

Beniamino Murgante Osvaldo Gervasi
Sanjay Misra Nadia Nedjah
Ana Maria A.C. Rocha David Taniar
Bernady O. Apduhan (Eds.)

LNCS 7335

Computational Science and Its Applications – ICCSA 2012

12th International Conference
Salvador de Bahia, Brazil, June 2012
Proceedings, Part III

3
Part III

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Beniamino Murgante Osvaldo Gervasi
Sanjay Misra Nadia Nedjah
Ana Maria A.C. Rocha David Taniar
Bernady O. Apduhan (Eds.)

Computational Science and Its Applications – ICCSA 2012

12th International Conference
Salvador de Bahia, Brazil, June 18-21, 2012
Proceedings, Part III

Volume Editors

Beniamino Murgante

University of Basilicata, Potenza, Italy, E-mail: beniamino.murgante@unibas.it

Oswaldo Gervasi

University of Perugia, Italy, E-mail: osvaldo@unipg.it

Sanjay Misra

Federal University of Technology, Minna, Nigeria, E-mail: smisra@futminna.edu.ng

Nadia Nedjah

State University of Rio de Janeiro, Brazil, E-mail: nadia@eng.uerj.br

Ana Maria A. C. Rocha

University of Minho, Braga, Portugal, E-mail: arocha@dps.uminho.pt

David Taniar

Monash University, Clayton, VIC, Australia, E-mail: david.taniar@infotech.monash.edu.au

Bernady O. Apduhan

Kyushu Sangyo University, Fukuoka, Japan, E-mail: bob@is.kyusan-u.ac.jp

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-31136-9

e-ISBN 978-3-642-31137-6

DOI 10.1007/978-3-642-31137-6

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012939389

CR Subject Classification (1998): C.2.4, C.2, H.4, F.2, H.3, D.2, F.1, H.5, H.2.8, K.6.5, I.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

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

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This four-part volume (LNCS 7333-7336) contains a collection of research papers from the 12th International Conference on Computational Science and Its Applications (ICCSA 2012) held in Salvador de Bahia, Brazil, during June 18–21, 2012. ICCSA is one of the successful international conferences in the field of computational sciences, and this year for the first time in the history of the ICCSA conference series it was held in South America. Previously the ICCSA conference series have been held in Santander, Spain (2011), Fukuoka, Japan (2010), Suwon, Korea (2009), Perugia, Italy (2008), Kuala Lumpur, Malaysia (2007), Glasgow, UK (2006), Singapore (2005), Assisi, Italy (2004), Montreal, Canada (2003), (as ICCS) Amsterdam, The Netherlands (2002), and San Francisco, USA (2001).

The computational science community has enthusiastically embraced the successive editions of ICCSA, thus contributing to making ICCSA a focal meeting point for those interested in innovative, cutting-edge research about the latest and most exciting developments in the field. We are grateful to all those who have contributed to the ICCSA conference series.

ICCSA 2012 would not have been made possible without the valuable contribution of many people. We would like to thank all session organizers for their diligent work, which further enhanced the conference level, and all reviewers for their expertise and generous effort, which led to a very high quality event with excellent papers and presentations. We specially recognize the contribution of the Program Committee and local Organizing Committee members for their tremendous support and for making this congress a very successful event. We would like to sincerely thank our keynote speakers, who willingly accepted our invitation and shared their expertise.

We also thank our publisher, Springer, for accepting to publish the proceedings and for their kind assistance and cooperation during the editing process.

Finally, we thank all authors for their submissions and all conference attendees for making ICCSA 2012 truly an excellent forum on computational science, facilitating the exchange of ideas, fostering new collaborations and shaping the future of this exciting field. Last, but certainly not least, we wish to thank our readers for their interest in this volume. We really hope you find in these pages interesting material and fruitful ideas for your future work.

We cordially invite you to visit the ICCSA website—<http://www.iccsa.org>—where you can find relevant information about this interesting and exciting event.

June 2012

Oswaldo Gervasi
David Taniar

Organization

ICCSA 2012 was organized by Universidade Federal da Bahia (Brazil), Universidade Federal do Recôncavo da Bahia (Brazil), Universidade Estadual de Feira de Santana (Brazil), University of Perugia (Italy), University of Basilicata (Italy), Monash University (Australia), and Kyushu Sangyo University (Japan).

Honorary General Chairs

Antonio Laganà	University of Perugia, Italy
Norio Shiratori	Tohoku University, Japan
Kenneth C.J. Tan	Qontix, UK

General Chairs

Oswaldo Gervasi	University of Perugia, Italy
David Taniar	Monash University, Australia

Program Committee Chairs

Bernady O. Apduhan	Kyushu Sangyo University, Japan
Beniamino Murgante	University of Basilicata, Italy

Workshop and Session Organizing Chairs

Beniamino Murgante	University of Basilicata, Italy
--------------------	---------------------------------

Local Organizing Committee

Frederico V. Prudente	Universidade Federal da Bahia, Brazil (Chair)
Mirco Ragni	Universidade Estadual de Feira de Santana, Brazil
Ana Carla P. Bitencourt	Universidade Federal do Recôncavo da Bahia, Brazil
Cassio Pigozzo	Universidade Federal da Bahia, Brazil
Angelo Duarde	Universidade Estadual de Feira de Santana, Brazil
Marcos E. Barreto	Universidade Federal da Bahia, Brazil
José Garcia V. Miranda	Universidade Federal da Bahia, Brazil

International Liaison Chairs

Jemal Abawajy	Deakin University, Australia
Marina L. Gavrilova	University of Calgary, Canada
Robert C.H. Hsu	Chung Hua University, Taiwan
Tai-Hoon Kim	Hannam University, Korea
Andrés Iglesias	University of Cantabria, Spain
Takashi Naka	Kyushu Sangyo University, Japan
Rafael D.C. Santos	National Institute for Space Research, Brazil

Workshop Organizers

Advances in High-Performance Algorithms and Applications (AHPAA 2012)

Massimo Cafaro	University of Salento, Italy
Giovanni Aloisio	University of Salento, Italy

Advances in Web-Based Learning (AWBL 2012)

Mustafa Murat Inceoglu	Ege University, Turkey
------------------------	------------------------

Bio-inspired Computing and Applications (BIOCA 2012)

Nadia Nedjah	State University of Rio de Janeiro, Brazil
Luiza de Macedo Mourell	State University of Rio de Janeiro, Brazil

Computer-Aided Modeling, Simulation, and Analysis (CAMSA 2012)

Jie Shen	University of Michigan, USA
Yuqing Song	Tianjing University of Technology and Education, China

Cloud Computing and Its Applications (CCA 2012)

Jemal Abawajy	University of Deakin, Australia
Osvaldo Gervasi	University of Perugia, Italy

Computational Geometry and Applications (CGA 2012)

Marina L. Gavrilova	University of Calgary, Canada
---------------------	-------------------------------

Chemistry and Materials Sciences and Technologies (CMST 2012)

Antonio Laganà University of Perugia, Italy

Cities, Technologies and Planning (CTP 2012)

Giuseppe Borruso University of Trieste, Italy
Beniamino Murgante University of Basilicata, Italy

Computational Tools and Techniques for Citizen Science and Scientific Outreach (CTTCS 2012)

Rafael Santos National Institute for Space Research, Brazil
Jordan Raddick and Johns Hopkins University, USA
Ani Thakar Johns Hopkins University, USA

Econometrics and Multidimensional Evaluation in the Urban Environment (EMEUE 2012)

Carmelo M. Torre Polytechnic of Bari, Italy
Maria Cerreta Università Federico II of Naples, Italy
Paola Perchinunno University of Bari, Italy

Future Information System Technologies and Applications (FISTA 2012)

Bernady O. Apduhan Kyushu Sangyo University, Japan

Geographical Analysis, Urban Modeling, Spatial Statistics (GEOG-AN-MOD 2012)

Stefania Bertazzon University of Calgary, Canada
Giuseppe Borruso University of Trieste, Italy
Beniamino Murgante University of Basilicata, Italy

International Workshop on Biomathematics, Bioinformatics and Biostatistics (IBBB 2012)

Unal Ufuktepe Izmir University of Economics, Turkey
Andrés Iglesias University of Cantabria, Spain

International Workshop on Collective Evolutionary Systems (IWCES 2012)

Alfredo Milani
Clement Leung

University of Perugia, Italy
Hong Kong Baptist University, Hong Kong

Mobile Communications (MC 2012)

Hyunseung Choo

Sungkyunkwan University, Korea

Mobile Computing, Sensing, and Actuation for Cyber Physical Systems (MSA4CPS 2012)

Moonseong Kim
Saad Qaisar

Korean intellectual Property Office, Korea
NUST School of Electrical Engineering and
Computer Science, Pakistan

Optimization Techniques and Applications (OTA 2012)

Ana Maria Rocha

University of Minho, Portugal

Parallel and Mobile Computing in Future Networks (PMCFUN 2012)

Al-Sakib Khan Pathan

International Islamic University Malaysia,
Malaysia

PULSES - Transitions and Nonlinear Phenomena (PULSES 2012)

Carlo Cattani
Ming Li
Shengyong Chen

University of Salerno, Italy
East China Normal University, China
Zhejiang University of Technology, China

Quantum Mechanics: Computational Strategies and Applications (QMCSA 2012)

Mirco Ragni
Frederico Vasconcellos

Universidade Federal de Bahia, Brazil

Prudente
Angelo Marconi Maniero

Universidade Federal de Bahia, Brazil
Universidade Federal de Bahia, Brazil

Ana Carla Peixoto Bitencourt

Universidade Federal do Recôncavo da Bahia,
Brazil

Remote Sensing Data Analysis, Modeling, Interpretation and Applications: From a Global View to a Local Analysis (RS 2012)

Rosa Lasaponara Institute of Methodologies for Environmental
Analysis, National Research Council, Italy
Nicola Masini Archaeological and Monumental Heritage
Institute, National Research Council, Italy

Soft Computing and Data Engineering (SCDE 2012)

Mustafa Matt Deris Universiti Tun Hussein Onn Malaysia, Malaysia
Tutut Herawan Universitas Ahmad Dahlan, Indonesia

Software Engineering Processes and Applications (SEPA 2012)

Sanjay Misra Federal University of Technology Minna,
Nigeria

Software Quality (SQ 2012)

Sanjay Misra Federal University of Technology Minna,
Nigeria

Security and Privacy in Computational Sciences (SPCS 2012)

Arijit Ukil Tata Consultancy Services, India

Tools and Techniques in Software Development Processes (TTSDP 2012)

Sanjay Misra Federal University of Technology Minna,
Nigeria

Virtual Reality and Its Applications (VRA 2012)

Oswaldo Gervasi University of Perugia, Italy
Andr es Iglesias University of Cantabria, Spain

Wireless and Ad-Hoc Networking (WADNet 2012)

Jongchan Lee
Sangjoon Park

Kunsan National University, Korea
Kunsan National University, Korea

Program Committee

Jemal Abawajy	Daekin University, Australia
Kenny Adamson	University of Ulster, UK
Filipe Alvelos	University of Minho, Portugal
Paula Amaral	Universidade Nova de Lisboa, Portugal
Hartmut Asche	University of Potsdam, Germany
Md. Abul Kalam Azad	University of Minho, Portugal
Michela Bertolotto	University College Dublin, Ireland
Sandro Bimonte	CEMAGREF, TSCF, France
Rod Blais	University of Calgary, Canada
Ivan Blecic	University of Sassari, Italy
Giuseppe Borruso	University of Trieste, Italy
Alfredo Buttari	CNRS-IRIT, France
Yves Caniou	Lyon University, France
José A. Cardoso e Cunha	Universidade Nova de Lisboa, Portugal
Leocadio G. Casado	University of Almeria, Spain
Carlo Cattani	University of Salerno, Italy
Mete Celik	Erciyes University, Turkey
Alexander Chemeris	National Technical University of Ukraine “KPI”, Ukraine
Min Young Chung	Sungkyunkwan University, Korea
Gilberto Corso Pereira	Federal University of Bahia, Brazil
M. Fernanda Costa	University of Minho, Portugal
Gaspar Cunha	University of Minho, Portugal
Carla Dal Sasso Freitas	Universidade Federal do Rio Grande do Sul, Brazil
Pradesh Debba	The Council for Scientific and Industrial Research (CSIR), South Africa
Frank Devai	London South Bank University, UK
Rodolphe Devillers	Memorial University of Newfoundland, Canada
Prabu Dorairaj	NetApp, India/USA
M. Irene Falcao	University of Minho, Portugal
Cherry Liu Fang	U.S. DOE Ames Laboratory, USA
Edite M.G.P. Fernandes	University of Minho, Portugal
Jose-Jesus Fernandez	National Centre for Biotechnology, CSIS, Spain
Maria Antonia Forjaz	University of Minho, Portugal
Maria Celia Furtado Rocha	PRODEB–PósCultura/UFBA, Brazil
Akemi Galvez	University of Cantabria, Spain
Paulino Jose Garcia Nieto	University of Oviedo, Spain
Marina Gavrilova	University of Calgary, Canada

Jerome Gensel	LSR-IMAG, France
Maria Giaoutzi	National Technical University, Athens, Greece
Andrzej M. Goscinski	Deakin University, Australia
Alex Hagen-Zanker	University of Cambridge, UK
Malgorzata Hanzl	Technical University of Lodz, Poland
Shanmugasundaram Hariharan	B.S. Abdur Rahman University, India
Eligius M.T. Hendrix	University of Malaga/Wageningen University, Spain/The Netherlands
Hisamoto Hiyoshi	Gunma University, Japan
Fermin Huarte	University of Barcelona, Spain
Andres Iglesias	University of Cantabria, Spain
Mustafa Inceoglu	EGE University, Turkey
Peter Jimack	University of Leeds, UK
Qun Jin	Waseda University, Japan
Farid Karimipour	Vienna University of Technology, Austria
Baris Kazar	Oracle Corp., USA
DongSeong Kim	University of Canterbury, New Zealand
Taihoon Kim	Hannam University, Korea
Ivana Kolingerova	University of West Bohemia, Czech Republic
Dieter Kranzlmüller	LMU and LRZ Munich, Germany
Antonio Laganà	University of Perugia, Italy
Rosa Lasaponara	National Research Council, Italy
Maurizio Lazzari	National Research Council, Italy
Cheng Siong Lee	Monash University, Australia
Sangyoun Lee	Yonsei University, Korea
Jongchan Lee	Kunsan National University, Korea
Clement Leung	Hong Kong Baptist University, Hong Kong
Chendong Li	University of Connecticut, USA
Gang Li	Deakin University, Australia
Ming Li	East China Normal University, China
Fang Liu	AMES Laboratories, USA
Xin Liu	University of Calgary, Canada
Savino Longo	University of Bari, Italy
Tinghuai Ma	NanJing University of Information Science and Technology, China
Sergio Maffioletti	University of Zurich, Switzerland
Ernesto Marcheggiani	Katholieke Universiteit Leuven, Belgium
Antonino Marvuglia	Research Centre Henri Tudor, Luxembourg
Nicola Masini	National Research Council, Italy
Nirvana Meratnia	University of Twente, The Netherlands
Alfredo Milani	University of Perugia, Italy
Sanjay Misra	Federal University of Technology Minna, Nigeria
Giuseppe Modica	University of Reggio Calabria, Italy

José Luis Montaña	University of Cantabria, Spain
Beniamino Murgante	University of Basilicata, Italy
Jiri Nedoma	Academy of Sciences of the Czech Republic, Czech Republic
Laszlo Neumann	University of Girona, Spain
Kok-Leong Ong	Deakin University, Australia
Belen Palop	Universidad de Valladolid, Spain
Marcin Paprzycki	Polish Academy of Sciences, Poland
Eric Pardede	La Trobe University, Australia
Kwangjin Park	Wonkwang University, Korea
Ana Isabel Pereira	Polytechnic Institute of Braganca, Portugal
Maurizio Pollino	Italian National Agency for New Technologies, Energy and Sustainable Economic Development, Italy
Alenka Poplin	University of Hamburg, Germany
Vidyasagar Potdar	Curtin University of Technology, Australia
David C. Prosperi	Florida Atlantic University, USA
Wenny Rahayu	La Trobe University, Australia
Jerzy Respondek	Silesian University of Technology Poland
Ana Maria A.C. Rocha	University of Minho, Portugal
Humberto Rocha	INESC-Coimbra, Portugal
Alexey Rodionov	Institute of Computational Mathematics and Mathematical Geophysics, Russia
Cristina S. Rodrigues	University of Minho, Portugal
Octavio Roncero	CSIC, Spain
Maytham Safar	Kuwait University, Kuwait
Haiduke Sarafian	The Pennsylvania State University, USA
Qi Shi	Liverpool John Moores University, UK
Dale Shires	U.S. Army Research Laboratory, USA
Takuo Suganuma	Tohoku University, Japan
Ana Paula Teixeira	University of Tras-os-Montes and Alto Douro, Portugal
Senhorinha Teixeira	University of Minho, Portugal
Parimala Thulasiraman	University of Manitoba, Canada
Carmelo Torre	Polytechnic of Bari, Italy
Javier Martinez Torres	Centro Universitario de la Defensa Zaragoza, Spain
Giuseppe A. Trunfio	University of Sassari, Italy
Unal Ufuktepe	Izmir University of Economics, Turkey
Mario Valle	Swiss National Supercomputing Centre, Switzerland
Pablo Vanegas	University of Cuenca, Ecuador
Piero Giorgio Verdini	INFN Pisa and CERN, Italy
Marco Vizzari	University of Perugia, Italy
Koichi Wada	University of Tsukuba, Japan

Krzysztof Walkowiak
 Robert Weibel
 Roland Wismüller
 Mudasser Wyne
 Chung-Huang Yang
 Xin-She Yang
 Salim Zabir
 Albert Y. Zomaya

Wroclaw University of Technology, Poland
 University of Zurich, Switzerland
 Universität Siegen, Germany
 SOET National University, USA
 National Kaohsiung Normal University, Taiwan
 National Physical Laboratory, UK
 France Telecom Japan Co., Japan
 University of Sydney, Australia

Sponsoring Organizations

ICCSA 2012 would not have been possible without tremendous support of many organizations and institutions, for which all organizers and participants of ICCSA 2012 express their sincere gratitude:



Universidade Federal da Bahia, Brazil
 (<http://www.ufba.br>)



Universidade Federal do Recôncavo da Bahia,
 Brazil
 (<http://www.ufrb.edu.br>)



Universidade Estadual de Feira de Santana,
 Brazil
 (<http://www.uefs.br>)



University of Perugia, Italy
 (<http://www.unipg.it>)



University of Basilicata, Italy
 (<http://www.unibas.it>)

XVI Organization



Monash University, Australia
(<http://monash.edu>)



Kyushu Sangyo University, Japan
(www.kyusan-u.ac.jp)



Brazilian Computer Society
(www.sbc.org.br)



Coordenação de Aperfeiçoamento de Pessoal de
Nível Superior (CAPES), Brazil
(<http://www.capes.gov.br>)



National Council for Scientific and
Technological Development (CNPq), Brazil
(<http://www.cnpq.br>)



SECRETARIA DE CIÊNCIA,
TECNOLOGIA E INOVAÇÃO



Fundação de Amparo à Pesquisa do Estado
da Bahia (FAPESB), Brazil
(<http://www.fapesb.ba.gov.br>)

Table of Contents – Part III

Workshop on Optimization Techniques and Applications (OTA 2012)

Incorporating Radial Basis Functions in Pattern Search Methods: Application to Beam Angle Optimization in Radiotherapy Treatment Planning	1
<i>Humberto Rocha, Joana M. Dias, Brigida C. Ferreira, and Maria do Carmo Lopes</i>	
On the Complexity of a Mehrotra-Type Predictor-Corrector Algorithm	17
<i>Ana Paula Teixeira and Regina Almeida</i>	
Design of Wood Biomass Supply Chains	30
<i>Tiago Costa Gomes, Filipe Pereira e Alvelos, and Maria Sameiro Carvalho</i>	
On Solving a Stochastic Programming Model for Perishable Inventory Control	45
<i>Eliugius M.T. Hendrix, Rene Haijema, Roberto Rossi, and Karin G.J. Pauls-Worm</i>	
An Artificial Fish Swarm Filter-Based Method for Constrained Global Optimization	57
<i>Ana Maria A.C. Rocha, M. Fernanda P. Costa, and Edite M.G.P. Fernandes</i>	
Solving Multidimensional 0–1 Knapsack Problem with an Artificial Fish Swarm Algorithm	72
<i>Md. Abul Kalam Azad, Ana Maria A.C. Rocha, and Edite M.G.P. Fernandes</i>	
Optimization Model of COTS Selection Based on Cohesion and Coupling for Modular Software Systems under Multiple Applications Environment	87
<i>Pankaj Gupta, Shilpi Verma, and Mukesh Kumar Mehlawat</i>	
A Derivative-Free Filter Driven Multistart Technique for Global Optimization	103
<i>Florabela P. Fernandes, M. Fernanda P. Costa, and Edite M.G.P. Fernandes</i>	

On Lower Bounds Using Additively Separable Terms in Interval B&B	119
<i>José L. Berenguel, Leocadio G. Casado, I. García, Eligius M.T. Hendrix, and F. Messine</i>	
A Genetic Algorithm for the Job Shop on an ASRS Warehouse	133
<i>José Figueiredo, José A. Oliveira, Luis Dias, and Guilherme A.B. Pereira</i>	
On Solving the Profit Maximization of Small Cogeneration Systems	147
<i>Ana C.M. Ferreira, Ana Maria A.C. Rocha, Senhorinha F.C.F. Teixeira, Manuel L. Nunes, and Luís B. Martins</i>	
Global Optimization Simplex Bisection Revisited Based on Considerations by Reiner Horst	159
<i>Eligius M.T. Hendrix, Leocadio G. Casado, and Paula Amaral</i>	
Application of Variance Analysis to the Combustion of Residual Oils ...	174
<i>Manuel Ferreira and José Carlos Teixeira</i>	
Warehouse Design and Planning: A Mathematical Programming Approach	187
<i>Carla A.S. Geraldés, Maria Sameiro Carvalho, and Guilherme A.B. Pereira</i>	
Application of CFD Tools to Optimize Natural Building Ventilation Design	202
<i>José Carlos Teixeira, Ricardo Lomba, Senhorinha F.C.F. Teixeira, and Pedro Lobarinhas</i>	
 Workshop on Mobile Communications (MC 2012)	
Middleware Integration for Ubiquitous Sensor Networks in Agriculture	217
<i>Junghoon Lee, Gyung-Leen Park, Min-Jae Kang, Ho-Young Kwak, Sang Joon Lee, and Jikwang Han</i>	
Usage Pattern-Based Prefetching: Quick Application Launch on Mobile Devices	227
<i>Hokwon Song, Changwoo Min, Jeehong Kim, and Young Ik Eom</i>	
EIMOS: Enhancing Interactivity in Mobile Operating Systems	238
<i>Sunwook Bae, Hokwon Song, Changwoo Min, Jeehong Kim, and Young Ik Eom</i>	

Development of Mobile Hybrid MedIntegraWeb App for Interoperation between u-RPMS and HIS	248
<i>Young-Hyuk Kim, Il-Kown Lim, Jae-Pil Lee, Jae-Gwang Lee, and Jae-Kwang Lee</i>	
A Distributed Lifetime-Maximizing Scheme for Connected Target Coverage in WSNs	259
<i>Duc Tai Le, Thang Le Duc, and Hyunseung Choo</i>	
Reducing Last Level Cache Pollution in NUMA Multicore Systems for Improving Cache Performance	272
<i>Deukhyeon An, Jeehong Kim, JungHyun Han, and Young Ik Eom</i>	
The Fast Handover Scheme for Mobile Nodes in NEMO-Enabled PMIPv6	283
<i>Changyong Park, Junbeom Park, Hao Wang, and Hyunseung Choo</i>	
A Reference Model for Virtual Resource Description and Discovery in Virtual Networks	297
<i>Yuemei Xu, Yanni Han, Wenjia Niu, Yang Li, Tao Lin, and Song Ci</i>	
TV Remote Control Using Human Hand Motion Based on Optical Flow System	311
<i>Soonmook Jeong, Taehoun Song, Keyho Kwon, and Jae Wook Jeon</i>	
Fast and Reliable Data Forwarding in Low-Duty-Cycle Wireless Sensor Networks	324
<i>Junseong Choe, Nguyen Phan Khanh Ha, Junguye Hong, and Hyunseung Choo</i>	
Workshop on Mobile-Computing, Sensing, and Actuation for Cyber Physical Systems (MSA4CPS 2012)	
Neural Network and Physiological Parameters Based Control of Artificial Pancreas for Improved Patient Safety	339
<i>Saad Bin Qaisar, Salman H. Khan, and Sahar Imtiaz</i>	
A Genetic Algorithm Assisted Resource Management Scheme for Reliable Multimedia Delivery over Cognitive Networks	352
<i>Salman Ali, Ali Munir, Saad Bin Qaisar, and Junaid Qadir</i>	
Performance Analysis of WiMAX Best Effort and ertPS Service Classes for Video Transmission	368
<i>Hassan Abid, Haroon Raja, Ali Munir, Jaweria Amjad, Aliya Mazhar, and Dong-Young Lee</i>	

Jump Oriented Programming on Windows Platform (on the x86)	376
<i>Jae-Won Min, Sung-Min Jung, Dong-Young Lee, and Tai-Myoung Chung</i>	
Cryptanalysis and Improvement of a Biometrics-Based Multi-server Authentication with Key Agreement Scheme	391
<i>Hakhyun Kim, Woongryul Jeon, Kwangwoo Lee, Yunho Lee, and Dongho Won</i>	
Rate-Distortion Optimized Transcoder Selection for Multimedia Transmission in Heterogeneous Networks	407
<i>Haroon Raja and Saad Bin Qaisar</i>	
Formal Probabilistic Analysis of Cyber-Physical Transportation Systems	419
<i>Atif Mashkooor and Osman Hasan</i>	
Workshop on Remote Sensing (RS 2012)	
DEM Reconstruction of Coastal Geomorphology from DINSAR	435
<i>Maged Marghany</i>	
Three-Dimensional Coastal Front Visualization from RADARSAT-1 SAR Satellite Data	447
<i>Maged Marghany</i>	
A New Self-Learning Algorithm for Dynamic Classification of Water Bodies	457
<i>Bernd Fichtelmann and Erik Borg</i>	
DEM Accuracy of High Resolution Satellite Images	471
<i>Mustafa Yanalak, Nebiye Musaoglu, Cengizhan Ipbuker, Elif Sertel, and Sinasi Kaya</i>	
Low Cost Pre-operative Fire Monitoring from Fire Danger to Severity Estimation Based on Satellite MODIS, Landsat and ASTER Data: The Experience of FIRE-SAT Project in the Basilicata Region (Italy)	481
<i>Antonio Lanorte, Fortunato De Santis, Angelo Aromando, and Rosa Lasaponara</i>	
Investigating Satellite Landsat TM and ASTER Multitemporal Data Set to Discover Ancient Canals and Acqueduct Systems	497
<i>Rosa Lasaponara and Nicola Masini</i>	
Using Spatial Autocorrelation Techniques and Multi-temporal Satellite Data for Analyzing Urban Sprawl	512
<i>Gabriele Nolè, Maria Danese, Beniamino Murgante, Rosa Lasaponara, and Antonio Lanorte</i>	

General Track on Information Systems and Technologies

A Framework for QoS Based Dynamic Web Services Composition	528
<i>Jigyasu Nema, Rajdeep Niyogi, and Alfredo Milani</i>	
Data Summarization Model for User Action Log Files	539
<i>Eleonora Gentili, Alfredo Milani, and Valentina Poggioni</i>	
User Modeling for Adaptive E-Learning Systems	550
<i>Birol Ciloglugil and Mustafa Murat Inceoglu</i>	
An Experimental Study of the Combination of Meta-Learning with Particle Swarm Algorithms for SVM Parameter Selection	562
<i>Péricles B.C. de Miranda, Ricardo B.C. Prudêncio, Andre Carlos P.L.F. de Carvalho, and Carlos Soares</i>	
An Investigation into Agile Methods in Embedded Systems Development	576
<i>Caroline Oliveira Albuquerque, Pablo Oliveira Antonino, and Elisa Yumi Nakagawa</i>	
Heap Slicing Using Type Systems	592
<i>Mohamed A. El-Zawawy</i>	
Using Autonomous Search for Generating Good Enumeration Strategy Blends in Constraint Programming	607
<i>Ricardo Soto, Broderick Crawford, Eric Monfroy, and Víctor Bustos</i>	
Evaluation of Normalization Techniques in Text Classification for Portuguese	618
<i>Merley da Silva Conrado, Víctor Antonio Laguna Gutiérrez, and Solange Oliveira Rezende</i>	
Extracting Definitions from Brazilian Legal Texts	631
<i>Edilson Ferneda, Hércules Antonio do Prado, Augusto Herrmann Batista, and Marcello Sandi Pinheiro</i>	
A Heuristic Diversity Production Approach	647
<i>Hamid Parvin, Hosein Alizadeh, Sajad Parvin, and Behzad Maleki</i>	
Structuring Taxonomies from Texts: A Case-Study on Defining Soil Classes	657
<i>Hércules Antonio do Prado, Edilson Ferneda, Francisco Carlos da Luz Rodrigues, Éder Martins de Souza, Osmar Abílio de Carvalho Jr., and Alfredo José Barreto Luiz</i>	
Exploring Fuzzy Ontologies in Mining Generalized Association Rules . . .	667
<i>Rodrigo Moura Juvenil Ayres, Marcela Xavier Ribeiro, and Marilde Terezinha Prado Santos</i>	

BTA: Architecture for Reusable Business Tier Components with Access Control	682
<i>Óscar Mortágua Pereira, Rui L. Aguiar, and Maribel Yasmina Santos</i>	
Analysing the PDDL Language for Argumentation-Based Negotiation Planning	698
<i>Ariel Monteserin, Luis Berdún, and Analía A. Amandi</i>	
Predicting Potential Responders in Twitter: A Query Routing Algorithm	714
<i>Cleyton Caetano de Souza, Jonathas José de Magalhães, Evandro Barros de Costa, and Joseana Macêdo Fechine</i>	
Towards a Goal Recognition Model for the Organizational Memory	730
<i>Marcelo G. Armentano and Analía A. Amandi</i>	
SART: A New Association Rule Method for Mining Sequential Patterns in Time Series of Climate Data	743
<i>Marcos Daniel Cano, Marilde Terezinha Prado Santos, Ana Maria H. de Avila, Luciana A.S. Romani, Agma J.M. Traina, and Marcela Xavier Ribeiro</i>	
Author Index	759

Incorporating Radial Basis Functions in Pattern Search Methods: Application to Beam Angle Optimization in Radiotherapy Treatment Planning

Humberto Rocha¹, Joana M. Dias^{1,2}, Brigida C. Ferreira^{3,4},
and Maria do Carmo Lopes^{3,4}

¹ INESC-Coimbra, Rua Antero de Quental, 199
3000-033 Coimbra, Portugal

² Faculdade de Economia, Universidade de Coimbra,
3004-512 Coimbra, Portugal

³ I3N, Departamento de Física, Universidade de Aveiro,
3810-193 Aveiro, Portugal

⁴ Serviço de Física Médica, IPOC-FG, EPE,
3000-075 Coimbra, Portugal

hrocha@mat.uc.pt, joana@fe.uc.pt, brigida@ua.pt,
mclopes@ipocoimbra.min-saude.pt

Abstract. The global optimization of black-box functions with many local minima occurs in many branches of science and engineering. There are many methods and heuristics to address this type of problems. However, for problems with expensive black-box functions, both in terms of cost or time, the number of function evaluations required by most of the methods or heuristics is prohibitive. The pattern search methods framework is suited to address this type of problems since it requires few function value evaluations to converge and have the ability to avoid local entrapment. The ability of this class of methods to obtain global minima depends on the incorporation of methods or heuristics for global optimization on their, so called, search step. We propose the use of radial basis functions both to influence the quality of the local minimizer found by the method and also to obtain a better coverage of the search space. Our approach is tailored for addressing the beam angle optimization (BAO) problem in intensity modulated radiation therapy treatment planning, but can be easily extended for other general problems. The BAO problem is quite difficult, and yet to be solved in a satisfactory way, since it is a highly non-convex optimization problem with many local minima. A couple of retrospective treated cases of head-and-neck tumors at the Portuguese Institute of Oncology of Coimbra is used to discuss the benefits of using our approach in the optimization of the BAO problem.

Keywords: Pattern Search Methods, Radial Basis Functions, Radiotherapy, IMRT, Beam Angle Optimization.

1 Introduction

The global optimization of black-box functions with many local minima occurs in many branches of science and engineering. Directional direct-search methods have been used to tackle this type of problems [1]. The pattern search methods framework is the most used and implemented class of directional direct-search methods [3]. Pattern search methods are organized around two steps at every iteration: the poll step and the search step. The poll step performs a local search in a neighborhood around the current iterate using the concepts of positive bases, and under the appropriate assumptions, it guarantees global convergence to stationary points. The search step consists of a finite search, free of rules, away from the current iterate, and the ability to obtain global minima depends on the incorporation of methods or heuristics for global optimization. An example of such hybridization is the use of particle swarm optimization in the search step of the pattern search methods framework [19]. However, for problems with expensive black-box functions, both in terms of cost or time, the number of function evaluations required for this type of strategies is prohibitive for obtaining results in an acceptable time frame or within the budget. The beam angle optimization (BAO) problem in intensity modulated radiation therapy treatment planning is such problem and will be used to illustrate the merits of our approach. The intensity modulated radiation therapy (IMRT) is a modern type of radiation therapy, whose inverse planning leads to complex optimization problems, including the BAO problem - the problem of deciding which incidence radiation beam angles should be used. The BAO problem is quite difficult, and yet to be solved in a satisfactory way, since it is a highly non-convex optimization problem with many local minima [5]. Therefore, methods that avoid being easily trapped in local minima should be used. Moreover, each function evaluation is time expensive so methods that require few function value evaluations should be used to tackle the BAO problem. The pattern search methods framework is suited to address the BAO problem since it requires few function value evaluations to converge and have the ability to avoid local entrapment. Here, we will discuss the benefits of incorporating radial basis functions in the pattern search methods framework for the optimization of the highly non-convex BAO problem. Radial basis functions are used both to influence the quality of the local minimizer found by the method and also to obtain a better coverage of the search space in amplitude. A couple of retrospective treated cases of head-and-neck tumors at the Portuguese Institute of Oncology of Coimbra is used to discuss the benefits of using our approach in the optimization of the BAO problem. Our approach is tailored to address this particular problem but it can be easily extended for other general problems. The paper is organized as follows. In the next section we describe the BAO problem. Radial basis functions interpolation and its use within the pattern search methods framework is presented in section 3. Clinical examples of head-and-neck cases used in the computational tests are presented in section 4. Section 5 presents the experimental results. In the last section we have the conclusions.

2 Beam Angle Optimization in IMRT Treatment Planning

The purpose of radiation therapy is to deliver a dose of radiation to the tumor volume to sterilize all cancer cells minimizing the collateral effects on the surrounding healthy organs and tissues. Typically, radiation is generated by a linear accelerator mounted on a gantry that can rotate along a central axis and is delivered with the patient immobilized on a couch that can rotate. The rotation of the couch combined with the rotation of the gantry allows radiation from almost any angle around the tumor. In IMRT the radiation beam is modulated by a multileaf collimator that enables the transformation of the beam into a grid of smaller beamlets of independent intensities. A common way to solve the inverse planning in IMRT optimization problems is to use a beamlet-based approach leading to a large-scale programming problem. Due to the complexity of the whole optimization problem, many times the treatment planning is divided into three smaller problems which can be solved sequentially: BAO problem, fluence map optimization (FMO) problem, and leaf sequencing problem. Here, we will focus our attention in the BAO problem, using coplanar angles, and we will assume that the number of beam angles is defined a priori by the treatment planner.

Many attempts to address the BAO problem can be found in the literature including simulated annealing [4], genetic algorithms [10], particle swarm optimization [12] or other heuristics incorporating a priori knowledge of the problem. Although those global heuristics can theoretically avoid local optima, globally optimal or even clinically better solutions can not be obtained without a large number of objective function evaluations. For that reason, many of the previous BAO studies are based on a variety of scoring methods or approximations to the FMO to gauge the quality of the beam angle set. When the BAO problem is not based on the optimal FMO solutions, the resulting beam angle set has no guarantee of optimality and has questionable reliability since it has been extensively reported that optimal beam angles for IMRT are often non-intuitive. Therefore, our approach for modeling the BAO problem, similarly to [2,5], uses the optimal solution value of the FMO problem as the measure of the quality for a given beam angle set. Thus, we will present the formulation of the BAO problem followed by the formulation of the FMO problem we used.

2.1 BAO Model

Let us consider n to be the fixed number of (coplanar) beam directions, i.e., n beam angles are chosen on a circle around the CT-slice of the body that contains the isocenter (usually the center of mass of the tumor). Typically, the BAO problem is formulated as a combinatorial optimization problem in which a specified number of beam angles is to be selected among a beam angle candidate pool. The continuous $[0^\circ, 360^\circ]$ gantry angles are generally discretized into equally spaced directions with a given angle increment, such as 5 or 10 degrees. We will consider a different approach for the formulation of the BAO problem.

All continuous $[0^\circ, 360^\circ]$ gantry angles will be considered instead of a discretized sample. Since the angle -5° is equivalent to the angle 355° and the angle 365° is the same as the angle 5° , we can avoid a bounded formulation. A basic formulation for the BAO problem is obtained by selecting an objective function such that the best set of beam angles is obtained for the function's minimum:

$$\begin{aligned} \min & f(\theta_1, \dots, \theta_n) \\ \text{s.t.} & (\theta_1, \dots, \theta_n) \in \mathbb{R}^n. \end{aligned}$$

Here, the objective $f(\theta_1, \dots, \theta_n)$ that measures the quality of the set of beam directions $\theta_1, \dots, \theta_n$ is the optimal value of the FMO problem for each fixed set of beam directions. Such functions have numerous local optima, which increases the difficulty of obtaining a good global solution. Thus, the choice of the solution method becomes a critical aspect for obtaining a good solution. Our formulation was mainly motivated by the ability of using a class of solution methods that we consider to be suited to successfully address the BAO problem: pattern search methods. The FMO model used is presented next.

2.2 FMO Model

For a given beam angle set, an optimal IMRT plan is obtained by solving the FMO problem - the problem of determining the optimal beamlet weights for the fixed beam angles. Many mathematical optimization models and algorithms have been proposed for the FMO problem, including linear models [17], mixed integer linear models [11] and nonlinear models [2].

Radiation dose distribution deposited in the patient, measured in Gray (Gy), needs to be assessed accurately in order to solve the FMO problem, i.e., to determine optimal fluence maps. Each structure's volume is discretized into voxels (small volume elements) and the dose is computed for each voxel using the superposition principle, i.e., considering the contribution of each beamlet. Typically, a dose matrix D is constructed from the collection of all beamlet weights, by indexing the rows of D to each voxel and the columns to each beamlet, i.e., the number of rows of matrix D equals the number of voxels (N_v) and the number of columns equals the number of beamlets (N_b) from all beam directions considered. Therefore, using matrix format, we can say that the total dose received by the voxel i is given by $\sum_{j=1}^{N_b} D_{ij}w_j$, with w_j the weight of beamlet j . Usually, the total number of voxels considered reaches the tens of thousands, thus the row dimension of the dose matrix is of that magnitude. The size of D originates large-scale problems being one of the main reasons for the difficulty of solving the FMO problem.

Here, we will use a convex penalty function voxel-based nonlinear model [2]. In this model, each voxel is penalized according to the square difference of the amount of dose received by the voxel and the amount of dose desired/allowed for the voxel. This formulation yields a quadratic programming problem with

only linear non-negativity constraints on the fluence values [17]:

$$\min_w \sum_{i=1}^{N_v} \frac{1}{v_S} \left[\underline{\lambda}_i \left(T_i - \sum_{j=1}^{N_b} D_{ij} w_j \right)_+^2 + \bar{\lambda}_i \left(\sum_{j=1}^{N_b} D_{ij} w_j - T_i \right)_+^2 \right]$$

$$s.t. \quad w_j \geq 0, \quad j = 1, \dots, N_b,$$

where T_i is the desired dose for voxel i , $\underline{\lambda}_i$ and $\bar{\lambda}_i$ are the penalty weights of underdose and overdose of voxel i , and $(\cdot)_+ = \max\{0, \cdot\}$. Although this formulation allows unique weights for each voxel, similarly to the implementation in [2], weights are assigned by structure only so that every voxel in a given structure has the weight assigned to that structure divided by the number of voxels of the structure (v_S). This nonlinear formulation implies that a very small amount of underdose or overdose may be accepted in clinical decision making, but larger deviations from the desired/allowed doses are decreasingly tolerated [2].

The FMO model is used as a black-box function. It is beyond the scope of this study to discuss if this formulation of the FMO problem is preferable to others. The conclusions drawn regarding BAO coupled with this nonlinear model are valid also if different FMO formulations are considered.

3 Radial Basis Function Interpolation and Its Use within the Pattern Search Methods Framework

For numerical approximation of multivariate functions, radial basis functions (RBFs) can provide excellent interpolants. For any finite data set in any Euclidean space, one can construct an interpolation of the data by using RBFs, even if the data points are unevenly and sporadically distributed in a high dimensional Euclidean space. However, RBF interpolant trends between and beyond the data points depend on the RBF used and may exhibit undesirable trends using some RBFs while the trends may be desirable using other RBFs. Numerical choice of the most adequate RBF for the problem at hand should be done instead of an usual a priori choice [15]. Next, we will formulate RBF interpolation problems and describe the strategy used to take advantage of the incorporation of RBF interpolants in the pattern search method framework applied to the BAO problem.

3.1 RBF Interpolation Problems

Let $f(\mathbf{x})$ be the true response to a given input vector \mathbf{x} (of n components) such that the value of f is only known at a set of N input vectors $\mathbf{x} = \mathbf{x}^1, \dots, \mathbf{x}^N$, i.e., only $f(\mathbf{x}^k)$ ($k = 1, \dots, N$) are known. An interpolation model $g(\mathbf{x})$ generated from a RBF $\varphi(t)$ can be represented in the following form:

$$g(\mathbf{x}) = \sum_{j=1}^N \alpha_j \varphi(\|\mathbf{x} - \mathbf{x}^j\|), \quad (1)$$

where α_j are the coefficients to be determined by interpolation conditions, $g(\mathbf{x}^k) = f(\mathbf{x}^k)$ ($k = 1, \dots, N$), $\|\mathbf{x} - \mathbf{x}^j\|$ denotes the parameterized distance between \mathbf{x} and \mathbf{x}^j defined as $\|\mathbf{x} - \mathbf{x}^j\| = \sqrt{\sum_{i=1}^n |\theta_i| (x_i - x_i^j)^2}$, and $\theta_1, \dots, \theta_n$ are scalars [15]. For fixed parameters θ_i , the coefficients $\alpha_1, \dots, \alpha_N$ in Eq. (1) can be calculated by solving the following linear system of interpolation equations:

$$\sum_{j=1}^N \alpha_j \varphi(\|\mathbf{x}^k - \mathbf{x}^j\|) = f(\mathbf{x}^k), \quad \text{for } k = 1, \dots, N. \quad (2)$$

The most popular examples of RBF [14] are cubic spline $\varphi(t) = t^3$, thin plate spline $\varphi(t) = t^2 \ln t$, multiquadric $\varphi(t) = \sqrt{1 + t^2}$, and Gaussian $\varphi(t) = \exp(-t^2)$. These RBFs can be used to model cubic, almost quadratic, and linear growth rates, as well as exponential decay, of the response for trend predictions. A unique interpolant is guaranteed for multiquadric and Gaussian RBFs, (i.e., the system matrix in Eq. (2) is nonsingular) even if the input vectors \mathbf{x}^j are few and poorly distributed, provided only that the input vectors are all different when $N > 1$. However, for cubic and thin plate spline RBFs, the system matrix in Eq. (2) might be singular [14]. An easy way to avoid this problem on the cubic and thin plate spline RBF interpolants is to add low-degree polynomials to interpolation functions in Eq. (1) (see [15]).

The constructed interpolant $g(\mathbf{x})$ in Eq. (1) depends on “subjective” choice of $\varphi(t)$, and model parameters $\theta_1, \dots, \theta_n$. While one can try all the possible choices of $\varphi(t)$ in search of a desirable interpolant, there are infinitely many choices for $\theta_1, \dots, \theta_n$. Mathematically, one could pick any fixed set of $\theta_1, \dots, \theta_n$ and construct the interpolation function for the given data. However, two different sets of $\theta_1, \dots, \theta_n$ will lead to two interpolation models that behave very differently between the input vectors $\mathbf{x}^1, \dots, \mathbf{x}^N$. Model parameter tuning for RBF interpolation aims at finding a set of parameters $\theta_1, \dots, \theta_n$ that results in the best prediction of the unknown response based on the available data. The prediction accuracy can be used as a criterion for choosing the best basis function $\varphi(t)$ and parameters θ_i . Cross-validation (CV) [18] was proposed to find $\varphi(t)$ and θ_i that lead to an approximate response model $g(\mathbf{x})$ with optimal prediction capability and proved to be effective [18]. The leave-one-out CV procedure is usually used in model parameter tuning for RBF interpolation [18]:

Algorithm 1. (Leave-one-out cross-validation for RBF interpolation)

1. Fix a set of parameters $\theta_1, \dots, \theta_n$.
2. For $j = 1, \dots, N$, construct the RBF interpolant $g_{-j}(\mathbf{x})$ of the data points $(\mathbf{x}^k, f(\mathbf{x}^k))$ for $1 \leq k \leq N, k \neq j$.
3. Use the following CV root mean square error as the prediction error:

$$E^{CV}(\theta_1, \dots, \theta_n) = \sqrt{\frac{1}{N} \sum_{j=1}^N (g_{-j}(\mathbf{x}^j) - f(\mathbf{x}^j))^2}. \quad (3)$$

The goal of model parameter tuning by CV is to find $\theta_1, \dots, \theta_n$ that minimize the CV error, $E^{CV}(\theta_1, \dots, \theta_n)$, so that the interpolation model has the highest prediction accuracy when CV error is the measure. Using different θ_i allows the model parameter tuning to scale each variable x_i based on its significance in modeling the variance in the response, thus, has the benefit of implicit variable screening built in the model parameter tuning.

3.2 Incorporation of RBF Models in the Pattern Search Methods Framework Tailored for the BAO Problem

Pattern search methods are directional direct search methods that belong to a broader class of derivative-free optimization methods, such that iterate progression is solely based on a finite number of function evaluations in each iteration, without explicit or implicit use of derivatives. Pattern search methods generate a sequence of non-increasing iterates $\{\mathbf{x}^k\}$ using positive bases (or positive spanning sets) and moving towards a direction that would produce a function decrease. A positive basis for \mathbb{R}^n can be defined as a set of nonzero vectors of \mathbb{R}^n whose positive combinations span \mathbb{R}^n (positive spanning set), but no proper set does. A positive spanning set contains at least one positive basis. It can be shown that a positive basis for \mathbb{R}^n contains at least $n + 1$ vectors and cannot contain more than $2n$ [8]. Positive basis with $n + 1$ and $2n$ elements are referred to as minimal and maximal positive basis, respectively. Commonly used minimal and maximal positive basis are $[I - e]$, with I being the identity matrix of dimension n and $e = [1 \dots 1]^\top$, and $[I - I]$, respectively.

One of the main features of positive bases (or positive spanning sets), that is the motivation for directional direct search methods, is that, unless the current iterate is at a stationary point, there is always a vector \mathbf{v}^i in a positive basis (or positive spanning set) that is a descent direction [8], i.e., there is an $\alpha > 0$ such that $f(\mathbf{x}^k + \alpha \mathbf{v}^i) < f(\mathbf{x}^k)$. This is the core of directional direct search methods and in particular of pattern search methods. The notions and motivations for the use of positive bases, its properties and examples can be found in [1, 8].

Pattern search methods framework is briefly presented next. Let us denote by \mathbf{V} the $n \times p$ matrix whose columns correspond to the p ($\geq n + 1$) vectors forming a positive spanning set. Given the current iterate \mathbf{x}^k , at each iteration k , the next point \mathbf{x}^{k+1} , aiming to provide a decrease of the objective function, is chosen from a finite number of candidates on a given mesh $M_k = \{\mathbf{x}^k + \alpha_k \mathbf{V} \mathbf{z} : \mathbf{z} \in \mathbb{Z}_+^p\}$, where α_k is the mesh-size (or step-size) parameter and \mathbb{Z}_+ is the set of nonnegative integers. Pattern search methods are organized around two steps at every iteration. The first step consists of a finite search on the mesh, free of rules, with the goal of finding a new iterate that decreases the value of the objective function at the current iterate. This step, called the search step, has the flexibility to use any strategy, method or heuristic, or take advantage of a priori knowledge of the problem at hand, as long as it searches only a finite number of points in the mesh. The search step provides the flexibility for a global search since it allows searches away from the neighborhood of the current iterate, and influences the quality of the local minimizer or stationary point found

by the method. If the search step fails to produce a decrease in the objective function, a second step, called the poll step, is performed around the current iterate. The poll step follows stricter rules and, using the concepts of positive bases, attempts to perform a local search in a mesh neighborhood around \mathbf{x}^k , $\mathcal{N}(\mathbf{x}^k) = \{\mathbf{x}^k + \alpha_k \mathbf{v} : \text{for all } \mathbf{v} \in P_k\} \subset M_k$, where P_k is a positive basis chosen from the finite positive spanning set \mathbf{V} . For a sufficiently small mesh-size parameter α_k , the poll step is guaranteed to provide a function reduction, unless the current iterate is at a stationary point [1]. So, if the poll step also fails to produce a function reduction, the mesh-size parameter α_k must be decreased. On the other hand, if both the search and poll steps fail to obtain an improved value for the objective function, the mesh-size parameter is increased or held constant.

The most common choice for the mesh-size parameter update is to half the mesh-size parameter at unsuccessful iterations and to keep it or double it at successful ones. Note that, if the initial mesh parameter is a power of 2, ($\alpha_0 = 2^l, l \in \mathbb{N}$), and the initial point is a vector of integers, using this common mesh update, all iterates will be a vector of integers until the mesh-size parameter becomes inferior to 1. This possibility is rather interesting for the BAO problem.

Recently, the efficiency of pattern search methods improved significantly by reordering the poll directions according to descent indicators built from simplex gradients [7]. Here, the poll directions are reordered according to the RBF model values. The most common approach for incorporating interpolation models in the search step consists of forming an interpolation model and finding its minimum. For example, in Custódio et al. [6], the search step computes a single trial point using minimum Frobenius norm quadratic models to be minimized within a trust region. The size of the trust region is coupled to the radius of the sample set. Thus, for an effective global search, the sample points should span all the search space. That could be achieved by using larger initial step-size parameters. However, since the BAO problem has many local minima and the number of sample points is scarce, the polynomial interpolation or regression models (usually quadratic models) used within the trust region struggle to find the best local minima. Therefore, starting with larger mesh-size parameters may lead to similar or worst results obtained when starting with smaller mesh-size parameters and at the cost of more function value evaluations [16]. An alternative and popular approach to keep small mesh-size parameters and still have a good coverage of the whole search space is to use a multi-start approach. However, the multi-start approach has the disadvantage of increasing the total number of function evaluations and with that the overall computational time. Moreover, the obtained good span of \mathbb{R}^2 in amplitude is only obtained by overlapping all the iterates giving the illusion that unusual beam angle configurations were tested while in fact only local searches around the initial beam angle configurations were performed. We adopted a different strategy, by considering a single starting point, a small initial mesh-size parameter, and trying to obtain a good span in amplitude of \mathbb{R}^2 by incorporating radial basis functions models in the search

step. The strategy sketched here is tailored for addressing the BAO problem and does not include the formal minimization of the RBF model:

Algorithm 2. (PSM framework using RBFs for the BAO problem)

- 0. Initialization** Set $k = 0$. Choose $\mathbf{x}^0 \in \mathbb{R}^n$, $\alpha_0 > 0$, and a positive spanning set \mathbf{V} .
- 1. Search step** If the number of evaluated points is not greater than $n + 1$ skip the search step. Otherwise, build a RBF model and while a decrease on the objective function value is not achieved, compute the RBFs trial points:
For each beam angle direction ($i = 1, \dots, n$)
- a. Evaluate the RBF model for every degree between the previous beam direction and the next one.
 - b. Find the minimum of those values that correspond to a beam direction that was not evaluated yet and, is at least 4 degrees away from a previously evaluated one, for the beam direction at stake.
 - c. Take as RBF trial point the current iterate updating the beam direction corresponding to the minimum found in b.
- If no RBF trial point correspond to a decrease on the objective function value, go to step 2 and the search step is declared unsuccessful. Otherwise, go to step 4 and both the search step and iteration are declared successful.
- 2. Poll step** This step is only performed if the search step is unsuccessful. If a RBF model was computed in the previous step then reorder the poll directions according to the RBF model values. If $f(\mathbf{x}^k) \leq f(\mathbf{x})$ for every \mathbf{x} in the mesh neighborhood $\mathcal{N}(\mathbf{x}^k)$, then go to step 3 and shrink M_k . Both poll step and iteration are declared unsuccessful. Otherwise, choose a point $\mathbf{x}^{k+1} \in \mathcal{N}(\mathbf{x}^k)$ such that $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ and go to step 4. Both poll step and iteration are declared successful.
- 3. Mesh reduction** Let $\alpha_{k+1} = \frac{1}{2} \times \alpha_k$. Set $k = k + 1$ and return to step 1.
- 4. Mesh expansion** Let $\alpha_{k+1} = \alpha_k$. Set $k = k + 1$ and return to step 1.

Our main goal for using a RBF model in the search step of the pattern search methods framework is to properly explore the search space in amplitude without a random criteria. Therefore, each beam direction is tested every degree between the previous beam direction and the next one as stated in step a of the RBF trial points computation. A proper minimization is unnecessary since we are interested in integer beam angle directions. Step b of the RBF trial points computation within search step forces the algorithm to consider only directions in regions not yet explored which is the main goal of the RBF models here (directions that are less than 4 degrees apart are considered to be clinically equivalent). The maximum number of points computed in the search step is n (e.g. if the search step is unsuccessful). More conservative strategies could be adopted considering, e.g., only the best of the RBF trial points.

The benefits of using RBFs in the pattern search methods framework for the optimization of the BAO problem are illustrated using a set of clinical examples of head-and-neck cases that are presented next.

4 Head-and-Neck Clinical Examples

Two clinical examples of retrospective treated cases of head-and-neck tumors at the Portuguese Institute of Oncology of Coimbra (IPOC) are used to test the incorporation of RBF models in a pattern search methods framework. The selected clinical examples were signalized at IPOC as complex cases where proper target coverage and organ sparing, in particular parotid sparing, proved to be difficult to obtain with the typical 7-beam equispaced coplanar treatment plans. The patients' CT sets and delineated structures were exported via Dicom RT to a freeware computational environment for radiotherapy research (see Figure [11](#)). Since the head-and-neck region is a complex area where, e.g., the parotid glands are usually in close proximity to or even overlapping with the target volume, careful selection of the radiation incidence directions can be determinant to obtain a satisfying treatment plan.

The spinal cord and the brainstem are some of the most critical organs at risk (OARs) in the head-and-neck tumor cases. These are serial organs, i.e., organs such that if only one subunit is damaged, the whole organ functionality is compromised. Therefore, if the tolerance dose is exceeded, it may result in functional damage to the whole organ. Thus, it is extremely important not to exceed the tolerance dose prescribed for these type of organs. Other than the spinal cord and the brainstem, the parotid glands are also important OARs. The parotid gland is the largest of the three salivary glands. A common complication due to parotid glands irradiation is xerostomia (the medical term for dry mouth due to lack of saliva). This decreases the quality of life of patients undergoing radiation therapy of head-and-neck, causing difficulties to swallow. The parotids are parallel organs, i.e., if a small volume of the organ is damaged, the rest of the organ functionality may not be affected. Their tolerance dose depends strongly on the fraction of the volume irradiated. Hence, if only a small fraction of the organ is irradiated the tolerance dose is much higher than if a larger fraction is irradiated. Thus, for these parallel structures, the organ mean dose is generally used instead of the maximum dose as an objective for inverse planning optimization.

In general, the head-and-neck region is a complex area to treat with radiotherapy due to the large number of sensitive organs in this region (e.g., eyes, mandible, larynx, oral cavity, etc.). For simplicity, in this study, the OARs used for treatment optimization were limited to the spinal cord, the brainstem and the parotid glands.

The tumor to be treated plus some safety margins is called planning target volume (PTV). For the head-and-neck cases in study it was separated in two parts with different prescribed doses: PTV1 and PTV2. The prescription dose for the target volumes and tolerance doses for the OARs considered in the optimization are presented in Table [11](#).

The parotid glands are in close proximity to or even overlapping with the PTV which helps explaining the difficulty of parotid sparing. Adequate beam directions can help on the overall optimization process and in particular in parotid sparing.

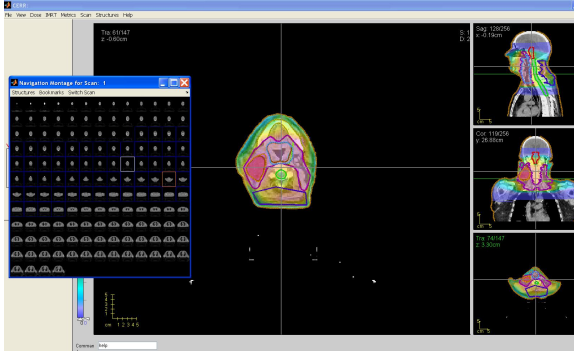


Fig. 1. Illustration of the structures visualized in CERR

Table 1. Prescribed doses for all the structures considered for IMRT optimization

Structure	Mean dose	Max dose	Prescribed dose
Spinal cord	–	45 Gy	–
Brainstem	–	54 Gy	–
Left parotid	26 Gy	–	–
Right parotid	26 Gy	–	–
PTV1	–	–	70.0 Gy
PTV2	–	–	59.4 Gy
Body	–	80 Gy	–

5 Results

Our tests were performed on a 2.66Ghz Intel Core Duo PC with 3 GB RAM. In order to facilitate convenient access, visualization and analysis of patient treatment planning data, as well as dosimetric data input for treatment plan optimization research, the computational tools developed within MATLAB and CERR – computational environment for radiotherapy research [9] are used widely for IMRT treatment planning research. We used CERR 3.2.2 version and MATLAB 7.4.0 (R2007a). The dose was computed using CERR’s pencil beam algorithm (QIB). An automatized procedure for dose computation for each given beam angle set was developed, instead of the traditional dose computation available from IMRTP module accessible from CERR’s menubar. This automatization of the dose computation was essential for integration in our BAO algorithm. To address the convex nonlinear formulation of the FMO problem we used a trust-region-reflective algorithm (*fmincon*) of MATLAB 7.4.0 (R2007a) Optimization Toolbox.

We choose to implement the use of RBFs taking advantage of the availability of an existing pattern search methods framework implementation used successfully by us to tackle the BAO problem [16] – the last version of SID-PSM [6,7].

The spanning set used was the positive spanning set $([e - e I - I])$, with I being the identity matrix and $e = [1 \dots 1]^T$. Each of these directions corresponds to, respectively, the rotation of all incidence directions clockwise, the rotation of all incidence directions counter-clockwise, the rotation of each individual incidence direction clockwise, and the rotation of each individual incidence direction counter-clockwise. The initial mesh-size parameter was set to $\alpha_0 = 4$ since larger values increase the number of function evaluations with no benefits [16]. Since the initial points were integer vectors, all iterates will have integer values as long as the mesh parameter does not become less than one. Therefore, the stopping criteria adopted was the mesh parameter becoming less than one.

The RBFs incorporation into the pattern search methods framework was tested using two clinical examples of retrospective treated cases of head-and-neck tumors at the Portuguese Institute of Oncology of Coimbra (IPOC). A typical head-and-neck treatment plan consists of radiation delivered from five to nine equally spaced coplanar orientations around the patient. Treatment plans with seven equispaced coplanar beams were used at IPOC and are commonly used in practice to treat head-and-neck cases [2]. Therefore, treatment plans of seven coplanar orientations were obtained using our BAO algorithms, denoted *SID-PSM* and *PSM-RBF*, whether the algorithm used was the pattern search framework alone or incorporating RBFs, respectively. These treatment plans were compared with the typical 7-beam equispaced coplanar treatment plans denoted *equi*.

The main goal of the present work is to verify the contribution of the incorporation of RBF models in pattern search methods applied to the optimization of the BAO problem, both in terms of optimal function value found and appropriate search space coverage. Beforehand, we need to decide which RBF is better and should be used for the BAO problem. The CV error of an interpolation model can be a useful and objective tool to decide which RBF model is better. We used the MATLAB code *fminsearch*, an implementation of the Nelder-Mead [13] multidimensional search algorithm, to minimize the CV error $E^{CV}(\theta_1, \dots, \theta_n)$ in Eq. (3) and to find the best model parameters $\theta_1, \dots, \theta_n$. Instead of choosing a priori which RBF should be used, the RBF model used at each iteration is the one that yields the smallest CV error, and consequently the RBF model with the highest prediction accuracy.

The objective function value decrease versus the number of function evaluations required is presented in Fig. 2 to compare the performances of *SID-PSM* and *PSM-RBF*. By simple inspection we conclude that *PSM-RBF* leads to better optimal objective function values compared to *SID-PSM*. The results are presented in terms of number of function evaluations instead of overall computational time since for different dose engines, beamlet optimization methods or even other objective function strategies, the overall computational time may have a totally different magnitude. Dose computation using QIB consumed most of the overall computational time. In average it took two and five hours to run the BAO optimization using the *SID-PSM* and the *PSM-RBF* algorithms, respectively. Our objective is to emphasize the small number of function evaluations required by

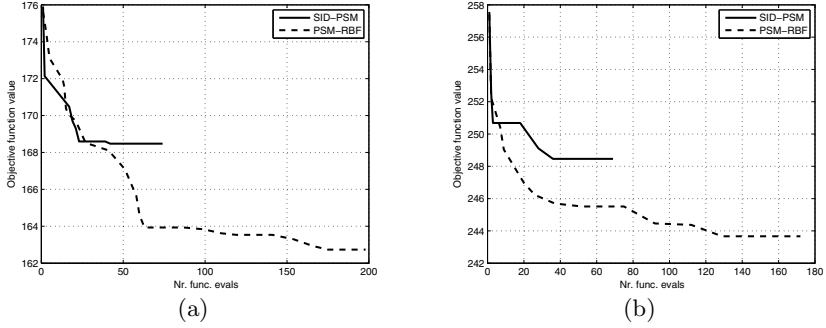


Fig. 2. History of the 7-beam angle optimization process using *SID-PSM* and *PSM-RBF* for cases 1 and 2, [2\(a\)](#) and [2\(b\)](#) respectively

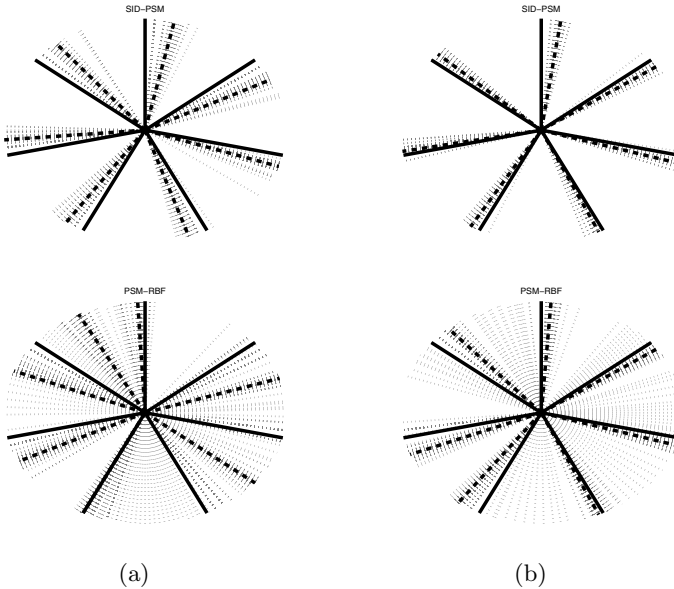


Fig. 3. History of the 7-beam angle optimization process using *SID-PSM* and *PSM-RBF* for cases 1 and 2, [3\(a\)](#) and [3\(b\)](#) respectively. Initial angle configuration, optimal angle configuration and intermediate angle configurations are displayed with solid, dashed and dotted lines, respectively.

pattern search methods, compared to most of the global search methods, heuristics or strategies, even when using RBFs within the search step.

The history of the 7-beam angle optimization process using *SID-PSM* and *PSM-RBF*, in terms of beam directions tested, for each case, is presented in Fig. [3](#). By simple inspection we can verify that the sequence of iterates are better distributed

by amplitude in \mathbb{R}^2 when using *PSM-RBF*, with a more appropriate coverage in amplitude of the whole search space.

Despite the improvement in FMO value, the quality of the results can be perceived considering a variety of metrics. Typically, results are judged by their cumulative dose-volume histogram (DVH). The DVH displays the fraction of a structure's volume that receives at least a given dose. Another metric usually used for plan evaluation is the volume of PTV that receives 95% of the prescribed dose. Typically, 95% of the PTV volume is required. DVH results for the two cases are displayed in Fig. 4. Since parotids are the most difficult organs to spare, and all the treatment plans fulfill the maximum dose requirements for the spinal cord and the brainstem, for clarity, the DVHs only include the targets and the parotids and were split in left and right parotid. The asterisks indicate 95% of PTV volumes versus 95% of the prescribed doses. We can verify that all treatment plans obtained a satisfactory target coverage. However, as expected, the main differences reside in parotid sparing with clear advantage for the optimized treatment plans. In average, *SID-PSM* treatment plans reduced the parotid's mean dose irradiation in 0.8 Gy compared to the *equi* treatment

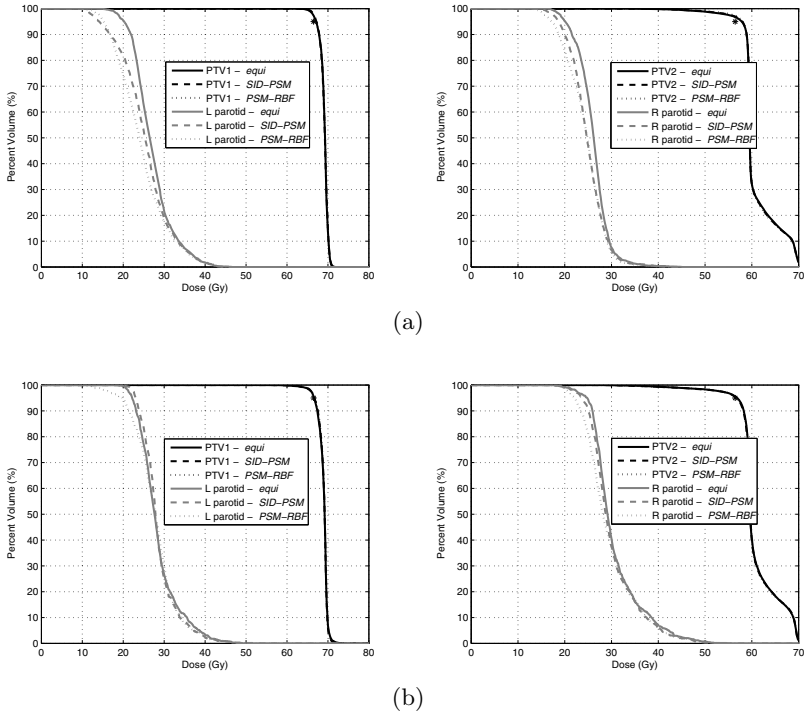


Fig. 4. Cumulative dose volume histogram comparing the results obtained by *equi*, *SID-PSM* and *PSM-RBF* for cases 1 and 2, 4(a) and 4(b) respectively

plans while *PSM-RBF* treatment plans reduced the parotid's mean dose irradiation in 1.5 Gy compared to the *equi* treatment plans. The differences between *SID-PSM* treatment plans and *PSM-RBF* treatment plans, concerning parotid sparing, show a clear advantage for the *PSM-RBF* treatment plans. The results displayed in Fig. 4 confirm the benefits of using the optimized beam directions, in particular using the directions obtained and used in *PSM-RBF* treatment plan.

6 Conclusions

The benefits of a tailored incorporation of RBFs in a pattern search methods framework were tested for the BAO problem using a couple of clinical head-and-neck cases. The BAO problem is a continuous global highly non-convex optimization problem known to be extremely challenging and yet to be solved satisfactorily. Pattern search methods are suited for the BAO problem since they require few function value evaluations and, similarly to other derivative-free optimization methods, have the ability to avoid local entrapment. The pattern search methods approach seems to be similar to neighborhood search approaches in which the neighborhood is constructed using the pattern search method. However, local neighborhood search approaches are only similar to the poll step of the pattern search methods framework. The existence of a search step with the flexibility to use any strategy, method or heuristic, or take advantage of a priori knowledge of the problem at hand, is an advantage that was explored successfully in this work. We have shown that a beam angle set can be locally improved in a continuous manner using pattern search methods. Moreover, it was shown that the incorporation of RBFs in the search step leads to an improvement of the local solution obtained. For numerical approximation of multivariate functions, RBFs can provide excellent interpolants, even if the data points available are unevenly and sporadically distributed. For the retrospective tumor cases tested, our RBFs tailored approach showed a positive influence on the quality of the local minimizer found and a clearly better coverage of the whole search space in amplitude. The improvement of the local solutions in terms of objective function value corresponded, for the head-and-neck cases tested, to high quality treatment plans with good target coverage and with improved organ sparing, in particular better parotid sparing. Moreover, we have to highlight the low number of function evaluations required to obtain locally optimal solutions, which is a major advantage compared to other global heuristics. This advantage should be even more relevant when considering non-coplanar directions since the number of possible directions to consider increase significantly. The efficiency on the number of function value computations is of the utmost importance for the optimization of other general expensive highly non-convex black-box functions.

Acknowledgements. This work was supported by FEDER funds through the COMPETE program and Portuguese funds through FCT under project grant PTDC/EIA-CCO/121450/2010 and by FCT under project grant PEst-C/EEI/UI0308/2011.

References

1. Alberto, P., Nogueira, F., Rocha, H., Vicente, L.N.: Pattern search methods for user-provided points: Application to molecular geometry problems. *SIAM J. Optim.* 14, 1216–1236 (2004)
2. Aleman, D.M., Kumar, A., Ahuja, R.K., Romeijn, H.E., Dempsey, J.F.: Neighborhood search approaches to beam orientation optimization in intensity modulated radiation therapy treatment planning. *J. Global Optim.* 42, 587–607 (2008)
3. Audet, C., Dennis Jr., J.E.: Analysis of generalized pattern search methods. *SIAM J. Optim.* 13, 889–903 (2003)
4. Bortfeld, T., Schlegel, W.: Optimization of beam orientations in radiation therapy: some theoretical considerations. *Phys. Med. Biol.* 38, 291–304 (1993)
5. Craft, D.: Local beam angle optimization with linear programming and gradient search. *Phys. Med. Biol.* 52, 127–135 (2007)
6. Custódio, A.L., Rocha, H., Vicente, L.N.: Incorporating minimum Frobenius norm models in direct search. *Comput. Optim. Appl.* 46, 265–278 (2010)
7. Custódio, A.L., Vicente, L.N.: Using sampling and simplex derivatives in pattern search methods. *SIAM J. Optim.* 18, 537–555 (2007)
8. Davis, C.: Theory of positive linear dependence. *Am. J. Math.* 76, 733–746 (1954)
9. Deasy, J.O., Blanco, A.I., Clark, V.H.: CERR: A Computational Environment for Radiotherapy Research. *Med. Phys.* 30, 979–985 (2003)
10. Ehrgott, M., Holder, A., Reese, J.: Beam selection in radiotherapy design. *Linear Algebra Appl.* 428, 1272–1312 (2008)
11. Lee, E.K., Fox, T., Crocker, I.: Integer programming applied to intensity-modulated radiation therapy treatment planning. *Ann. Oper. Res.* 119, 165–181 (2003)
12. Li, Y., Yao, D., Yao, J., Chen, W.: A particle swarm optimization algorithm for beam angle selection in intensity modulated radiotherapy planning. *Phys. Med. Biol.* 50, 3491–3514 (2005)
13. Nelder, J.A., Mead, R.: A simplex method for function minimization. *Comput. J.* 7, 308–313 (1965)
14. Powell, M.: Radial Basis Function Methods for Interpolation to Functions of Many Variables. *HERMIS: Int. J. Computer Maths & Appl.* 3, 1–23 (2002)
15. Rocha, H.: On the selection of the most adequate radial basis function. *Appl. Math. Model.* 33, 1573–1583 (2009)
16. Rocha, H., Dias, J.M., Ferreira, B.C., Lopes, M.C.: Beam angle optimization using pattern search methods: initial mesh-size considerations. In: *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems* (2012)
17. Romeijn, H.E., Ahuja, R.K., Dempsey, J.F., Kumar, A., Li, J.: A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planing. *Phys. Med. Biol.* 48, 3521–3542 (2003)
18. Tu, J.: Cross-validated Multivariate Metamodeling Methods for Physics-based Computer Simulations. In: *Proceedings of the IMAC-XXI* (2003)
19. Vaz, A.I.F., Vicente, L.N.: A particle swarm pattern search method for bound constrained global optimization. *J. Global Optim.* 39, 197–219 (2007)

On the Complexity of a Mehrotra-Type Predictor-Corrector Algorithm

Ana Paula Teixeira¹ and Regina Almeida²

¹ Department of Mathematics
University of Trás-os-Montes e Alto Douro
P-5000-911 Vila Real, Portugal
CIO. Faculty of Sciences, University of Lisbon, Portugal
ateixeir@utad.pt

² Department of Mathematics
University of Trás-os-Montes e Alto Douro
P-5000-911 Vila Real, Portugal
CIDMA. University of Aveiro, Portugal
ralmeida@utad.pt

Abstract. Based on the good computational results of the feasible version of the Mehrotra's predictor-corrector variant algorithm presented by Bastos and Paixão, in this paper we discuss its complexity. We prove the efficiency of this algorithm by showing its polynomial complexity and, consequently, its Q -linearly convergence.

We start by proving some technical results which are used to discuss the step size estimate of the algorithm.

It is shown that, at each iteration, the step size computed by this Mehrotra's predictor-corrector variant algorithm is bounded below, for $n \geq 2$, by $\frac{1}{200n^4}$; consequently proving that the algorithm has $O(n^4 |\log(\epsilon)|)$ iteration complexity.

Keywords: Linear Programming, predictor-corrector variant, interior-point methods, Mehrotra-type algorithm, polynomial complexity, Q -linear convergence.

1 Introduction

Since Karmarkar's paper [6], many researchers, for example: Freund and Jarre in [4], Güler and Ye in [5], Kojima et al. in [7], Lustig et al. in [8–10], McShane et al. in [11], Salahi et al. in [13], Wright in [14], Ye in [15] and Zhang et al. in [16, 17], devoted their attention to the study of Interior-Point Methods. Predictor-Corrector methods are one of the most studied variants of interior-point methods, being Mehrotra's predictor-corrector algorithm [12] used in several optimization packages.

In general, classical predictor-corrector algorithms perform four line searches by iteration, two of them after obtaining the predictor direction to estimate the duality measure and the other two after computing the final direction. Bastos [2] and Bastos and Paixão [3] presented a new feasible predictor-corrector Linear

Programming variant of Mehrotra's algorithm [12], that just makes two line searches per iteration and so it has the advantage of decreasing the running time of each iteration. The major differences between this new variant and the classical predictor-corrector for Linear Programming are: the predictor direction is computed as in the primal-dual methods, it uses the same duality measure both for the predictor and the corrector directions and no line search is needed to obtain the duality measure. The authors showed that this new version was computationally more efficient than the original one for the class of problems studied in that work.

The complexity of several variants of Mehrotra's algorithm has been studied by many researchers; for example, Zhang and Zhang in [16] and Salahi et al. in [13] proved the polynomial complexity of some Mehrotra-type predictor-corrector variants. Recently, Almeida et al. [1] also analyzed the complexity of another Mehrotra-type predictor-corrector variant algorithm.

Bastos and Paixão in [3], presented specialized versions of an interior-point algorithm for transportation and assignment problems. To validate their algorithm they carried out some computational experiment: sixty instances for the transportation problems, all of them with fifty origins and fifty destinations were generated. Twelve different classes of test problems were considered and five different instances were randomly generated for each one of them. Using these instances, they obtained the correspondent versions of the assignment problems. In this case, four different classes of test problems were considered, each one of them with fifteen instances. The performed computational experience has shown that both the Mehrotra and the new predictor-corrector variant were the best algorithms (among all the ones considered in that paper) for the studied classes of problems. For the transportation problems case, although the Mehrotra variant usually takes a smaller number of iterations and requires less computing time than the new variant, the last one usually presents an inferior gap. For the assignment problems, the Mehrotra variant usually takes a smaller number of iterations but requires more computing time than the new variant.

Based on the good computational results of the new Linear Programming feasible algorithm presented in [3], in this paper we discuss the theoretical efficiency of that algorithm, establishing a complexity bound.

Let us now present some concepts and notation that will be used. Consider $M_{m \times n}(\mathbb{R})$ as the set of the real $m \times n$ matrices. The standard primal-dual pair of Linear Programming problems is

$$\begin{array}{ll}
 \text{(P)} \min_x c^T x & \text{(D)} \max_{u,v} b^T u \\
 \text{s.t. } Ax = b & \text{s.t. } A^T u + v = c \\
 x \geq 0 & v \geq 0
 \end{array}$$

with $A \in M_{m \times n}(\mathbb{R})$, $c, x, v \in \mathbb{R}^n$ and $b, u \in \mathbb{R}^m$. The set of primal-dual feasible solutions of (P) and (D) is given by

$$\mathcal{F} = \{(x, u, v) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n : (x, v) \geq (0, 0), Ax = b, A^T u + v = c\},$$

while its set of primal-dual strictly feasible solutions is

$$\mathcal{F}^\circ = \{(x, u, v) \in \mathcal{F} : (x, v) > (0, 0)\}.$$

The predictor-corrector algorithm in [2, 3] uses the negative infinity norm neighborhood defined by

$$\mathcal{N}_\infty^-(\gamma) = \{(x^k, u^k, v^k) \in \mathcal{F}^\circ : x_i^k v_i^k \geq \gamma \mu_k, i \in \mathcal{I}\},$$

where $\gamma \in]0, 1[$ is a constant, $\mathcal{I} = \{1, 2, \dots, n\}$ and

$$\mu_k = \frac{c^T x^k - b^T u^k}{\theta(n)}, \quad \text{with} \quad \theta(n) = \begin{cases} n^2, & n \leq 5000 \\ n\sqrt{n}, & n > 5000 \end{cases}. \quad (1)$$

The affine direction of this algorithm $(\Delta x^a, \Delta u^a, \Delta v^a)$ is obtained by solving the system

$$\begin{cases} A\Delta X^a e & = 0 \\ A^T \Delta U^a e + \Delta V^a e & = 0 \\ V^k \Delta X^a e + X^k \Delta V^a e = \mu_k e - X^k V^k e \end{cases} \quad (2)$$

and the corrector direction $(\Delta x, \Delta u, \Delta v)$ is obtained by solving the system

$$\begin{cases} A\Delta X e & = 0 \\ A^T \Delta U e + \Delta V e & = 0 \\ V^k \Delta X e + X^k \Delta V e = \mu_k e - \Delta X^a \Delta V^a e - X^k V^k e \end{cases} \quad (3)$$

where e denotes the vector with all components equal to one and $X^k = \text{diag}(x^k)$ is a diagonal matrix with the elements of the vector x^k in the diagonal. Analogously, the matrices $V^k, \Delta X, \Delta U, \Delta V, \Delta X^a, \Delta U^a$ and ΔV^a are obtained by using the elements of the correspondent vectors in the diagonal.

This algorithm can be formalized in *Algorithm 1*.

Throughout this paper, not only are valid the definitions and notation previously mentioned but, we also consider $x \cdot v$ to represent the componentwise product of the vectors x and v , $\|\cdot\|$ to denote the 2-norm of vectors and the index sets

$$\mathcal{I}_+ = \{i \in \mathcal{I} : \Delta x_i^a \Delta v_i^a > 0\}, \quad \mathcal{I}_- = \{i \in \mathcal{I} : \Delta x_i^a \Delta v_i^a < 0\}.$$

For simplicity of notation, we omit the iteration index of the triples that represent the affine and the corrector directions.

In this paper, we use $\gamma = \frac{1}{2}$ in the negative infinity norm neighborhood, i.e., $\mathcal{N}_\infty^-(\frac{1}{2})$ and we consider $\lambda_k = \min\{\alpha_p^k, \alpha_d^k\}$.

Algorithm 1.

Require: $(x^0, u^0, v^0) \in \mathcal{N}_\infty^-(\gamma)$;

Set $k = 0$;

while the termination criteria is not satisfied **do**

(a) Compute μ_k using (1);

(b) Obtain the affine direction $(\Delta x^a, \Delta u^a, \Delta v^a)$ by solving (2);

(c) Obtain the corrector direction $(\Delta x, \Delta u, \Delta v)$ by solving (3);

(d) Compute primal and dual step sizes, respectively,

$$\alpha_p^k = \max\{\alpha > 0 : x^k + \alpha \Delta x \geq 0\}, \quad \alpha_d^k = \max\{\alpha > 0 : v^k + \alpha \Delta v \geq 0\};$$

(e) Compute

$$x^{k+1} = x^k + 0.9995\alpha_p^k \Delta x, \quad (u^{k+1}, v^{k+1}) = (u^k, v^k) + 0.9995\alpha_d^k (\Delta u, \Delta v);$$

(f) Set $k = k + 1$;

end while

2 Technical Results

In this section we prove some technical results which are used, in Section 3, in order to discuss the step size estimate of *Algorithm 1*. Throughout this section Lemmas 5.1 and 5.3 of [14] are frequently used. For simplicity, we transcribe these results in the Appendix.

Lemma 1. *Let $(\Delta x^a, \Delta u^a, \Delta v^a)$ be the affine direction of Algorithm 1. Then, for all $i \in \mathcal{I}_+$,*

$$\Delta x_i^a \Delta v_i^a \leq \frac{1}{4} \left(1 - \frac{\mu_k}{x_i^k v_i^k} \right)^2 x_i^k v_i^k.$$

Proof. Using the third condition of (2), we have

$$\frac{\Delta x_i^a}{x_i^k} + \frac{\Delta v_i^a}{v_i^k} = \frac{\mu_k}{x_i^k v_i^k} - 1. \quad (4)$$

Since

$$0 \leq \left(\frac{\Delta x_i^a}{x_i^k} - \frac{\Delta v_i^a}{v_i^k} \right)^2 = \left(\frac{\Delta x_i^a}{x_i^k} \right)^2 + \left(\frac{\Delta v_i^a}{v_i^k} \right)^2 - 2 \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k},$$

then

$$\left(\frac{\Delta x_i^a}{x_i^k} \right)^2 + \left(\frac{\Delta v_i^a}{v_i^k} \right)^2 \geq 2 \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k}. \quad (5)$$

Therefore, using

$$\left(\frac{\Delta x_i^a}{x_i^k} + \frac{\Delta v_i^a}{v_i^k}\right)^2 = \left(\frac{\Delta x_i^a}{x_i^k}\right)^2 + \left(\frac{\Delta v_i^a}{v_i^k}\right)^2 + 2\frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k},$$

equality (4) and inequality (5), we have

$$\left(\frac{\mu_k}{x_i^k v_i^k} - 1\right)^2 \geq 4\frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k}.$$

Lemma 2. Let $(\Delta x^a, \Delta u^a, \Delta v^a)$ be the affine direction of Algorithm 1. Then,

$$\sum_{i \in \mathcal{I}_+} \Delta x_i^a \Delta v_i^a = \sum_{i \in \mathcal{I}_-} |\Delta x_i^a \Delta v_i^a| \leq \frac{\mu_k}{4} \theta(n).$$

Proof. Since $\Delta x^{aT} \Delta v^a = 0$, from Lemma 5.1 of [14], p.87], then

$$0 = \sum_{i \in \mathcal{I}} \Delta x_i^a \Delta v_i^a = \sum_{i \in \mathcal{I}_+} \Delta x_i^a \Delta v_i^a + \sum_{i \in \mathcal{I}_-} \Delta x_i^a \Delta v_i^a.$$

Therefore,

$$\sum_{i \in \mathcal{I}_+} \Delta x_i^a \Delta v_i^a = \sum_{i \in \mathcal{I}_-} |\Delta x_i^a \Delta v_i^a|.$$

Using Lemma 1, we have

$$\begin{aligned} \sum_{i \in \mathcal{I}_+} \Delta x_i^a \Delta v_i^a &\leq \frac{1}{4} \sum_{i \in \mathcal{I}_+} \left(\frac{\mu_k}{x_i^k v_i^k} - 1\right)^2 x_i^k v_i^k \\ &= \frac{1}{4} \sum_{i \in \mathcal{I}_+} \left(x_i^k v_i^k + \frac{\mu_k^2}{x_i^k v_i^k} - 2\mu_k\right) \\ &\leq \frac{1}{4} \sum_{i \in \mathcal{I}_+} \left(x_i^k v_i^k + \frac{2\mu_k^2}{\mu_k} - 2\mu_k\right) \\ &\leq \frac{1}{4} \left(\mu_k \theta(n) + \sum_{i \in \mathcal{I}_+} (2\mu_k - 2\mu_k)\right) \\ &\leq \frac{\mu_k}{4} \theta(n). \end{aligned}$$

Using the above technical lemmas, we obtain the following proposition which gives an upper bound estimate of the 2-norm of the inner product of the vectors Δv and Δx .

Proposition 1. *Let $(\Delta x, \Delta u, \Delta v)$ be the solution of (3) and $(x^k, u^k, v^k) \in \mathcal{N}_\infty^-(\frac{1}{2})$ be the current iterate. Then*

$$\|\Delta v \cdot \Delta x\| \leq 4n^4 \mu_k.$$

Proof. Multiplying the third equation of (3) by $(X^k V^k)^{-\frac{1}{2}}$, we obtain

$$\begin{aligned} (X^k)^{-\frac{1}{2}} (V^k)^{\frac{1}{2}} \Delta x + (V^k)^{-\frac{1}{2}} (X^k)^{\frac{1}{2}} \Delta v &= \\ &= (X^k V^k)^{-\frac{1}{2}} (\mu_k e - \Delta X^a \Delta V^a e - X^k V^k e). \end{aligned} \quad (6)$$

Taking the first term of (6) and

$$D = (V^k)^{-\frac{1}{2}} (X^k)^{\frac{1}{2}},$$

we have

$$(X^k)^{-\frac{1}{2}} (V^k)^{\frac{1}{2}} \Delta x + (V^k)^{-\frac{1}{2}} (X^k)^{\frac{1}{2}} \Delta v = D^{-1} \Delta x + D \Delta v.$$

Let us denote

$$w_1 = D^{-1} \Delta x, \quad w_2 = D \Delta v, \quad W_1 = \text{diag}(w_1), \quad W_2 = \text{diag}(w_2).$$

Using Lemma 5.3 of [14, p.88] follows

$$\begin{aligned} \|W_1 W_2 e\| &= \|\Delta v \cdot \Delta x\| \\ &\leq 2^{-\frac{3}{2}} \|w_1 + w_2\|^2 \\ &= 2^{-\frac{3}{2}} \left\| (X^k V^k)^{-\frac{1}{2}} (\mu_k e - \Delta X^a \Delta V^a e - X^k V^k e) \right\|^2 \\ &= 2^{-\frac{3}{2}} \left(\mu_k^2 \sum_{i \in \mathcal{I}} \frac{1}{x_i^k v_i^k} + \theta(n) \mu_k + \sum_{i \in \mathcal{I}} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k} - 2n \mu_k \right. \\ &\quad \left. - 2\mu_k \sum_{i \in \mathcal{I}} \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k} + 2 \sum_{i \in \mathcal{I}} \Delta x_i^a \Delta v_i^a \right). \end{aligned}$$

For $x_i^k v_i^k \geq \frac{1}{2} \mu_k$, we have

$$\mu_k^2 \sum_{i \in \mathcal{I}} \frac{1}{x_i^k v_i^k} \leq 2\mu_k n.$$

Now, let us analyze the term

$$\sum_{i \in \mathcal{I}} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k} = \sum_{i \in \mathcal{I}_+} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k} + \sum_{i \in \mathcal{I}_-} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k}. \quad (7)$$

For the case $i \in \mathcal{I}_+$, using Lemma [1](#) and the fact that (x^k, u^k, v^k) belongs to $\mathcal{N}_\infty^-(\frac{1}{2})$, we get

$$\begin{aligned} \sum_{i \in \mathcal{I}_+} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k} &\leq \sum_{i \in \mathcal{I}_+} \frac{1}{16} \left(\frac{\mu_k}{x_i^k v_i^k} - 1 \right)^4 x_i^k v_i^k \\ &= \frac{1}{16} \sum_{i \in \mathcal{I}_+} \left(\frac{\mu_k^4}{(x_i^k v_i^k)^3} - 4 \frac{\mu_k^3}{(x_i^k v_i^k)^2} + 6 \frac{\mu_k^2}{x_i^k v_i^k} - 4\mu_k + x_i^k v_i^k \right) \\ &\leq \frac{1}{16} \sum_{i \in \mathcal{I}_+} \left(8\mu_k - 4 \frac{\mu_k^3}{(x_i^k v_i^k)^2} + 12\mu_k - 4\mu_k + x_i^k v_i^k \right) \\ &\leq \frac{1}{16} \sum_{i \in \mathcal{I}_+} (16\mu_k + x_i^k v_i^k) \\ &\leq \mu_k \left(n + \frac{\theta(n)}{16} \right). \end{aligned}$$

For $i \in \mathcal{I}_-$, using Lemma [2](#), we have

$$\begin{aligned} \sum_{i \in \mathcal{I}_-} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k} &\leq \frac{2}{\mu_k} \sum_{i \in \mathcal{I}_-} (\Delta x_i^a \Delta v_i^a)^2 \\ &\leq \frac{2}{\mu_k} \left(\sum_{i \in \mathcal{I}_-} |\Delta x_i^a \Delta v_i^a| \right)^2 \\ &\leq \frac{2}{\mu_k} \left(\frac{\theta(n)}{4} \mu_k \right)^2 \\ &\leq \frac{\mu_k}{8} (\theta(n))^2. \end{aligned}$$

Therefore, we have for [7](#)

$$\sum_{i \in \mathcal{I}} \frac{(\Delta x_i^a \Delta v_i^a)^2}{x_i^k v_i^k} \leq \left(n + \frac{\theta(n)}{16} + \frac{(\theta(n))^2}{8} \right) \mu_k.$$

Using Lemma 2, the fact that (x^k, u^k, v^k) belongs to $\mathcal{N}_\infty^-(\frac{1}{2})$ and the definition of \mathcal{I}_- and \mathcal{I}_+ , we obtain

$$\begin{aligned}
-2\mu_k \sum_{i \in \mathcal{I}} \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k} &\leq -2\mu_k \sum_{i \in \mathcal{I}_+} \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k} - 2\mu_k \sum_{i \in \mathcal{I}_-} \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k} \\
&\leq -2\mu_k \sum_{i \in \mathcal{I}_-} \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k} \\
&\leq 2\mu_k \sum_{i \in \mathcal{I}_-} 2 \frac{|\Delta x_i^a \Delta v_i^a|}{\mu_k} \\
&\leq 4 \sum_{i \in \mathcal{I}_-} |\Delta x_i^a \Delta v_i^a| \\
&\leq 4\mu_k \frac{1}{4} \theta(n) \\
&= \mu_k \theta(n).
\end{aligned}$$

Since from Lemma 5.1 of [14, p.87] we get

$$\Delta x^{aT} \Delta v^a = 0,$$

then it follows

$$\|\Delta v \cdot \Delta x\| \leq \left(\frac{33}{16} \theta(n) + \frac{1}{8} (\theta(n))^2 + n \right) \mu_k.$$

In conclusion, estimating $\theta(n)$ by n^2 , we have

$$\|\Delta v \cdot \Delta x\| \leq 4n^4 \mu_k.$$

Analogously to Lemma 5.1 of [14, p.87], by a straightforward calculation, we have the following result.

Lemma 3. *Let μ_k satisfy (1). Then*

$$\mu_{k+1} = \left(1 - \lambda_k \left(1 - \frac{n}{\theta(n)} \right) \right) \mu_k.$$

Proof. Since

$$c^T x^{k+1} - b^T u^{k+1} = (v^{k+1})^T x^{k+1},$$

where $x^{k+1} = x^k + \lambda_k \Delta x$ and $v^{k+1} = v^k + \lambda_k \Delta v$, we get

$$\begin{aligned}
c^T x^{k+1} - b^T u^{k+1} &= (v^{k+1})^T x^{k+1} \\
&= (v^k + \lambda_k \Delta v)^T (x^k + \lambda_k \Delta x) \\
&= v^{kT} x^k + \lambda_k \left(\Delta v^T x^k + v^{kT} \Delta x \right) + \lambda_k^2 \Delta v^T \Delta x
\end{aligned}$$

which is equivalent to

$$c^T x^{k+1} - b^T u^{k+1} = v^{kT} x^k + \lambda_k \left(x^{kT} \Delta v + v^{kT} \Delta x \right) + \lambda_k^2 \Delta v^T \Delta x. \quad (8)$$

Using the second equation of (3), $\Delta u^T A = -(\Delta v)^T$, we obtain

$$\Delta u^T A \Delta x = -(\Delta v)^T \Delta x. \quad (9)$$

Therefore, from the first equation of (3) and (9), we get

$$\Delta v^T \Delta x = 0. \quad (10)$$

Introducing the third equation of (3), we have

$$\begin{aligned}
x^{kT} \Delta v + v^{kT} \Delta x &= [V^k \Delta X e + X^k \Delta V e]^T e \\
&= [\mu_k e - \Delta X^a \Delta V^a e - X^k V^k e]^T e \\
&= n\mu_k - \Delta v^{aT} \Delta x^a - v^{kT} x^k.
\end{aligned} \quad (11)$$

Applying (10) and (11) to (8), we obtain

$$c^T x^{k+1} - b^T u^{k+1} = v^{kT} x^k + \lambda_k \left(n\mu_k - \Delta v^{aT} \Delta x^a - v^{kT} x^k \right). \quad (12)$$

Due to $\Delta v^{aT} \Delta x^a = 0$ and

$$v^{kT} x^k = c^T x^k - b^T u^k,$$

equation (12) can be rewritten as

$$c^T x^{k+1} - b^T u^{k+1} = (1 - \lambda_k) (c^T x^k - b^T u^k) + \lambda_k n\mu_k.$$

Using (1), we obtain the desired relation

$$\mu_{k+1} = \left(1 - \lambda_k + \frac{n\lambda_k}{\theta(n)} \right) \mu_k.$$

3 Polynomial Complexity

In this section we prove that *Algorithm 1* is Q -linearly convergent. More precisely, we prove that at each iteration the step size computed by this algorithm is bounded below by $\frac{1}{200n^4}$, for $n \geq 2$. Consequently, we prove that *Algorithm 1* has $O(n^4 |\log(\epsilon)|)$ iteration complexity.

In the next result we discuss the step size estimate of the algorithm to establish its worst case iteration complexity, in order to analyze the asymptotic behavior of *Algorithm 1*.

Theorem 1. *Suppose that the current iterate (x^k, u^k, v^k) belongs to $\mathcal{N}_\infty^- \left(\frac{1}{2}\right)$. Let the solution of (13) be $(\Delta x, \Delta u, \Delta v)$. Then the maximum step size λ_k , that keeps $(x^{k+1}, u^{k+1}, v^{k+1})$ in $\mathcal{N}_\infty^- \left(\frac{1}{2}\right)$, satisfies*

$$\lambda_k \geq \frac{1}{200n^4}, \quad k \geq 0, \quad n \geq 2.$$

Proof. In order to find the maximum nonnegative λ_k for which

$$x_i^{k+1} v_i^{k+1} \geq \frac{1}{2} \mu_{k+1},$$

with $i \in \mathcal{I}$, we define

$$t = \max_{i \in \mathcal{I}_+} \left\{ \frac{\Delta x_i^a \Delta v_i^a}{x_i^k v_i^k} \left(\frac{\mu_k}{x_i^k v_i^k} - 1 \right)^{-2} \right\}. \quad (13)$$

Since $\Delta x^{aT} \Delta v^a = 0$, we have $\mathcal{I}_+ \neq \emptyset$. Let $i \in \mathcal{I}_+$, then

$$\begin{aligned} x_i^{k+1} v_i^{k+1} &= x_i^k v_i^k + \lambda_k (\mu_k - x_i^k v_i^k - \Delta x_i^a \Delta v_i^a) + \lambda_k^2 \Delta x_i \Delta v_i \\ &= (1 - \lambda_k) x_i^k v_i^k + \lambda_k \mu_k - \lambda_k \Delta x_i^a \Delta v_i^a + \lambda_k^2 \Delta x_i \Delta v_i. \end{aligned}$$

Applying Proposition 1, (13) and

$$\left(\frac{\mu_k}{x_i^k v_i^k} - 1 \right)^2 \leq 1 + \frac{\mu_k^2}{(x_i^k v_i^k)^2}$$

we get

$$\begin{aligned} x_i^{k+1} v_i^{k+1} &\geq (1 - \lambda_k) x_i^k v_i^k + \lambda_k \mu_k - \lambda_k t x_i^k v_i^k \left(1 + \frac{\mu_k^2}{(x_i^k v_i^k)^2} \right) + \lambda_k^2 \Delta x_i \Delta v_i \\ &\geq (1 - \lambda_k (1 + t)) x_i^k v_i^k + (1 - 2t) \lambda_k \mu_k - \lambda_k^2 \Delta x_i \Delta v_i \\ &\geq (1 - \lambda_k (1 + t)) x_i^k v_i^k + (1 - 2t) \lambda_k \mu_k - \lambda_k^2 (4n^4) \mu_k. \end{aligned}$$

Since each iterate must belong to $\mathcal{N}_\infty^-\left(\frac{1}{2}\right)$, if we take $1 - \lambda_k(1+t) > 0$, we obtain

$$x_i^{k+1}v_i^{k+1} \geq (1 - \lambda_k(1+t))\frac{1}{2}\mu_k + (1-2t)\lambda_k\mu_k - 4\lambda_k^2n^4\mu_k.$$

By Lemma [1](#) we get $t \leq \frac{1}{4}$ and, consequently, $\frac{1}{1+t} \geq \frac{4}{5}$. Therefore, we take $\lambda_k \in [0, \frac{4}{5}]$. In order to guaranty that the next iterate belongs to $\mathcal{N}_\infty^-\left(\frac{1}{2}\right)$, we consider

$$(1 - \lambda_k(1+t))\frac{1}{2}\mu_k + (1-2t)\lambda_k\mu_k - 4\lambda_k^2n^4\mu_k \geq \frac{1}{2}\mu_{k+1}.$$

Using Lemma [3](#), we have

$$(1 - \lambda_k(1+t))\frac{1}{2}\mu_k + (1-2t)\lambda_k\mu_k - 4\lambda_k^2n^4\mu_k \geq \frac{1}{2}\left(1 - \lambda_k\left(1 - \frac{n}{\theta(n)}\right)\right)\mu_k,$$

which is equivalent to

$$\lambda_k\left(2 - 5t - \frac{n}{\theta(n)}\right)\frac{\mu_k}{2} \geq 4\lambda_k^2n^4\mu_k. \tag{14}$$

Using Lemma [1](#) and the definition of $\theta(n)$ presented in [\(11\)](#), we obtain, for $n \geq 2$,

$$\left(2 - 5t - \frac{n}{\theta(n)}\right) \geq 2 - \frac{5}{4} - \frac{1}{\sqrt{n}} \geq \frac{3\sqrt{n} - 4}{4\sqrt{n}} \geq \frac{1}{25}. \tag{15}$$

From [\(14\)](#) and [\(15\)](#), we conclude, for $n \geq 2$,

$$\lambda_k \geq \min\left\{\frac{4}{5}, \frac{1}{200n^4}\right\} = \frac{1}{200n^4}.$$

The following theorem gives an upper bound for the number of iterations in which *Algorithm 1* stops.

Theorem 2. *Let $\epsilon \in]0, 1[$. *Algorithm 1* stops after at most $O(n^4|\log(\epsilon)|)$ iterations with a solution for which $x^T v \leq \epsilon$.*

Proof. Considering the definition of μ_k presented in [\(11\)](#) and Lemma [3](#), we obtain

$$\mu_{k+1} = \left(1 - \lambda_k + \frac{n\lambda_k}{\theta(n)}\right)\mu_k \leq \left(1 - \lambda_k\left(1 - \frac{1}{\sqrt{n}}\right)\right)\mu_k.$$

Using Theorem [1](#), we have

$$\mu_{k+1} \leq \left(1 - \lambda_k\left(1 - \frac{1}{\sqrt{n}}\right)\right)\mu_k \leq \left(1 - \frac{1,45 \times 10^{-3}}{n^4}\right)\mu_k.$$

Applying Theorem 3.2 of [\[14\]](#), p. 61] (see Appendix) we complete the proof.

Acknowledgments. A. Teixeira is financially supported by the Unit CIO-Operations Research Center (MATH-LVT-Lisboa-152), based at the Faculty of Sciences, University of Lisbon, financed by the Portuguese Foundation for Science and Technology (FCT), within project PEst-OE/MAT/UI0152. R. Almeida was financially supported by *FEDER* funds through *COMPETE*-Operational Programme Factors of Competitiveness and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* (University of Aveiro) and the Portuguese Foundation for Science and Technology (FCT), within project PEst-C/MAT/UI4106/2011 with *COMPETE* number FCOMP-01-0124-FEDER-022690.

References

1. Almeida, R., Bastos, F., Teixeira, A.: On polynomiality of a predictor-corrector variant algorithm. In: International Conference on Numerical Analysis and Applied Mathematica 2010 - ICNAAM 2010. AIP Conference Proceedings, vol. 2, pp. 959–963. Springer, New York (2010)
2. Bastos, F.: Problemas de transporte e métodos de ponto interior, Dissertação de Doutoramento, Universidade Nova de Lisboa, Lisboa (1994)
3. Bastos, F., Paixão, J.: Interior-point approaches to the transportation and assignment problems on microcomputers. *Investigação Operacional* 13(1), 3–15 (1993)
4. Freund, R.W., Jarre, F.: A QMR-based interior-point algorithm for solving linear programs, Numerical Analysis Manuscript, 94-19. AT&T Bell Laboratories, Murray Hill, N.J. (1994)
5. Güler, O., Ye, Y.: Convergence behavior of interior-point algorithms. *Mathematical Programming* 60, 215–228 (1993)
6. Karmarkar, N.K.: A new polynomial-time algorithm for linear programming. *Combinatorica* 4, 373–395 (1984)
7. Kojima, M., Megiddo, N., Mizuno, S.: A primal-dual infeasible-interior point algorithm for linear programming. *Mathematical Programming, Series A* 61, 261–280 (1993)
8. Lustig, I.J., Marsten, R.E., Shanno, D.F.: Computational experience with a primal-dual interior point method for linear programming. *Linear Algebra and Its Applications* 152, 191–222 (1991)
9. Lustig, I.J., Marsten, R.E., Shanno, D.F.: On implementing Mehrotra’s predictor-corrector interior-point method for linear programming. *SIAM Journal on Optimization* 2(3), 435–449 (1992)
10. Lustig, I.J., Marsten, R.E., Shanno, D.F.: Interior point method for linear programming: Computational state of the art. *ORSA Journal on Computing* 6, 1–14 (1994)
11. McShane, K.A., Monma, C.L., Shanno, D.F.: An implementation of a primal-dual interior point method for linear programming. *ORSA Journal on Computing* 1, 70–83 (1989)
12. Mehrotra, S.: On the implementation of a primal-dual interior point method. *SIAM J. Optimization* 2, 575–601 (1992)
13. Salahi, M., Peng, J., Terlaky, T.: On Mehrotra-Type Predictor-Corrector Algorithms. *SIAM Journal on Optimization* 18(4), 1377–1397 (2007)
14. Wright, S.J.: *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia (1997)

15. Ye, Y.: Interior Point Algorithms, Theory and Analysis. John Wiley and Sons, Chichester (1997)
16. Zhang, Y., Zhang, D.: On polynomiality of the Mehrotra-type predictor-corrector interior-point algorithms. *Mathematical Programming* 68, 303–318 (1995)
17. Zhang, Y., Tapia, R.A., Dennis, J.E.: On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms. *SIAM J. Optimization* 2, 304–324 (1992)

Appendix

In this appendix we transcribe Theorem 3.2 and Lemmas 5.1 and 5.3 of [14].

Theorem 3.2 [14, p.61]. *Let $\epsilon \in]0, 1[$ be given. Suppose that our algorithm for solving the Karush-Kuhn-Tucker conditions associated with the standard primal-dual pair of Linear Programming problems generates a sequence of iterates that satisfies*

$$\mu_{k+1} \leq \left(1 - \frac{\delta}{n^\omega}\right) \mu_k, \quad k = 0, 1, \dots$$

for some positive constants δ and ω . Suppose too that the starting point (x^0, u^0, v^0) satisfies

$$\mu_0 \leq \frac{1}{\epsilon^\kappa}$$

for some positive constant κ . Then there exists an index K with

$$K = O(n^\omega |\log(\epsilon)|)$$

such that

$$\mu_k \leq \epsilon \quad \text{for all } k \geq K.$$

Lemma 5.1. [14, p.87] *Let the step $(\Delta x^a, \Delta u^a, \Delta v^a)$ be defined by [2]. Then*

$$\Delta v^{aT} \Delta x^a = 0.$$

Lemma 5.3. [14, p.88] *Let u and v be any two vectors in \mathbb{R}^n with $u^T v \geq 0$. Then*

$$\|UVe\| \leq 2^{-\frac{3}{2}} \|u + v\|^2,$$

where $U = \text{diag}(u_1, \dots, u_n)$ and $V = \text{diag}(v_1, \dots, v_n)$.

Design of Wood Biomass Supply Chains

Tiago Costa Gomes¹, Filipe Pereira e Alvelos², and Maria Sameiro Carvalho²

¹ Algoritmi Research Center, University of Minho, Portugal
tiago.gomes@dps.uminho.pt

² Department of Production and Systems, University of Minho, Portugal
{falvelos,sameiro}@dps.uminho.pt

Abstract. The purpose of this paper is to propose a mathematical programming approach to minimize the total cost in a biomass supply chain. A company that collects material from forests, transforms it into chipped product, stores and delivers it to its customers is considered. For tackling all the aspects of the supply chain management, a mixed integer programming model that supports tactical and operational decisions was developed and optimized using a general purpose solver. The model was implemented in C++ and several computational tests have been performed.

Keywords: Supply chain, Biomass, Optimization, Mixed integer programming.

1 Introduction

Recent trends in energy all over the world highlight the need of using efficient approaches in the management of energy resources, not only by changing energy consumption and emission habits but also, by using new technologies, tools and methodologies. New challenges have to be addressed and new strategies have to be devised to have a cost-effective use of energy resources, in particular, renewable resources.

In order to change the current situation, a large number of countries are focusing on increasing the consumption of renewable and environmentally friendly energies to allow a reduction of the dependence on fossil energy and emissions of greenhouse gases [1-4]. In [1], the European Commission defined ambitious energy objectives for 2020: *“to reduce greenhouse gas emissions by 20%, to increase the share of renewable energy to 20% and to make a 20% improvement in energy efficiency”*.

There are several types of renewable energy sources such as the wind energy, hydropower, solar energy, geothermal energy, wave energy and bioenergy. The use of biomass has increased in recent years and its share is expected to increase in the near future. Biomass is a general term that incorporates biological material such as wood (harvested wood, wood waste, etc) or biodegradable wastes. The wood biomass can be included in various economic activities, such as: in the production of derivatives timber products, for heating, for electricity generation, or in the transport sector.

According to [5,6] “woody biomass is biomass from trees, bushes and shrubs. This definition includes forest and plantation wood, wood processing industry by-products and residues, and used wood”. The biomass action plan of European Union is defined in [7] with the aim to encourage Member States to implement measures to promote biomass industry by establishing possible guidelines for developing national biomass action plans.

In general, wood biomass is considered a renewable and clean energy with a carbon neutral cycle: the trees capture CO_2 from the atmosphere; forest fuel is fired and CO_2 is released back into the atmosphere. This cycle repeats successively. For this reason, the burning of biomass may not cause more emissions of greenhouse gases. In [8] the need to balance deforestation and reforestation activities to balance between emissions and inventory of CO_2 is stressed.

Biomass energy has been considered a successful alternative to fossil fuels but its competitiveness is highly dependent on the efficiency of its supply chain. In fact, unlike fossil fuels, biomass is distributed over widespread areas and therefore, the location of wood terminals, the position of facilities to pre-process products; inventory and storage decisions, the location of bioenergy plants/consumers and transportation operations (both on-farm/forest and on road transportation) become critical factors in designing and planning new biomass systems. To be profitable the business of biomass it is necessary to optimize the supply chain, from the collection of raw materials until the end-user. Transport costs account for about 50% of total costs, in some cases can reach 65% [9,10]. The correct definition of the links in the supply chain and the optimal planning of the flow of materials are crucial in order to meet the demand at the desired time while minimizing total costs.

Additionally, other characteristics of biomass systems add extra complexity to the design and planning of its supply chain [11]: both supply and demand are seasonal, supply product is time and weather sensitive; products have variable moisture content and low bulk density. An efficient biomass supply chain must balance costs, customer service and sustainability to achieve high levels of performance and decision support models are required to analyze and optimize such a complex system.

In this work, a mixed integer programming model, capable of incorporating, tactical and operational decisions in a biomass supply chain, is developed. While minimizing total system costs, this approach tries to incorporate some of the specificities and complexities of these logistics systems.

This article is divided into five sections. The subsequent section presents a review of existing models and describes, in detail, the problem addressed. In the third section, the mathematical model is presented and discussed and in the fourth section, computational results are presented and analyzed. Finally, in the last section, some conclusions are drawn and some guidelines for future work are proposed.

2 Problem Description

In the last decades several approaches have been proposed in the literature to study and analyze general purpose supply chains. More recent models proposed in the

literature, based on mathematical programming models, allow the integration of decisions from different levels (e.g. strategic and tactical) demonstrating the potential savings of integrated approaches in the design of global logistics systems. In some cases these approaches are general purpose models, difficult to apply to very complex and specific supply chains. Others approaches are applied studies difficult to be used in a wide variety of systems.

In terms of biomass supply chains, the number of scientific publications has been increasing more recently supporting the greater concern on renewable energy resources.

In [12] the special characteristics of waste biomass supply chains are presented and it is stressed that logistics and supply chain management are areas of critical importance for the successful utilization of energetic biomass. Furthermore, in that work, authors stress the need of “comprehensive waste biomass supply chain approaches”. Although, focusing in the waste biomass problem, the questions raised have large similarities with wood biomass, in particular in what concerns the decision making process: i) at strategic decision level, the network configuration problem – sourcing (selection of collection sites, selection of the types of wood biomass to collect, etc), location and capacity of terminals, storage and production facilities – required to balance seasonal variations of demand and supply; sustainability of the supply chain; ii) at tactical decision level: decisions concerning pre-treatment (technology, timing and place of pre-treatment – particularly relevant in the wood biomass supply chain is the definition of the best timing for pre-treatment, namely the chipping process, since it has a large impact on transportation decisions and costs) - inventory management; fleet and transportation management and iii) at operational level: operations planning (such as chipping operations; inventory control and vehicle scheduling operations);

From the comprehensive analysis undertaken by these authors it is clear that the majority of published work focus only a particular aspect of the decision making process and that integrated models to incorporate strategic, tactic and/or operational decisions are still required to allow a systemic understanding of biomass systems.

A mixed integer programming model of the biomass supply chain is proposed [13] in which integrates both strategic and tactical decisions. The problem addressed consists of a multi-period (monthly based), multi-product with capacitated facilities defined over a two-echelon structure: suppliers (forest areas, saw mill, external suppliers), terminals (for storage) and customers (heating plants) with known demand for forest fuel. Decisions concerning the choice of suppliers and terminals, timing and location of chipping, storage quantities and flows among facilities are to be addressed. The mathematical model is solved using a heuristic approach and allows evaluation of alternative strategies and scenarios providing support in the decision making process.

The model proposed in this research can be seen as an extension of model of [13]. The main difference concerns the inclusion of the operational decision level, in addition to the tactical level and a daily based time scale is used to incorporate transportation operations (e.g. how many vehicles are required) and processing operations (allocation of resources to collecting and chipping operations). Additionally, as transportation costs constitute one of the most important fractions of

the total costs, a more realistic approach is adopted: transportation costs are not a linear function of units transported, but a staircase function where each step corresponds to a used vehicle which allows a closer approximation to real costs.

Next, the problem analyzed will be described in more detail.

A company that collects material from forests and distributes chipped product for heating purposes is considered. Raw product consists of harvests material from forest areas (usually forest residues such as branches, bushes and shrubs, dead trees, etc.) obtained from several forest areas. Raw material from forest areas requires pre-treatment: wood residues have to be transformed into chipped product (wood chips) “sub rectangular shaped pieces with a defined particle size produced by mechanical treatment with, usually, knives” [6]. This pre-treatment process can be undertaken either in forest areas or in facilities with chipping capacity. Chipped products has to be moved to customers either directly or using intermediate facilities. In this sector of activity, supply and demand are seasonal, varying throughout the year due to climatic variations. Because of this seasonality the use of inventories is of most importance for a good level of service.

Figure 1 represents a general scheme of the supply chain, where the rectangles represent the supply chain elements; the arrows represent the operations and the circles the availability of certain material in each period.

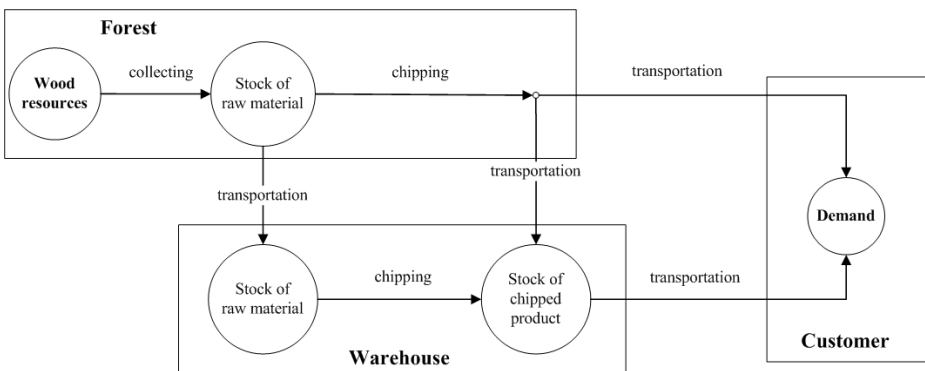


Fig. 1. Scheme of supply chain

The collection of wood resources is carried out in forest areas and raw materials can be stored or transformed in chipped product before they are transported either to a customer or to a warehouse (storage facility). Intermediate warehouses serve to improve supply and demand coordination, particularly relevant when seasonality of both supply and demand are an important issue and to improve the management of chipped and transportation operations. Stores can accommodate raw material, as well as chipped product. They can also be used to undertake chip operations, through a fixed chipper or a mobile chipper. Finally, customers receive products already processed in order to meet their demands.

Supply and demand quantities are known, resources and capacities are limited and, an important aspect of the problem is the optimal management of equipments and

human resources involved in the tasks of collecting, chipping, transporting and storing the biomass. It is assumed that operations duration is one day.

For tackling all these aspects of the supply chain design, a mixed integer programming model that supports tactical and operational decisions was developed. The model allows the minimization of total costs involved.

For a given planning horizon divided in time periods, the model addresses the following decisions:

- selection of collection sites (whether or not use a supplier/forest area);
- location of storage facilities (whether or not use an intermediate warehouse);
- type and quantity of different type of products to be supplied;
- pre-treatment decisions (whether or not use pre-treatment at forest; location of fixed and mobile chippers);
- storage decision (quantities to be stored at all time periods, in the different warehouses);
- transportation flows between different links in the chain.

In the following section the mathematical programming model will be described.

3 Mathematical Programming Model

3.1 Decision Variables and Constraints

A set I of forest areas, a set W of potential warehouses, a set J of customers, a set P of products, and a set T of time periods are considered.

The collection of raw materials at forest areas involve two sets of decision variables: the first set gives the volume of each raw material collected at each forest area and time period and it is denoted by X_i^{pt} , $\forall p \in P$, $\forall t \in T$, $\forall i \in I$. The second set of decision variables corresponds to the binary variables Y_i^t which are equal to 1 if area i , $\forall i \in I$, used for collection at time period t , $\forall t \in T$, and 0 otherwise. Additionally, associated with forest areas and teams that are available to collect them the following parameters are defined: the volume of raw material p available at area i is denoted by V_i^p ; at each time period, there are N_i collecting teams, each one with the capacity of collecting a volume of V_i .

The following constraints define feasible ways of collecting the raw materials at the forest areas.

$$\sum_{t \in T} X_i^{pt} \leq V_i^p, \forall i \in I, \forall p \in P \quad (1)$$

$$X_i^{pt} \leq V_i^p Y_i^t, \forall i \in I, \forall p \in P, \forall t \in T \quad (2)$$

$$\sum_{i \in I} Y_i^t \leq N_i, \forall t \in T \quad (3)$$

$$\sum_{p \in P} X_i^{pt} \leq Vi, \forall i \in I, \forall t \in T \quad (4)$$

Constraints (1) state that the collected volume cannot exceed the existent volume (which was assumed to be the same for all the planning horizon). Constraints (2) link the two types of decision variables: if any volume is collected then the corresponding binary variable is 1. Constraints (3) set the limit of the number of areas where biomass is collected at each period to the number of available teams. Constraints (4) set a limit to the volume (over all raw materials) collected at each area in each period; this limit is given by the capacity of one team.

After being collected, raw materials can be stored and/or chipped (with a mobile chipper) near the forest area, or sent to a warehouse. The following decision variables are defined to accommodate these alternatives:

Xn_i^{pt} - volume of raw material p from area i stored near the forest area at period t , $\forall i \in I, \forall p \in P, \forall t \in T$;

Xn_{iw}^{pt} - volume of raw material p send from area i to warehouse w at period t , $\forall i \in I, \forall p \in P, \forall t \in T, \forall w \in W$;

Xm_i^{pt} - volume of raw material p from area i chipped by a mobile chipper near the forest area at period t , $\forall i \in I, \forall p \in P, \forall t \in T$.

Constraints (5) assure that, at any time period, what exists of any raw material in any area (which may come from the previous period or from collection in the current period) will be kept near the forest, chipped, or transported to a warehouse. These constraints may be seen as (time) flow conservation constraints.

$$Xn_i^{pt} + X_i^{pt} = Xn_i^{p,t+1} + Xm_i^{p,t+1} + \sum_{w \in W} Xn_{iw}^{p,t+1}, \quad (5)$$

$$\forall i \in I, \forall p \in P, \forall t = 1, \dots, T - 1$$

When a raw material is chipped near the forest, it is sent to a warehouse or directly to a customer in the next period (the storage near the forest of chipped products is not allowed). The following decision variables are defined:

Xc_{iw}^{pt} - volume of chipped product p sent from area i to warehouse w at period t , $\forall i \in I, \forall p \in P, \forall w \in W, \forall t \in T$;

Xc_{ij}^{pt} - volume of chipped product p sent from area i to customer j at period t , $\forall i \in I, \forall p \in P, \forall j \in J, \forall t \in T$.

The following constraints relate the volume of chipped products near the forest with the volume sent to the warehouses and/or customer.

$$Xm_i^{pt} = \sum_{w \in W} Xc_{iw}^{p,t+1} + \sum_{j \in J} Xc_{ij}^{p,t+1}, \forall i \in I, \forall p \in P, \forall t = 1, \dots, T - 1 \quad (6)$$

An important aspect of the proposed model is the explicit consideration of two types of chippers: mobile and fixed.

Mobile chippers can be used near the forest areas or in the warehouses and in different places in different periods. Their number is represented by Nm and the capacity of each one (measured in volume of product chipped) is represented by Vm . There were defined two sets of integer decision variables to represent the number of mobile chippers placed at a forest area i in period t , Zm_i^t , $\forall i \in I$, $\forall t \in T$, and the number of mobile chippers placed at warehouse w in period t , Zm_w^t , $\forall w \in W$, $\forall t \in T$. A set of continuous variables, Xm_w^{pt} , $\forall p \in P$, $\forall t \in T$, $\forall w \in W$, representing the volume of raw material p chipped in period t in warehouse w by a mobile chipper were also defined.

The following constraints are related with mobile chippers.

$$\sum_{i \in I} Zm_i^t + \sum_{w \in W} Zm_w^t \leq Nm, \forall t \in T \quad (7)$$

$$\sum_{p \in P} Xm_i^{pt} \leq Vm Zm_i^t, \forall i \in I, \forall t \in T \quad (8)$$

$$\sum_{p \in P} Xm_w^{pt} \leq Vm Zm_w^t, \forall w \in W, \forall t \in T \quad (9)$$

$$Zm_w^t \leq Nm Y_w, \forall w \in W, \forall t \in T \quad (10)$$

Constraints (7) assure that the number of mobile chippers used at each period does not exceed the number of existing mobile chippers. Constraints (8) and (9) guarantee that the chipped volume (over all raw materials) at each time period does not exceed the available capacity of chippers placed at the forest area (constraints (8)) and at the warehouse (constraints (9)). Constraints (10) state that mobile chippers can be placed only at existing warehouses. The decision to open a warehouse w is associated with a binary variable: $Y_w = 1$ if warehouse w is open, and $Y_w = 0$, otherwise, $\forall w \in W$.

Fixed chippers can only be located at warehouses. A binary variable, Yf_w , is associated with the decision of placing a fixed chipper in a warehouse: $Yf_w = 1$ if a fixed chipper is placed at warehouse w , $Yf_w = 0$ otherwise, $\forall w \in W$. A set of continuous variables, Xf_w^{pt} , $\forall p \in P$, $\forall t \in T$, $\forall w \in W$, representing the volume of raw material p chipped in period t in warehouse w by a fixed chipper are defined. The number of available fixed chippers is represented by Nf and the capacity of each one (measured in volume of product chipped) is represented by Vf .

The following constraints are related with fixed chippers.

$$\sum_{p \in P} Xf_w^{pt} \leq Vf Yf_w, \forall w \in W, \forall t \in T \quad (11)$$

$$\sum_{w \in W} Yf_w \leq Nf \quad (12)$$

$$Yf_w \leq Y_w, \forall w \in W \quad (13)$$

Constraints (11) assure that the chipped volume (over all raw materials) at each time period does not exceed the available capacity of fixed chippers (for all warehouses). Constraints (12) state that the number of fixed chippers installed at the warehouses does not exceed the number of available fixed chippers. Constraints (13) state that a fixed chipper can only be installed in an existent warehouse.

To model the stock at the warehouses the additional decision variables are required:

Xn_w^{pt} - volume of the stock of raw material p at period t in warehouse w , $\forall p \in P$, $\forall t \in T$, $\forall w \in W$;

Xc_w^{pt} - volume of the stock chipped product p at period t in warehouse w , $\forall p \in P$, $\forall t \in T$, $\forall w \in W$.

The following constraints are flow conservation constraints with respect to each warehouse and each raw material (constraints 14) and each chipped product (constraints 15). Variables Xc_{wj}^{pt} represent the volume of chipped product p sent from the warehouse w to the customer j at period t , $\forall p \in P$, $\forall w \in W$, $\forall j \in J$, $\forall t \in T$.

$$Xn_w^{pt} + \sum_{i \in I} Xn_{iw}^{pt} = Xn_w^{p,t+1} + Xm_w^{p,t+1} + Xf_w^{p,t+1}, \quad (14)$$

$$\forall w \in W, \forall p \in P, \forall t = 1, \dots, T-1$$

$$Xc_w^{pt} + \sum_{i \in I} Xc_{iw}^{pt} + Xm_w^{pt} + Xf_w^{pt} = Xc_w^{p,t+1} + \sum_{j \in J} Xc_{wj}^{p,t+1}, \quad (15)$$

$$\forall w \in W, \forall p \in P, \forall t = 1, \dots, T-1$$

Representing the capacity of product and chipped product at warehouse w , $\forall w \in W$, as Vn_w and Vc_w , respectively, the following two sets of constraints state that the stock cannot exceed them.

$$\sum_{p \in P} Xc_w^{pt} \leq Vc_w Y_w, \forall w \in W, \forall t \in T \quad (16)$$

$$\sum_{p \in P} Xn_w^{pt} \leq Vn_w Y_w, \forall w \in W, \forall t \in T \quad (17)$$

The demand of a customer j , $\forall j \in J$, in period t , $\forall t \in T$, of chipped product p , $\forall p \in P$, is denoted by V_j^{pt} . The following set of constraints as assure the demand of all customers in all periods is satisfied.

$$\sum_{w \in W} Xc_{wj}^{pt} + \sum_{i \in I} Xc_{ij}^{pt} \leq V_j^{p,t+1}, \forall j \in J, \forall p \in P, \forall t \in T \quad (18)$$

The last set of constraints is related with the number of trips required to send the product between different locations. There are four types of trips: from a forest area to a warehouse (raw material or chipped product), from an area to a customer, from a warehouse to a customer. The following decision variables are associated with the number of trips of each

type required: Zn_{iw}^{pt} , Zc_{iw}^{pt} , Zc_{ij}^{pt} and Zc_{wj}^{pt} . Assuming the capacity of each vehicle is Vk , the following constraints relate the volume sent with the number of trips required.

$$Xn_{iw}^{pt} \leq Vk Zn_{iw}^{pt}, \forall i \in I, \forall w \in W, \forall p \in P, \forall t \in T \quad (19)$$

$$Xc_{iw}^{pt} \leq Vk Zc_{iw}^{pt}, \forall i \in I, \forall w \in W, \forall p \in P, \forall t \in T \quad (20)$$

$$Xc_{wj}^{pt} \leq Vk Zc_{wj}^{pt}, \forall w \in W, \forall j \in J, \forall p \in P, \forall t \in T \quad (21)$$

$$Xc_{ij}^{pt} \leq Vk Zc_{ij}^{pt}, \forall i \in I, \forall j \in J, \forall p \in P, \forall t \in T \quad (22)$$

3.2 Objective Function

The objective function aims at minimizing the total cost, which can be divided in four components: collecting costs, chipping costs, storage costs, and transportation costs. For most of them two types of costs are considered: a fixed cost which is incurred if the activity is performed, no matter what its level is, and a variable cost which is proportional to the level of the activity. Although this cost structure is nonlinear, the way the decision variables were defined and linked (through constraints) allows the construction of a linear objective function.

The following parameters related with the collecting costs are defined:

Ciq - fixed cost for collecting raw material from a forest area in a time period;

Civ - cost for collecting one unit of volume of raw material from a forest area in a time period.

The collecting cost, Ci , is

$$Ci = \sum_{i \in I} \sum_{t \in T} Ciq Y_i^t + \sum_{i \in I} \sum_{p \in P} \sum_{t \in T} Civ X_i^{pt}$$

And the following parameters related with the chipping costs are used:

Cmq - fixed cost of using a mobile chipper in a period;

Cmv - cost for chipping one unit of volume of any raw material with a mobile chipper at a period;

Cfq - cost of installation of a fixed chipper at a warehouse;

Cfv - cost for chipping one unit of volume of any raw material with a fixed chipper at a period.

The chipping cost, Cc , is

$$\begin{aligned} Cc = & \sum_{i \in I} \sum_{t \in T} Cmq Zm_i^t + \sum_{w \in W} \sum_{t \in T} Cmq Zm_w^t + \\ & \sum_{i \in I} \sum_{p \in P} \sum_{t \in T} Cmv Xm_i^{pt} + \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} Cmv Xm_w^{pt} + \\ & \sum_{w \in W} Cfq Yf_w + \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} Cfv Xf_w^{pt} \end{aligned}$$

The following parameters associated with the storage costs are defined:

Swq - fixed cost incurred when opening one warehouse;

Siv - cost of storing one unit of volume of raw material near the forest during one period;

Snv - cost of storing one unit of volume of raw material in a warehouse during one period;

Scv - cost of storing one unit of volume of chipped product in a warehouse during one period.

The cost related with storage, Cw , is

$$Cw = \sum_{i \in I} \sum_{p \in P} \sum_{t \in T} Siv Xn_i^{pt} + \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} Scv Xc_w^{pt} + \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} Snv Xn_w^{pt} + \sum_{w \in W} Swq Y_w$$

Lastly, the transportation cost is related with the following parameters:

Ck - cost of one Km run by a vehicle;

d_{ij} - distance (in Km) between forest area i and customer j , $\forall i \in I$, $\forall j \in J$;

d_{iw} - distance (in Km) between forest area i and warehouse w , $\forall i \in I$, $\forall w \in W$;

d_{wj} - distance (in Km) between warehouse w and customer j , $\forall w \in W$, $\forall j \in J$.

To exemplify how the transportation cost is structured an example of the transportation cost function is shown in Fig. 2. The example concerns the cost of transport between forest area i and customer j for the chipped product p and time period t , with the maximum number of trips is assumed to be three.

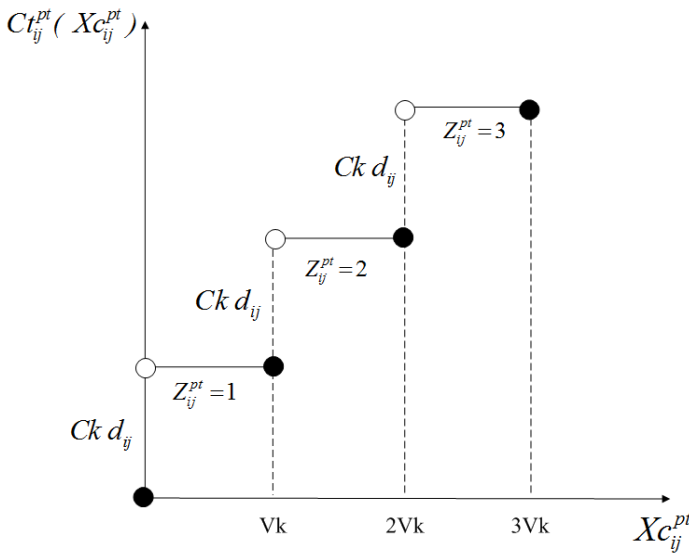


Fig. 2. Example of a transportation cost structure

The cost related with transportation, C_t , is

$$C_t = \sum_{i \in I} \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} C_k d_{iw} Z n_{iw}^{pt} + \sum_{i \in I} \sum_{w \in W} \sum_{p \in P} \sum_{t \in T} C_k d_{iw} Z c_{iw}^{pt} + \sum_{w \in W} \sum_{j \in J} \sum_{p \in P} \sum_{t \in T} C_k d_{wj} Z c_{wj}^{pt} + \sum_{i \in I} \sum_{j \in J} \sum_{p \in P} \sum_{t \in T} C_k d_{ij} Z c_{ij}^{pt}$$

Summing up, the objective function is

$$\text{Min } Z = C_i + C_c + C_w + C_t$$

4 Computation Results

The mixed integer programming model was implemented in C++ using the callable library of Cplex 12.1 which uses the branch-and-cut method. The computer used to perform tests was an Intel® Core™ 2 Duo 2.00GHz CPU and had 2.00GB of RAM.

A set of instances was generated based on the information provided in [14].

The sets of parameters that has a major impact on the size of the model were varied to check their behavior for instances of small size and for larger instances (closer to reality). Thus, the model was tested with 5 and 150 areas of forest, with 1 and 4 potential warehouses, with 3 and 7 customer, with 1 and 4 products, and with 7, 30, 90 and 150 time periods. In practice, a time period is assumed to be one day.

Two tables of computational results are presented below: Table 1 and Table 2. Each table has four columns (left) with the four main parameters of the model ($|I|$ - number of forests area, $|W|$ - number of potential warehouses, $|J|$ - number of customers, $|P|$ - number of products) and two sets of six columns, referring to two periods of time. For each time period the first three columns correspond to the size of the model (Lines - number of constraints, Columns - number of variables, Non-zero - the number of elements different from zero of the matrix associated with constraints) and the latter three referring to the results obtained (Z - value of the best solution found, $\text{GAP} = 100 |Z - \text{LI}| / |Z| \%$ where LI corresponds to the lower bound for the value of the optimal solution given by Cplex, Time - runtime in seconds). Note that, for example, a GAP of 0.04% means that the solution has a cost higher than the optimal solution by no more than 0.04%. Two stopping criteria were defined, one for the maximum time (3600 seconds) and the other referring to GAP (10^{-4}).

For most instances, the application was stopped because it reaches the time limit. When in column Z appears the symbol * or ** it means that no integer solution has been reached or that there was a mistake due to lack of memory of Cplex, respectively.

For instances with fewer time periods (Table 1.) it was almost always possible to obtain quality solutions on a very acceptable time of one hour. The exception was the instance signaled with * in the Table 1. Note that half of these instances have 150 forest areas.

Table 1. Results for the instances with the time period equal to 7 and 30

I	W	J	P	T					
				Lines	Columns	Non-zero	Z	GA P	Time
7									
5	1	3	1	513	604	1,445	206,880	0.01%	122
5	1	3	4	1,689	2,179	5,204	597,071	0.04%	3,600
5	1	7	1	757	940	2,173	406,504	0.06%	3,600
5	1	7	4	2,665	3,523	8,116	1,477,708	0.02%	3,600
5	4	3	1	1,029	1,261	3,050	209,389	0.01%	131
5	4	3	4	3,429	4,726	11,156	587,944	0.04%	3,600
5	4	7	1	1,381	1,765	4,222	409,068	0.01%	3,600
5	4	7	4	4,837	6,742	15,844	1,489,029	0.01%	3,600
150	1	3	1	11,968	15,829	35,810	198,940	4.48%	3,604
150	1	3	4	41,419	56,989	130,484	687,353	0.75%	3,602
150	1	7	1	17,432	24,285	53,358	333,517	0.87%	3,605
150	1	7	4	63,275	90,813	200,676	1,546,040	1.60%	3,605
150	4	3	1	20,314	28,666	61,340	195,933	3.48%	3,601
150	4	3	4	74,479	108,256	232,136	687,189	1.05%	3,604
150	4	7	1	25,886	37,290	79,332	333,042	0.82%	3,602
150	4	7	4	96,767	142,752	304,104	1,535,634	1.59%	3,605
30									
				Lines	Columns	Non-zero	Z	GA P	Time
5	1	3	1	2,008	2,582	6,183	842,725	0.17%	3,600
5	1	3	4	6,496	9,332	22,293	3,584,108	0.30%	3,600
5	1	7	1	2,896	4,022	9,211	1,938,850	0.25%	3,600
5	1	7	4	10,048	15,092	34,405	6,755,222	0.41%	3,600
5	4	3	1	3,904	5,378	13,032	803,498	0.28%	3,602
5	4	3	4	12,721	20,228	47,772	3,453,325	0.59%	3,601
5	4	7	1	5,176	7,538	17,884	1,943,465	0.24%	3,608
5	4	7	4	17,809	28,868	67,180	8,066,747	0.18%	3,601
150	1	3	1	46,813	67,832	153,938	851,028	0.59%	3,603
150	1	3	4	159,616	244,232	561,113	3,926,461	7.97%	3,607
150	1	7	1	66,261	104,072	227,146	2,053,567	1.79%	3,604
150	1	7	4	237,408	389,192	853,945	8,952,701	3.28%	3,609
150	4	3	1	76,549	122,828	264,752	837,781	1.85%	3,604
150	4	3	4	277,201	463,928	1,002,452	*		
150	4	7	1	96,381	159,788	339,784	1,967,425	1.32%	3,605
150	4	7	4	356,529	611,768	1,302,580	8,952,701	3.28%	3,600

Table 2. Results for the instances with the time period equal to 90 and 150

I	W	J	P	T					
				Lines	Columns	Non-zero	Z	GA P	Time
90									
5	1	3	1	5,908	7,742	18,543	2,561,156	0.24%	3,600
5	1	3	4	19,036	27,992	66,873	12,048,978	0.23%	3,601
5	1	7	1	8,476	12,062	27,571	6,755,222	0.41%	3,600
5	1	7	4	29,308	45,272	102,985	26,472,796	0.15%	3,601
5	4	3	1	11,404	16,118	39,072	2,600,874	0.30%	3,601
5	4	3	4	36,961	60,668	143,292	10,975,805	0.30%	3,605
5	4	7	1	15,076	22,598	53,524	6,175,014	0.27%	3,601
5	4	7	4	51,649	86,588	201,100	25,656,848	0.32%	3,605
150	1	3	1	137,713	203,492	462,098	2,787,543	1.01%	3,628
150	1	3	4	467,956	732,692	1,684,493	**		
150	1	7	1	193,641	312,212	680,506	*		
150	1	7	4	691,668	1,167,572	2,558,125	**		
150	4	3	1	223,249	368,468	795,392	2,758,312	3.57%	3,612
150	4	3	4	806,041	1,391,768	3,011,972	**		
150	4	7	1	280,281	479,348	1,019,224	6,487,737	4.75%	3,621
150	4	7	4	1,034,169	1,835,288	3,907,300	**		
150									
				Lines	Columns	Non-zero	Z	GA P	Time
5	1	3	1	9,808	12,902	30,903	4,815,603	0.80%	3,601
5	1	3	4	31,576	46,652	111,453	19,092,615	0.32%	3,602
5	1	7	1	14,056	20,102	45,931	11,304,564	0.44%	3,601
5	1	7	4	48,568	75,452	171,565	43,780,685	0.16%	3,603
5	4	3	1	18,904	26,858	65,112	4,315,313	0.72%	3,603
5	4	3	4	61,201	101,108	238,812	17,599,099	0.28%	3,604
5	4	7	1	24,976	37,658	89,164	10,324,530	0.34%	3,600
5	4	7	4	85,489	144,308	335,020	42,573,105	0.32%	3,607
150	1	3	1	228,613	339,152	770,258	4,874,245	3.34%	3,612
150	1	3	4	776,296	1,221,152	2,807,873	**		
150	1	7	1	321,021	520,352	1,133,866	**		
150	1	7	4	1,145,928	1,945,952	4,262,305	**		
150	4	3	1	369,949	614,108	1,326,032	*		
150	4	3	4	1,334,881	2,319,608	5,021,492	**		
150	4	7	1	464,181	798,908	1,698,664	**		
150	4	7	4	1,711,809	3,058,808	6,512,020	**		

For instances with more time periods (Table 2.) in cases with fewer forests it has always been possible to reach satisfactory solutions. For instances with 150 forest areas, with 90 periods it was still possible to find some quality solutions in the time available. For instances with 150 time periods solutions were not obtained due to the model dimension (impracticable for Cplex).

5 Conclusions

This paper proposed a mixed integer programming model that supports, in an integrated approach, operational and tactical decisions for the management of a wood biomass supply chain. The model was implemented in C++ using the callable library of Cplex 12.1. Computational tests were performed in 64 instances based on information from the literature, with 5 and 150 areas of forest, with 1 and 4 potential warehouses, with 3 and 7 customers, with 1 and 4 products, and with 7, 30, 90 and 150 periods of time.

For instances of small and medium-scale, solutions were obtained in an hour of computing time, whose value is close to the value of optimal solution (in most cases the difference is less than 1%).

For larger instances, with 150 forest areas and with 90 periods, it was possible to find some quality solutions on the available time, but with 150 forest areas and 150 periods there was not obtained a solution because the model has a dimension impracticable for Cplex.

As future work it is planned to use more efficient resolution techniques to solve larger problems and a more expeditious approach. At the same time it also is intended to introduce in the model some more characteristics of this supply chain, such as the seasonality of supply and the change of products characteristics over time with impact on the economic performance of the biomass supply chain.

Acknowledgments. This work has been partially funded by FCT (Fundação para a Ciência e a Tecnologia - Portugal) through the PhD. grant SFRH / BD / 44601 / 2008 of the first author and through project PTDC/EIA-EIA/100645/2008 "SearchCol: Meta-heuristic Search by Column generation".

References

1. European Union, C: Green Paper - A European Strategy for Sustainable, Competitive and Secure Energy (2006)
2. United Nations: Kyoto Protocol to the United Nations Framework Convention on Climate Change (1998)
3. European Parliament and Council: Directive 2009/28/EC amending and subsequently repealing 2001/77/EC and 2003/30/EC so on the promotion of the use of energy from renewable sources. Official Journal of the European Union L 140, 16–62 (2009)
4. European Parliament and Council: Directive 2009/29/EC amending Directive 2003/87/EC so as to improve and extend the greenhouse gas emission allowance trading scheme of the Community. Official Journal of the European Union L 140, 63–87 (2009)
5. CEN/TS: EN 14588 – Solid biofuels, Terminology, definitions and description (2009)
6. EUBIONET III, Alakangas, E.: EN 14961-1 - Classification of biomass origin in European solid biofuel standard (2010)
7. European Union, C: Biomass action plan (2005)
8. Johnson, E.: Goodbye to carbon neutral: Getting biomass footprints right. Environmental Impact Assessment Review 29, 165–168 (2009)

9. Tekes: Developing technology for large-scale production of forest chips - Wood Energy Technology Programme 1999–2003 (2004)
10. EUBIONET II: Final result-oriented report - Efficient trading of biomass fuels and analysis of fuel supply chains and business for market actors by networking (2008)
11. Sokhansanj, S., Kumar, A., Turhollow, A.F.: Development and implementation of integrated biomass supply analysis and logistics model (IBSAL). *Biomass and Bioenergy* 30, 838–847 (2006)
12. Iakovou, E., Karagiannidis, A., Vlachos, D., Toka, A., Malamakis, A.: Waste biomass-to-energy supply chain management: A critical synthesis. *Waste Management* 30(10), 1860–1870 (2010)
13. Gunnarsson, H., Ronnqvist, M., Lundgren, J.T.: Supply chain modelling of forest fuel. *European Journal of Operational Research* 158, 103–123 (2004)
14. EUBIONET II: Factsheet 17 - Wood chips for a district heating plant - case district heating plant in Central Carinthia, Austria (2007)

On Solving a Stochastic Programming Model for Perishable Inventory Control*

Eligius M.T. Hendrix^{1,2}, Rene Haijema^{2,3}, Roberto Rossi⁴,
and Karin G.J. Pauls-Worm²

¹ Computer Architecture, Universidad de Málaga

² Operations Research and Logistics, Wageningen University

³ TI Food and Nutrition, Wageningen

{eligius.hendrix, rene.haijema, karin.pauls}@wur.nl

⁴ Business School, University of Edinburgh

robros@gmail.com

Abstract. This paper describes and analyses a Stochastic Programming (SP) model that is used for a specific inventory control problem for a perishable product. The decision maker is confronted with a non-stationary random demand for a fixed shelf life product and wants to make an ordering plan for a finite horizon that satisfies a service level constraint. In literature several approaches have been described to generate approximate solutions. The question dealt with here is whether exact approaches can be developed that generate solutions up to a guaranteed accuracy. Specifically, we look into the implications of a Stochastic Dynamic Programming (SDP) approach.

Keywords: Stochastic Programming, Dynamic Programming, Inventory control, Perishable products, Service level constraint.

1 Introduction

We consider a production planning problem over a finite horizon of T periods of a perishable product with a fixed shelf life of J periods. The demand is uncertain and non-stationary such that one produces to stock. To keep waste due to outdating low, one issues the oldest product first, i.e. FIFO issuance. A service level applies to guarantee that the probability of not being out-of-stock is higher than α in every period $t \in \{1, 2, \dots, T\}$. Any unmet demand is backlogged. In [9], a Stochastic Programming model has been introduced that uses a chance constraint to generate order policies for this planning problem. Specifically, an MILP re-formulation to generate an approximate solution and an enumeration method based on Sample Average Approximation are investigated in that paper.

* This paper has been supported by The Spanish Ministry of Science and Innovation (project TIN2008-01117) and Junta de Andalucía (P11-TIC-7176), in part financed by the European Regional Development Fund (ERDF). Eligius M.T. Hendrix is a fellow of the Spanish “Ramon y Cajal” contract program, co-financed by the European Social Fund. The study is co-funded by the TIFN (project RE002).

The question is here, whether methods can be developed based on the properties of the model to generate solutions for a realistic time horizon and shelf life based on Stochastic Dynamic Programming (SDP). The target is to generate solutions for the instances used in [9] which relates to a practical planning problem with a time horizon of $T = 12$ and shelf life of the perishable products of $J = 3$ periods. This is followed by questions on how solution approaches behave for varying settings of the problem.

This paper is organised as follows. Section 2 describes the underlying SP model. Section 3 describes a conceptual solution approach based on Stochastic Dynamic Programming (SDP) and the properties of the underlying problems to be solved. In Sect. 4 we illustrate the approach with several instances. Section 5 summarises our findings.

2 Stochastic Programming Model

The model is summarised from an optimisation perspective. As much as possible, in the used symbols, we distinguish between model parameters (exogenous in lower case letters) and decision variables (capital letters) that are characterised by direct decision variables and dependent stock variables. Capitals are also used for upper bounds on the indices.

Indices

t period index, $t = 1, \dots, T$, with T the time horizon
 j age index, $j = 1, \dots, J$, with J the fixed shelf life

Data

d_t Normally distributed demand, cumulative distribution function (cdf) F_t
 k fixed ordering cost, euro
 c procurement cost, euro/ton
 h inventory cost, euro/ton
 w disposal cost, euro/ton, is negative when having salvage value
 α service level

Here we have assumed the product demand to be continuous, but the model can be modified easily to deal with the discrete demand case.

Variables

$Q_t \geq 0$ ordered and delivered quantity at beginning period t , ton
 I_{jt} Inventory of age j at end of period t ,
 with the initial closing inventory levels fixed to $I_{j0} = 0$, and
 variable I_{1t} is free (as we assume backlogging and FIFO issuing),
 whereas $I_{jt} \geq 0$ for $j = 2, \dots, J$.

The inventory variables (apart from the initial fixed levels I_{j0}) are random variables due to the stochastic nature of demand. If the order decision Q_t depends on the inventory levels $(I_{1,t-1}, \dots, I_{J-1,t-1})$, Q_t is also a random variable.

In the notations below $P(\cdot)$ denotes a probability and $E(\cdot)$ is the expected value operator. The total expected costs over the finite horizon is to be minimised:

$$E \left(\sum_{t=1}^T \left(h \sum_{j=1}^{J-1} I_{jt}^+ + g(Q_t) + wI_{Jt} \right) \right) = \sum_{t=1}^T E \left(g(Q_t) + h \sum_{j=1}^{J-1} I_{jt}^+ + wI_{Jt} \right), \quad (1)$$

where procurement cost is given by the function

$$g(x) = k + cx, \quad \text{if } x > 0, \quad \text{and } g(0) = 0 \quad (2)$$

and the notation $y^+ = \max\{0, y\}$ is used. The chance constraint is

$$P(I_{1t} \geq 0) \geq \alpha, \quad t = 1, \dots, T \quad (3)$$

and the dynamics of the inventory of the items of different ages is described by

$$I_{1t} = Q_t - (d_t - \sum_{j=1}^{J-1} I_{j,t-1})^+, \quad t = 1, \dots, T \quad (4)$$

and

$$I_{jt} = \left(I_{j-1,t-1} - (d_t - \sum_{i=j}^{J-1} I_{i,t-1})^+ \right)^+, \quad t = 1, \dots, T, j = 2, \dots, J. \quad (5)$$

These dynamics equations describe the FIFO issuing policy and imply that I_{1t} is a free variable, whereas I_{jt} is nonnegative for the older vintages $j = 2, \dots, J$. A feasible order strategy $Q_t(I)$ of the SP model fulfills the nonnegativity aspects and equations (3), (4) and (5). An optimal solution also minimises (1).

The approaches described in [9], in fact look for the best static moments of ordering and translate the expected value of I_t and Q_t into an appropriate order-up-to level S_t . For periods where $E(Q_t)$ is positive an order is placed at the beginning of that period of size $S_t - \sum_{j=1}^{J-1} I_{j,t-1}$. The result of the approaches is a static-dynamic solution, where the order moments are fixed but the order size depends on a fixed S_t and the actual value of $\sum_{j=1}^{J-1} I_{jt}$. For practical reasons one may add a correction for the expected waste till the next review period. Such a policy is not necessarily an optimal strategy but is feasible and practically useful.

An optimal state dependent policy assumes that both the order moment and the order quantity depend on the actual value of the stock levels of all vintages $\sum_{j=1}^{J-1} I_{jt}$. That is an optimal rule is a function $Q_t(I_{j,t-1}, \dots, I_{J-1,t-1})$. The first studies to such an stock-age dependent rule date back to the early seventies. For an overview see Nahmias [8] and Karaesmen et al. [7]. In these early studies, some analytical results are presented for stylised models that allow for mathematical analysis. Practical results for more realistic models are achieved at the

beginning of this century by [2], [4], and [3]. In these papers, an optimal strategy is determined by Stochastic Dynamic Programming (SDP) for both periodic (stationary) and non-stationary problems with a positive lead time of one period under the lost sales assumption. In order to solve large scale problems, aggregation of states is used in that paper. In the current paper, we follow a similar approach that solves large scale problems without aggregation of states. Moreover, we add a service level constraint and restrict ourselves to a finite horizon setting with zero lead time and backlogging of unmet demand.

3 SDP Approach

Stochastic Dynamic programming is an appropriate technique to approach (II), as the problem is clearly separable in t . Ingredients of a stochastic dynamic program are the state (and state space), the action (and action space), and a state transition function, next to a contribution function and an objective function. The interested reader is referred to the books [1] and [6] for an introduction in (Stochastic) Dynamic Programming. With respect to the state space, first notice that the waste I_{Jt} does not influence future decisions, as it is not further available; it is not a state variable. The state values are given by $X_t = (I_{1,t}, \dots, I_{J-1,t})$ in $(J - 1)$ -dimensional space. In this $(J - 1)$ -dimensional space, the transition is provided by (4) and (5), so abstractly we have a state transition function Φ :

$$X_t = \Phi(X_{t-1}, Q_t, d_t), \quad t = 1, \dots, T \quad (6)$$

and we are looking for a rule $Q_t(X_{t-1})$. To facilitate notation, it is convenient to denote the total available inventory at the beginning of period t to fulfil demand in that period by

$$Y_t = \sum_{j=1}^{J-1} X_{j,t-1}.$$

Now the chance constraint including the nonnegativity of Q_t , can be written as

$$Q_t \geq (\Gamma_t^{-1}(\alpha) - Y_t)^+. \quad (7)$$

The waste I_{Jt} is a function of the inventory at the beginning of the period and the demand; $I_{Jt} = f(X_{t-1}, d_t)$. We can write the expected contribution to the objective function in period t as function of state X_{t-1} and decision Q_t :

$$EC(X_{t-1}, Q_t) = g(Q_t) + E\{wf(X_{t-1}, d_t) + h\mathbf{1}^T \Phi(X_{t-1}, Q_t, d_t)\}, \quad (8)$$

where $\mathbf{1}$ is the all-ones vector. Notice that cost and transition functions are not time dependent in the presented model, although the same approach holds for time dependent procurement cost.

The SDP objective function can be written down in a conceptual way via the Bellmann equation using a value function V :

$$V_t(X) = \min_Q (EC(X, Q) + E[V_{t+1}(\Phi(X, Q, d_t))]), \quad (9)$$

subject to Q fulfilling (7). The argmin of (9) represents the optimum strategy $Q_t(X)$. So starting with a valuation $V_T(X)$ for every possible closing inventory state X , one can compute the valuations backward to know $V_{t-1}(\cdot), \dots, V_1(\cdot)$ and the optimizing order quantity function $Q_t(\cdot)$. The final result $Q_t(X)$ tells us how much to order in period t given the state of inventory is X . It represents a decision function or a rule that results into the minimum expected cost $V_1(X_0)$ over the time horizon. In a rolling horizon situation, only the optimal decision $Q_1(X_0)$ is implemented and after observing the new inventory levels and updating the demand predictions for the next T periods the SDP may be executed once again to determine the next order quantity.

Our focus is on how to code the computations of the SDP approach to compute $Q_t(X)$ efficiently. Therefore we first look into the easier characteristics of the case where d_t is deterministic in Section 3.1. In Section 3.2 we go into practical and theoretical considerations for the stochastic SDP approach.

3.1 Solution Properties for Deterministic Demand

For the case known as variable demand, where d_t is given for a finite horizon, waste and backlogging can be avoided. As we are dealing with a perishable product, it can be derived that the order Q_t consists of a sum of future demand for an integer number of periods:

$$Q_t \in \{0, d_t, d_t + d_{t+1}, \dots, d_t + d_{t+1} + \dots + d_{t+J-1}\}.$$

Notice, that this is optimal in a situation with backlogging. In a lost sales situation, the decision maker can simply order less depending on the interpretation of

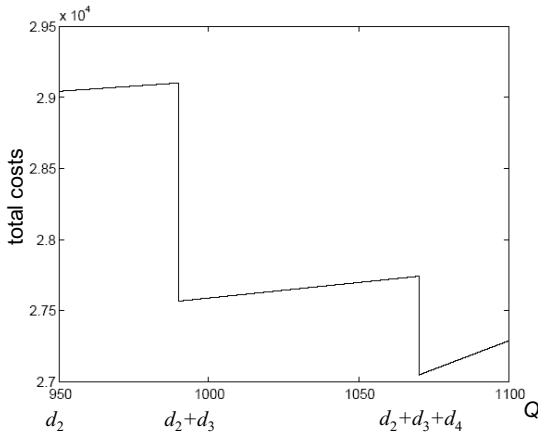


Fig. 1. Minimum total cost at $t = 2$ for $I_{1t} = I_{2t} = 0$ to the end of the planning horizon as function of order decision Q

the service level constraint. Using (deterministic) Dynamic programming (DP), a solution of the Bellmann equation

$$V_t(X) = \min_{Q_t} (g(Q_t) + V_{t+1}(\Phi(X, Q_t, d_t))), \tag{10}$$

can easily be found; if demand is larger than the inventory on hand, $Y_t < d_t$, order a sum of future demands and else, do not order at all. Practically, as no analytical expression can be derived, the implementation of the value functions V_t , consists of discretising the state space \mathcal{X} and using interpolation within (10) to evaluate the state one arrives at taking a decision Q_t .

In order to implement a discretisation of the state space, one should have a clue about the range of the state variables X_j . In the deterministic case one can take the range $[0, U_{jt}]$ of X_{jt} and choose upper bound U_{jt} as

$$U_{jt} \geq d_{t+1} + \dots + d_{t+1+J-j}, \quad t = 1, \dots, T.$$

Implicitly $d_k = 0$ here if $k > T$. We illustrate with the following instance.

Table 1. Varying demand d_t and upper bounds for the inventory, $J = 3$

t	1	2	3	4	5	6	7	8	9	10	11	12
d_t	1900	950	40	80	30	150	800	950	1100	350	150	700
$I_{1t} \leq$	990	120	110	180	950	1750	2050	1450	500	850	700	0
$I_{2t} \leq$	950	40	80	30	150	800	950	1100	350	150	700	0

Example 1. Consider an instance with a perishability of $J = 3$ periods and costs given by $c = 2, h = 1, w = 4$ and $k = 3000$. Table 1 provides the data for the demand d_t and also upper bounds for the inventory if inventory costs and waste are minimised; $I_{1t} \leq d_{t+1} + d_{t+2}$ and $I_{2t} \leq d_{t+1}$. Taking a global upper bound for the inventory implies using $U_1 = \max_t \{d_{t+1} + d_{t+2}\} = 2050$ and $U_2 = \max_{t \geq 2} d_{t+1} = 1100$.

Implementation of the DP approach discretising the state space with steps of 10, leads to the optimal decision sequence that can easily be verified: $Q_1 = 2890, Q_4 = 260, Q_7 = 1750, Q_9 = 1600, Q_{12} = 700$ resulting in a total cost of 32360. In a discretised state space, the DP approach requires solving (10) for each grid point in that space. An illustration is given in Fig. 1, which shows the objective value of (10) for $t = 2$ and $X = (0, 0)$ for varying values of Q starting at d_2 . The first slope denotes the inventory cost of one period, the second slope the inventory cost of two periods and the third one represents the increasing cost of waste caused by ordering more than $d_2 + d_3 + d_4$.

The figure shows that in fact one is iteratively minimising a global optimisation problem, see 5. For the deterministic case, it has easy to determine candidates for the minimum points. As will be illustrated, the stochastic model inherits the global optimisation character, but the minimum points are not that easy to determine.

3.2 Solution Properties for Stochastic Demand

In the stochastic model, the demand is a random variable with known distribution functions Γ_t . The stochastic model allows negative inventory values in 5% of the cases when a 95% service level applies. One may expect additional costs due to additional production runs when demand appeared to be too big and due to product waste, which may be inevitable.

There are several complications to deal with when we are confronted with stochastic demand. How to deal with the probability distribution of the demand? How to bound the state space? How to solve (9) iteratively?

For notational ease, assume that the demand is Normally distributed with the same coefficient of variation cv over all periods: $d_t \sim \mu_t \times (1 + cv \times N(0, 1))$. In this way, demand is fulfilled with a probability of $\alpha\%$, if at the beginning of the period more than $s_t := \Gamma_t^{-1}(\alpha) = (1 + cvG^{-1}(\alpha))\mu_t$ of the product is available, where G is the cdf of the standard normal distribution. For the illustration, Table 2 shows the so-called safety stock s_t and μ_t that corresponds to the data of Example 1 and a $cv = 0.33$, service level $\alpha = 95\%$. The analogy with the deterministic case is that no order is required if the current stock Y_t at the beginning of the period is larger than safety stock s_t .

Table 2. 95% safety level s_t for mean demand μ_t and $cv = 0.33$

t	1	2	3	4	5	6	7	8	9	10	11	12
μ_t	1900	950	40	80	30	150	800	950	1100	350	150	700
s_t	2931	1389	62	123	46	231	1234	1466	1697	540	231	1080

One can discretise the space of possible outcomes of the stochastic demand by using the quantiles of the normal distribution. Practically this works by using an equidistant grid over the probability range $[0, 1]$ with a step p and generating a discrete outcome space $\{\Gamma^{-1}(p), \Gamma^{-1}(2p), \Gamma^{-1}(3p), \dots, \Gamma^{-1}(1 - p)\}$. The consequence of this operation is that the outcome space is truncated by the p -quantiles and every outcome has the same probability of occurrence.

The expected values of the cost and valuation of the state one arrives at, is approximated by an average using the discrete outcomes and the probability. In the deterministic example, the step size in the grid was chosen such that the transition leads to other grid points. An additional complication here is that the transition of all possible outcomes will lead to points in between the grid points requiring the use of interpolation. If the ranges of the state space are not chosen large enough, also extrapolation is required.

The ranges for the stock are less easy to derive than in the deterministic case. Conceptually, there are no bounds, as d_t may not have a bounded support. However, due to the truncation of the outcome space, we have $d_t \in [dmin_t, dmax_t] = [G^{-1}(p) + 1, G^{-1}(1 - p) + 1] \times \mu_t \times cv$. As it makes no sense to order more than $Qmax_t = dmax_t + dmax_{t+1} + \dots + dmax_{t+J-1} - Y_t$, we have an upper bound on the decision Q_t . Using $dmin_t$ as the minimum that will be demanded, we also have bounds on the inventory. For instance $I_{1t} \leq Qmax_t - dmin_t$. The inventory of one period old can never get more negative than $s_t - dmax_t$.

Finally, for every grid point X in the state space, we should now approximate (9) by interpolation of V_t given the discretised values of d_t and find the best order quantity $Q_t(X)$. The difficulty to do so is illustrated in Fig. 2. It shows

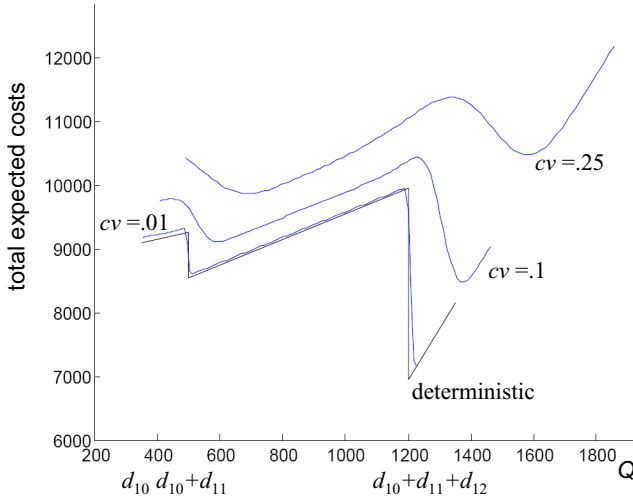


Fig. 2. Minimum total expected cost at $t = 10$, $I_{1t} = I_{2t} = 0$ to the end of the planning horizon as function of order decision Q for different variation coefficient values

problem (9) for the last three periods of Example 1. We are again dealing with a one-dimensional global optimisation problem for each grid point. Notice that with increasing uncertainty the function to be minimised gets more smooth, leads to higher expected cost and gives the tendency to order for less periods ahead. The figure also illustrates, that when the cv goes up from .1 to .25 the best order quantity is no longer in the range of ordering for 2 periods ahead. For the minimisation, first the best value for Q was determined over a set of gridpoints within a range of Q . Then from that point, a local search procedure FMINBND of MATLAB was called to finetune the best value.

4 Comparative Study

The next question is what is the quality of the described approach in terms of effectiveness and efficiency. With respect to effectiveness, one can estimate the expected cost of the strategy Q_t , by running a simulation with a large number of pseudo randomly generated demand series. In [9], $N = 5000$ series were used to estimate the expected cost with an accuracy of 99%. Moreover, one can test, whether the chosen strategy is really feasible with respect to the chance constraint by keeping track on the number of times that a negative inventory level is reached.

With respect to efficiency, in a decision support environment with a rolling planning horizon, the optimisation should be repeated on a daily basis. That means that calculation times should at least be smaller than say an hour in order to fill in new data and place the order. The calculation time depends on the chosen grid density in the state space and is linear in T . It also depends on the goodness of the implementation, the programming language and the platform on which it is run.

We first illustrate the approach with the base case. Then we consider again what happens if variation is going up.

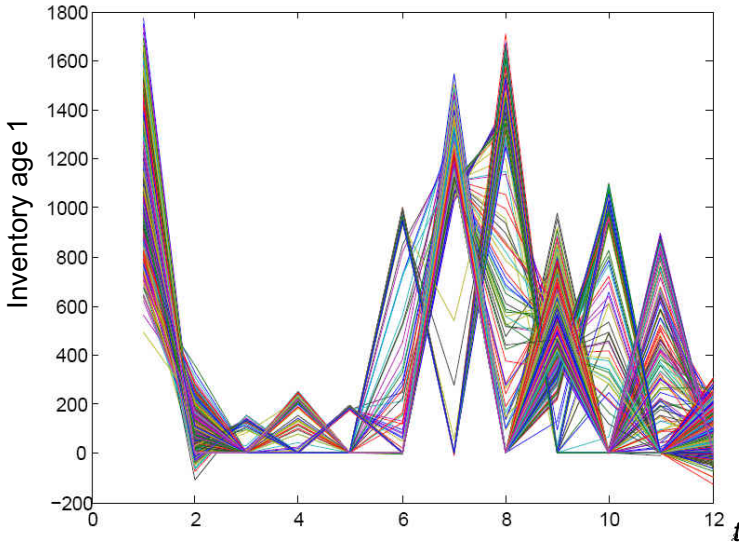


Fig. 3. Development of inventory of age 1 when following the SDP strategy for 2000 generated demand patterns, $cv = 0.1$

Example 2. Consider again the case of Example 1 with $cv = 0.1$. The SDP procedure was implemented in MATLAB on a standard PC. The state space consist of a grid of $100 \times 60 = 6000$ points for which the value function is determined for $t = 3, \dots, 12$. For $t = 2$, $X_2 = 0$, so only 100 values are determined and finally $V_1(0) = 3776$ and $Q_1(0) = 3060$ are determined. For the demand d_t , 50 outcomes are generated and evaluated. The calculation of the complete decision table took 500 seconds.

For illustration and evaluation, in total $N = 2000$ demand series were generated for the 12 periods according to the demand distribution $d_t \sim N(\mu_t, 0.1\mu_t)$. The outcome of the SDP procedure Q_t is evaluated for all repetitions providing an average cost of 38126 which is close to the prediction by the SDP of $V_1(0) = 3776$. As can be observed in Fig. 3, negative inventory levels are only reached at the end of period 2 and at the end of the time horizon. In fact, negative levels are reached at the end of period 2 in 2.2% of the series and in 4.6%

of the series at the end of the horizon. That means that chance constraints for the individual periods do not seem binding. Figure 4 provides the development of the inventory of two periods old. One can observe products left at the end of the horizon. However, it is not counted for waste as the items can still be sold in period $T + 1$.

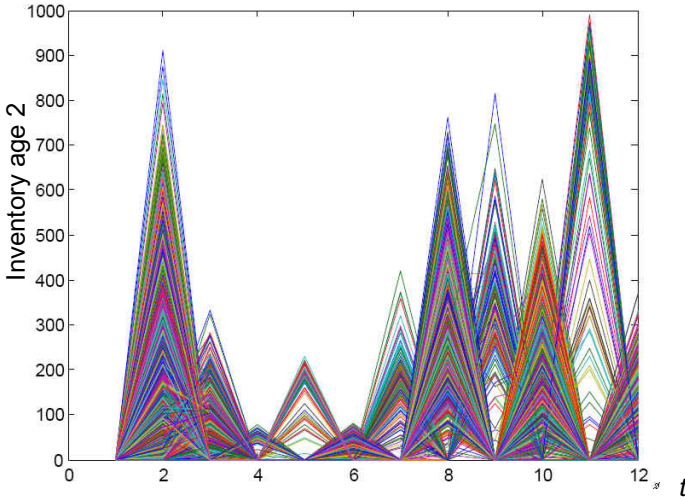


Fig. 4. Development of inventory of age 2 periods when following the SDP strategy for 2000 generated demand patterns, $cv = 0.1$

We now experiment further with the base case testing the robustness of the algorithm and the behaviour of the model with increasing variation modelled by the coefficient of variation cv . The estimates are based on 5000 repetitions (runs) of a pseudo randomly generated demand series. The results are summarised in Table 3. Over the 5000 repetitions the average order quantities and waste are measured to approximate the resulting mean of the generated order policy $Q_t(X)$. Also the number of runs that lead to out of stock (negative inventory) is also counted for each period in order to check the fulfilment of the service level constraint. In fact, the computational effort of all three generated scenarios is in principle the same, keeping the number of grid points equal.

What can typically be learned from the observations, is that increasing uncertainty leads to ordering for less periods in the optimal strategy. Due to the nonstationarity of the demand, one can see that the critical periods with respect to non-negative demand are shifting to other periods when the variation increases. The expected behaviour of more emergency orders (with on average higher fixed order costs), more waste, and higher stock due to higher safety stocks, can also be observed.

Table 3. Average values base case over 5000 runs, increasing uncertainty, %oos: percentage of runs that lead to backlog in that period (negative inventory)

Cost	$cv = 0.1$			$cv = 0.25$			$cv = 0.33$		
	38126			43141			46097		
t	Q	waste	%oos	Q	waste	%oos	Q	waste	%oos
1	3060	0	0	2681	0	0	2931	0	0
2	64	0	1.7%	592	0	4.2%	538	0	2.3%
3	8	192	0	15	124	0	13	299	0
4	204	13	0	64	204	0	107	226	0
5	56	1	0	194	3	0	194	2	0
6	66	30	0.7%	106	19	1.2%	98	36	1%
7	1952	0	0.6%	2211	0	0	2341	0	0
8	159	0	0.2%	76	2	1%	74	2	1.4%
9	1213	0	0	976	35	2.3%	924	100	4.4%
10	135	5	0.1%	165	0	0	199	0	0
11	277	71	0.4%	141	121	0	172	149	0
12	457	0	4%	779	0	4.6%	826	1	4.5%

5 Conclusions

A description is given of a Stochastic Dynamic programming model to generate finite horizon (T periods) production policies for a perishable (lifetime J) product confronted with a non-stationary demand. The properties of the model and the potential of an SDP approach to come to an optimum strategy are investigated.

In an implementation setting of the concept of dynamic programming, we investigated the boundaries of state space, decision space and the discretisation of the state and outcome space of the random process. The iterative solution of the Bellmann equation for each grid point in the state space, implies solving a one-dimensional Global optimisation problem. With a higher variation of demand, this problem becomes more smooth.

The total computing time depends polynomially on the grid density chosen in the $(J - 1)$ -dimensional state space. However, it increases only linearly in the time horizon T of the problem. For an MILP approximation of the policy as suggested in [9], this is more cumbersome, as the computing time increases in principle exponential in the time horizon due the increase in binary variables. A MATLAB implementation of the code requires in the order of magnitude of minutes and optimal strategy for a given time horizon of $T = 12$ periods, that coincides with the practical problem the research is founded on.

Experimenting with the model and the coefficient of variation that quantifies the uncertainty in demand, one can observe that optimal orders have the tendency to cover less periods if uncertainty goes up. Of course also the cost is going up due to more waste, higher safety stocks and more emergency production

orders. The model can be practically used in a rolling horizon setting, where only the first order is carried out and a new plan is generated based on new demand forecasts.

References

1. Bellman, R.: *Dynamic Programming*. Princeton University Press (1957)
2. Blake, J.T., Thompson, S., Smith, S., Anderson, D., Arellano, R., Bernard, D.: Using dynamic programming to optimize the platelet supply chain in nova scotia. In: Dlouhý, M., Prague, C.R.O. (eds.) *Proceedings of the 29th Meeting of the European Working Group on Operational Research Applied to Health Services*, pp. 47–65 (2003)
3. Haijema, R., van Dijk, N.M., van der Wal, J., Smit Sibinga, C.: Blood platelet production with breaks: Optimization by SDP and Simulation. *International Journal of Production Economics* 121, 467–473 (2009), doi:10.1016/j.ijpe.2006.11026
4. Haijema, R., van der Wal, J., van Dijk, N.M.: Blood platelet production: Optimization by dynamic programming and simulation. *Computers and Operations Research* 34(3), 760–779 (2007), doi:10.1016/j.cor.2005.03.023
5. Hendrix, E.M.T., Toth, B.G.: *Introduction to Nonlinear and Global Optimization*. Springer, New York (2010)
6. Howard, R.A.: *Dynamic Programming and Markov Processes*. MIT Press, Cambridge University (1960)
7. Karaesmen, I., Scheller-Wolf, A., Deniz, B.: Planning Production and Inventories in the Extended Enterprise. In: *Managing Perishable and Aging Inventories: Review and Future Research Directions*. International Series in Operations Research & Management Science, vol. 151, ch. 15, pp. 393–436. Springer (2011)
8. Nahmias, S.: Perishable inventory theory: A review. *Operations Research* 30, 680–708 (1982)
9. Pauls-Worm, K.G.J., Rossi, R., Haijema, R., Hendrix, E.M.T.: Non-stationary inventory control for a perishable product. *OR Spektrum* (2012)

An Artificial Fish Swarm Filter-Based Method for Constrained Global Optimization

Ana Maria A.C. Rocha^{1,4}, M. Fernanda P. Costa^{2,3},
and Edite M.G.P. Fernandes⁴

- ¹ Department of Production and Systems, School of Engineering
aarocho@dps.uminho.pt
- ² Department of Mathematics and Applications, School of Sciences
- ³ Mathematics R&D Centre
mfc@math.uminho.pt
- ⁴ Algoritmi R&D Centre
emgpf@dps.uminho.pt
- University of Minho, 4710-057 Braga, Portugal

Abstract. An artificial fish swarm algorithm based on a filter methodology for trial solutions acceptance is analyzed for general constrained global optimization problems. The new method uses the filter set concept to accept, at each iteration, a population of trial solutions whenever they improve constraint violation or objective function, relative to the current solutions. The preliminary numerical experiments with a well-known benchmark set of engineering design problems show the effectiveness of the proposed method.

Keywords: Global optimization, Swarm intelligence, Artificial Fish Swarm, Filter Method.

1 Introduction

In this paper, a new method for handling general nonlinear constraints in a global optimization problem is proposed. The method is based on the implementation of a filter methodology within a population-based swarm intelligence algorithm for solving continuous nonlinear constrained global optimization problems in the form:

$$\begin{aligned} & \underset{x \in \Omega}{\text{minimize}} && f(x) \\ & \text{subject to} && g_j(x) \leq 0, j = 1, \dots, p \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $g_j : \mathbb{R}^n \rightarrow \mathbb{R}$ are nonlinear continuous functions and $\Omega = \{x \in \mathbb{R}^n : -\infty < l_k \leq x_k \leq u_k < +\infty, k = 1, \dots, n\}$. Problems with equality constraints can be reformulated in the above form using a small tolerance. This is a common procedure in stochastic methods for global optimization. Since we do not assume convexity, problem (1) may have several minima and convergence to the global minimum is not guaranteed by some classical gradient-based algorithms. Derivative-free deterministic and stochastic methods are available to

solve global optimization problems, in particular bound constrained problems. Algorithms that use nature inspired or swarm intelligence principles are common in the literature, see for example [5, 14, 16, 21–23, 30, 35, 36, 42]. A recent artificial life computing algorithm that simulates fish swarm behaviors has been used in different contexts [18–20, 39, 41]. The algorithm known as the artificial fish swarm (AFS) algorithm has shown to be competitive with other global solution methods [30].

Most stochastic as well as deterministic methods for global optimization were firstly developed for unconstrained or simple bound constrained problems. Then, they were extended to more general constrained problems by modifying the solution procedures or by applying penalty function methods. Constraint-handling techniques for global optimization can be classified according to the below referred categories.

- Methods based on penalty functions, where the constraint violation is combined with the objective function to define the penalty function that aims at penalizing infeasible solutions [7, 10, 24, 27, 34]. Augmented Lagrangian techniques are particular cases of penalty methods [8, 31, 32].
- Methods based on multi-objective optimization concepts, where both constraint violation and objective function are goals to be minimized separately [1, 3, 17]. The dominance concept of optimality in the multi-objective optimization field is used to accept trial solutions.
- Methods based on biasing feasible over infeasible solutions, where constraint violation and objective function are used separately and optimized by some sort of order being the violation the most important [6, 13, 33, 40, 42].
- Methods that give superiority to feasible solutions, in which feasible solutions are always better than the infeasible ones [13, 22, 29].
- Methods based on preserving feasibility of solutions, where infeasible points are discarded or repaired [9, 42].
- Methods that use an ensemble of (selected) constraint handling techniques, where each technique has its own subpopulation and it is chosen according to the characteristics of the problem to be solved and the stage of the iterative process [26].

Although penalty function methods are probably the most known constraint handling technique, a penalty function depends, in general, on a penalty parameter. Unfortunately, the choice of a suitable value for the penalty parameter is a critical issue since it depends on the optimal solution of the problem. Fletcher and Leyffer [15] proposed a filter method as an alternative to penalty functions to guarantee global convergence in algorithms for nonlinear optimization. This technique incorporates the concept of non-dominance from the multi-objective programming to build a filter that is able to accept solutions if they improve either the objective function or the constraint violation, instead of a linear combination of those two measures. Therefore, the filter replaces the use of penalty functions, avoiding the update of their penalty parameters. The filter methodology has already been used with sequential quadratic programming and interior point methods for solving nonlinear optimization problems, see for example

[11, 12, 15, 28, 37, 38]. Convergence to a local optimal solution, although not necessarily a global one, has been guaranteed whatever the initial approximation. A derivative-free pattern search filter method for nonlinear constrained optimization has already been proposed [4]. However, the therein convergence analysis requires specific problem structure. In the field of global optimization, Hedar and Fukushima present in [17] a hybrid simulated annealing method that uses a filter set concept to accept trial solutions exploring both feasible and infeasible regions.

In this paper, we are particularly interested in using the filter methodology within the AFS algorithm to efficiently solve constrained global optimization problems. The algorithm does not compute or approximate any derivatives or penalty parameters, and will be hereafter denoted by AFSFilter. The method uses the filter set concept to accept, at each iteration, a population of trial solutions whenever they improve constraint violation or objective function relative to the current solutions. This is the first attempt to incorporate the filter methodology into a population-based algorithm to handle the constraints of the problem.

In nature, fishes desire to stay close to the swarm, protecting themselves from predators and looking for food, and to avoid collisions within the group. These behaviors inspire mathematical modelers aiming to solve efficiently optimization problems. The main fish swarm behaviors are the following:

- i) *random* behavior - in general, fish swims randomly in water looking for food and other companions;
- ii) *searching* behavior - this is a basic biological behavior since fish tends to the food; when fish discovers a region with more food, by vision or sense, it goes directly and quickly to that region;
- iii) *swarming* behavior - when swimming, fish naturally assembles in groups which is a living habit in order to guarantee the existence of the swarm and avoid dangers;
- iv) *chasing* behavior - when a fish, or a group of fishes, in the swarm discovers food, the others in the neighborhood find the food dangling quickly after it.

The artificial fish is a fictitious entity of a true fish. Its movements are simulations and interpretations of the above listed fish behaviors [20, 39]. The environment in which the artificial fish moves, searching for the minimum, is the feasible search space of the minimization problem. Considering the problem that is addressed in the paper, the feasible search space is the set $\Xi = \{x \in \mathbb{R}^n : g_j(x) \leq 0, j = 1, \dots, p \text{ and } l_k \leq x_k \leq u_k, k = 1, \dots, n\}$. The position of an artificial fish in the solution space is herein denoted by a point x (a vector in \mathbb{R}^n). We will use the words ‘fish’ and ‘point’ interchangeably throughout the paper.

The organization of the paper is as follows. In Section 2, the filter paradigm is briefly introduced. Section 3 describes the AFS algorithm and presents the details concerning with the use of the filter methodology within the AFS algorithm to handle the constraints of the problem. Section 4 describes the numerical experiments of this preliminary study and Section 5 presents the conclusions and ideas for future work.

Notation: $x^i \in \mathbb{R}^n$ is used to represent the i th point of a population, $x_k^i \in \mathbb{R}$ is the k th ($k = 1, \dots, n$) component of the point x^i of the population, and p_{size} is the number of points in the population. x^{best} is the best point of the population in the sense that it is better than any other point in the population (see Definition [1](#)). The objective function value of the best point of the population is denoted by f^{best} , and x^* is the global optimal solution. The iteration counter in the algorithm is t and t_{max} represents the maximum number of allowed iterations. $\|\cdot\|$ represents the Euclidean norm.

2 Filter Paradigm

The filter paradigm aims at accepting trial solutions of optimization problems if they improve either the constraint violation or the objective function, with respect to current solutions. The methodology appears naturally from the observation that an optimal solution of a nonlinear optimization problem like [1](#) aims at minimizing both constraint violation and objective function values,

$$\theta(x) \doteq \sum_{j=1}^p \max\{0, g_j(x)\} \quad \text{and} \quad f(x), \quad (2)$$

respectively. Thus, problem [1](#) is seen as a bi-objective problem, with two goals, where θ is the objective with the highest priority because we must ensure that $\theta(x^*) = 0$. The methodology uses the concept of non-dominance borrowed from the multi-objective optimization. In this context, a point x^i , or the corresponding pair $(\theta(x^i), f(x^i))$, is dominated by a point x^j , or the corresponding pair $(\theta(x^j), f(x^j))$, if

$$\theta(x^j) \leq \theta(x^i) \quad \text{and} \quad f(x^j) \leq f(x^i).$$

Clearly, a trial solution is considered better than the current solution if it is not dominated by the current solution. The filter \mathcal{F} is defined as a finite set of pairs $(\theta(x^j), f(x^j))$ that correspond to a collection of infeasible solutions x^j such that no filter entry is dominated by any of the others in the filter. The filter defines a forbidden region that does not accept solutions that are dominated by pairs in the current filter. Only solutions that are not dominated by any pair in the filter might be accepted. To avoid acceptance of a trial solution that corresponds to a pair that is arbitrarily close to the border of the filter, the acceptability condition of a solution x to the filter is:

$$\theta(x) \leq (1 - \gamma_\theta) \theta(x^j) \quad \text{or} \quad f(x) \leq f(x^j) - \gamma_f \theta(x^j) \quad (3)$$

for all points x^j in the current filter, i.e., for all filter entries $(\theta(x^j), f(x^j)) \in \mathcal{F}$, where $\gamma_\theta, \gamma_f \in (0, 1)$. A typical filter is shown in Figure [1](#). The shaded area represents the region dominated by the filter entries. According to the conditions in [3](#), all pairs that are below and to the left of the dashed line are acceptable to the filter. When a solution x is added to/included into the filter, then all the entries that are dominated by the new entry $(\theta(x), f(x))$ are removed from the

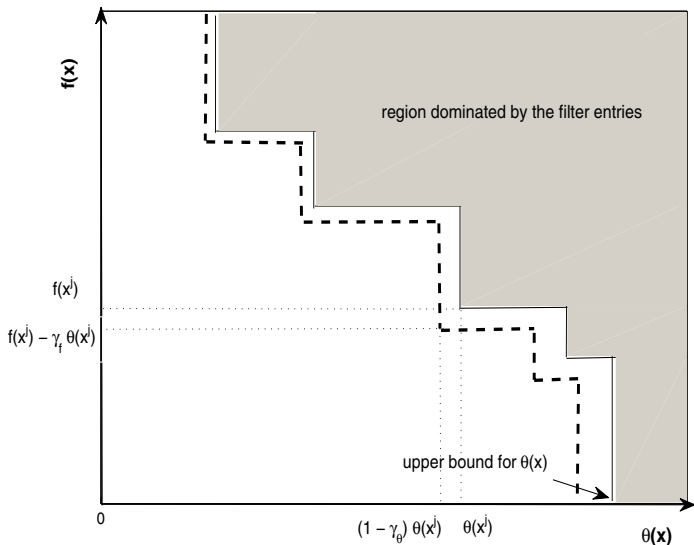


Fig. 1. A filter with four entries

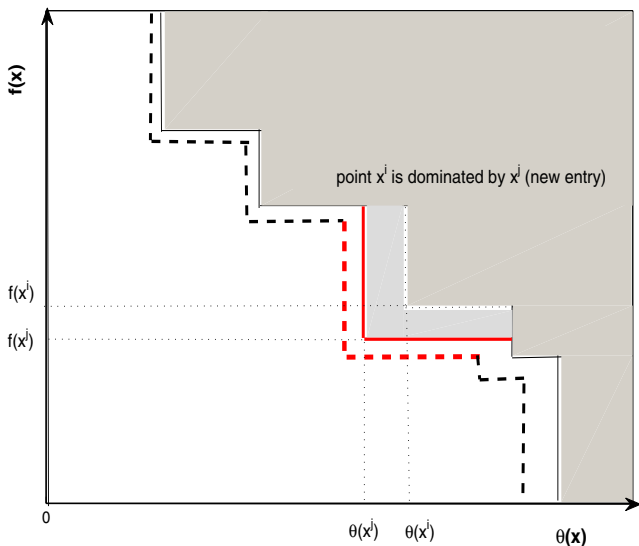


Fig. 2. The entry $(\theta(x^j), f(x^j))$ removes $(\theta(x^i), f(x^i))$

filter. We remark that an inclusion into the filter can occur only when $\theta(x) > 0$. Figure 2 shows that $(\theta(x^i), f(x^i))$ is removed since it is dominated by the new entry $(\theta(x^j), f(x^j))$.

3 The AFSFilter Method

The details concerned with the procedures of the AFS algorithm as well as the implementation of the filter methodology in the proposed population-based method are now described. The AFSFilter method uses:

- i) the artificial fish swarm algorithm to define, at each iteration, random movements and a set of trial solutions;
- ii) the filter methodology to define the acceptability conditions that are able to accept trial solutions according to their constraint violation and objective function values.

Progress towards the optimal solution is assessed by the filter methodology. Here, a global optimal solution $x^* \in \Omega$ such that

$$g_j(x^*) \leq 0, j = 1, \dots, p \text{ and } f(x^*) \leq f(x) \text{ for all } x \neq x^* \in \Omega$$

is to be found. In this context, when comparing points of the population, the following definition is used:

Definition 1. Let x^i and x^j be two points inside Ω . x^i is a better point than x^j if the following condition holds:

$$\theta(x^i) < \theta(x^j) \text{ or } (\theta(x^i) = \theta(x^j) \text{ and } f(x^i) < f(x^j)). \quad (4)$$

To define an appropriate movement for a point x^i in the population, the AFS algorithm relies on a crucial quantity: the ‘visual scope’ of the point. This represents the closed neighborhood with center x^i and radius equal to a positive quantity v . Based on the simple bounds of the variables, in the problem [\(II\)](#), v is defined by

$$v = \varsigma \max_{k \in \{1, \dots, n\}} (u_k - l_k),$$

where ς is a positive visual parameter. Relative to x^i , let

- I^i be the set of indices of the points inside the ‘visual scope’ ($i \notin I^i$),
- n_p^i be the number of points in its ‘visual scope’ ($n_p^i < p_{\text{size}}$).

If the condition $n_p^i / p_{\text{size}} \leq \kappa$ holds, where $\kappa \in (0, 1]$ is the crowd parameter, the ‘visual scope’ of x^i is said to be not crowded. Depending on the relative positions of the points in the population, the below three possible situations may occur.

1. When $n_p^i = 0$, the ‘visual scope’ is empty, and the point x^i , with no other points in its neighborhood to follow, has a *random* behavior.
2. When the ‘visual scope’ is crowded, the point has some difficulty in following any particular point, and has a *searching* behavior. It chooses randomly another point from the ‘visual scope’, hereafter denoted by x^{rand} , and moves towards it if x^{rand} is better than x^i (see condition [\(4\)](#)); otherwise it moves according to a *random* behavior.

3. When the ‘visual scope’ is not crowded, the point firstly tries the *chasing* behavior moving towards the best point inside the ‘visual scope’, denoted by x^{\min} , if this is better than x^i . Otherwise, the point first tries the *swarming* behavior moving towards the central point, c , of the ‘visual scope’. However, if c is not better than x^i , the point follows either a *searching* behavior or a *random* behavior depending on the point x^{rand} being better than x^i or not.

We remark that the described procedures carried out when the ‘visual scope’ is not crowded correspond to a modification that has been introduced in the original version of the AFS algorithm ([20, 39]) and has been shown to outperform the original algorithm, see for example [30, 32].

A simple formal description of the AFSFilter algorithm is presented below in Algorithm 1. The algorithm terminates with a successful message if the best solution found thus far is feasible and is within a certain percentage of accuracy of the best known optimal solution found in the literature, f^{opt} . The algorithm is also allowed to run for a maximum of t_{max} iterations.

Details related with the numerical computations to define point movements and translate the previously referred behaviors will be presented below. We remark that these movements have been devised to maintain the points satisfying the bound constraints of the problem, i.e., inside the set Ω .

Firstly, the initial population is randomly generated in the set Ω . Each point x^i in the population is componentwise computed by

$$x_k^i = l_k + \lambda(u_k - l_k), \text{ for } k = 1, \dots, n,$$

where u_k and l_k are the upper and lower bounds respectively of the set Ω , and λ is an independent uniform random number distributed in the range $[0, 1]$.

For each point x^i of the current population, the trial point y^i is generated according to a direction d and a step size $\alpha \in (0, 1]$,

$$y^i = x^i + \alpha d. \tag{5}$$

The procedure that decides if a trial solution is to be accepted and replaces the current solution is a filter method, as explained later on in Subsection 3.5.

3.1 Random Behavior

When a *random* behavior is invoked, the point x^i moves randomly and

$$y^i = x^i + \alpha \delta \text{ RNG},$$

where δ is a uniformly distributed number between -1 and 1 and RNG is a vector whose components denote the allowed range of movement towards the lower bound l_k , or the upper bound u_k , of the set Ω , for each component k .

3.2 Searching Behavior

When the ‘visual scope’ is crowded, the AFSFilter algorithm tries to follow the *searching* behavior. A point inside the ‘visual scope’ is randomly selected, x^{rand}

Algorithm 1. AFSFilter algorithm

Require: Random population $x^i \in \Omega$, for $i = 1, \dots, p_{\text{size}}$, $t_{\text{max}} > 0$, $0 < \epsilon_1, \epsilon_2 \ll 1$

- 1: Set $t = 0$
- 2: Compute f and θ for all x^i and select x^{best}
- 3: **while** ($|f^{\text{best}} - f^{\text{opt}}| > \epsilon_1 |f^{\text{opt}}| + \epsilon_2$ or $\theta^{\text{best}} > \epsilon_2$) and $t \leq t_{\text{max}}$ **do**
- 4: **for all** x^i **do**
- 5: Compute ‘visual scope’
- 6: **if** ‘visual scope’ is empty **then**
- 7: *Random* behavior
- 8: **else**
- 9: **if** ‘visual scope’ is crowded **then**
- 10: Select randomly x^{rand} from the ‘visual scope’
- 11: **if** x^{rand} is better than x^i **then**
- 12: *Searching* behavior
- 13: **else**
- 14: *Random* behavior
- 15: **end if**
- 16: **else**
- 17: **if** x^{min} is better than x^i **then**
- 18: *Chasing* behavior
- 19: **else**
- 20: Compute c
- 21: **if** c is better than x^i **then**
- 22: *Swarming* behavior
- 23: **else**
- 24: Select randomly x^{rand} from the ‘visual scope’
- 25: **if** x^{rand} is better than x^i **then**
- 26: *Searching* behavior
- 27: **else**
- 28: *Random* behavior
- 29: **end if**
- 30: **end if**
- 31: **end if**
- 32: **end if**
- 33: **end if**
- 34: *Filter line search* to decide if the trial point y^i is accepted
- 35: **end for**
- 36: Set $t = t + 1$
- 37: **end while**

($\text{rand} \in I^i$), and the point x^i is moved towards it to generate a trial point y^i if x^{rand} is better than x^i (see (4)). Here, the direction of movement to get y^i is defined as $d = x^{\text{rand}} - x^i$, recall (5). Otherwise, the point x^i follows a *random* behavior as previously described in Subsection 3.1

3.3 Chasing Behavior

According to Algorithm 1, *chasing* behavior is performed if the ‘visual scope’ is not crowded and the best point inside the ‘visual scope’ of x^i , x^{\min} , is better than x^i , in the sense of Definition 1. The direction to carry out the movement towards a trial point y^i is defined by $d = x^{\min} - x^i$.

3.4 Swarming Behavior

When the ‘visual scope’ of a point x^i is not crowded and x^{\min} is not better than x^i , the central point inside the ‘visual scope’ of x^i is computed by

$$c = \sum_{j \in I^i} x^j / n_p^i,$$

and compared with x^i . If c is better than x^i , then the *swarming* behavior follows and the corresponding trial point is computed using (5) where the direction of the movement is $d = c - x^i$. Otherwise, the *searching* behavior is tried (see Subsection 3.2).

3.5 The Filter Line Search Method

Here, we aim to show how the methodology of a filter as outlined in [15] can be adapted to this population-based AFS method. Each entry in the filter is defined by two components: $\theta(x)$ that aims to measure feasibility and $f(x)$ that measures optimality, as previously defined in (2). The proposed paradigm defines just one filter. After a search direction d has been computed, a step size should be determined, using (5), in a way that sufficient progress towards the optimal solution is obtained.

Backtracking Line Search. The step size $\alpha \in (0, 1]$ is determined by a backtracking line search technique. A decreasing sequence of step sizes $\{\alpha_j\}$ with $\lim_j \alpha_j = 0$ is tried, until a set of acceptance conditions are satisfied. This j denotes the iteration counter for the inner loop. A trial step size α_j can be accepted if the corresponding trial point $y^i = x^i + \alpha_j d$ is acceptable by the filter. When $\alpha_j > \alpha_{\min}$, the point y^i might be acceptable if sufficient progress in one of the two measures, relative to the value at the current point x^i , holds:

$$\theta(y^i) \leq (1 - \gamma_\theta) \theta(x^i) \text{ or } f(y^i) \leq f(x^i) - \gamma_f \theta(x^i). \quad (6)$$

However, when the current solution is (almost) feasible, in practice, if $\theta(x^i) \leq \theta_{\min}$, the trial point has to satisfy only the condition

$$f(y^i) \leq f(x^i) - \gamma_f \theta(x^i) \quad (7)$$

to be acceptable, in order to prevent convergence to feasible but non-optimal solutions, where $0 < \theta_{\min} \ll 1$. In particular, the corresponding solution/iteration

is denoted by f -type. To prevent cycling between points that improve either θ or f , at each iteration t , the algorithm maintains the filter \mathcal{F}_t that contains pairs (θ, f) that are prohibited for a successful trial point at iteration t . Thus, during the line search procedure, a trial point y^i is acceptable only if $(\theta(y^i), f(y^i)) \notin \mathcal{F}_t$.

The filter is initialized to $\mathcal{F}_0 \subseteq \{(\theta, f) \in \mathbb{R}^2 : \theta \geq \theta_{\max}\}$, where $\theta_{\max} > 0$ is the upper bound on θ , and later is augmented using the formula

$$\mathcal{F}_{t+1} = \mathcal{F}_t \cup \{(\theta, f) \in \mathbb{R}^2 : \theta > (1 - \gamma_\theta)\theta(x^i) \text{ and } f > f(x^i) - \gamma_f\theta(x^i)\}$$

only after every iteration in which the trial point satisfies (6). Since an inclusion of a point into the filter can occur only when $\theta > 0$, an f -type trial solution is never included into the filter.

Restoration Step. When it is not possible to find a step size $\alpha_j > \alpha_{\min}$ that satisfy one of the above referred conditions, the algorithm reverts to a restoration step. In this case, an approximate descent direction for θ or f , at x^i , is computed. The procedure that defines the descent direction is the following. Two exploring points e^1, e^2 are randomly generated in a small neighborhood of the point x^i and an approximate descent direction for θ or f is defined [17, 29]. When the current point x^i is feasible, the direction d is descent for f at x^i ; otherwise it is descent for θ at x^i . Thus

$$\Delta^j = \begin{cases} f(e^j) - f(x^i), & \text{if } \theta(x^i) \leq \theta_{\min} \\ \theta(e^j) - \theta(x^i), & \text{otherwise} \end{cases}$$

where $\|e^j - x^i\| \leq \varepsilon$ for $j = 1, 2$ and a very small positive constant ε , and

$$d = -\frac{1}{\sum_{k=1}^2 |\Delta^k|} \sum_{j=1}^2 \Delta^j \frac{e^j - x^i}{\|e^j - x^i\|}. \quad (8)$$

Based on the direction (8), the trial solution $y^i = x^i + d$ is accepted if it remains inside Ω ; otherwise a projection onto Ω is carried out.

4 Numerical Experiments

In this section, the numerical results of a preliminary study running a benchmark set of engineering problems are reported. A comparison with the results available in the literature is also included. The six chosen engineering design problems are the most common in the literature.

- The welded beam design problem has four design variables and seven inequality constraints [6, 17, 29, 34, 40]. The objective is to minimize the cost of a welded beam, subject to the constraints on the shear stress, bending stress, buckling load on the bar, end deflection of the beam and side constraints.
- In the speed reducer design problem [6, 29, 40], the weight of the speed reducer is to be minimized subject to the constraints on bending stress of the gear teeth, surface stress, transverse deflections of the shafts and stress in the shafts. The problem has seven variables and 11 inequality constraints.

- The tension/compression spring design problem has three continuous variables and four constraints [6, 17, 29, 40], and aims to minimize the weight of a tension/compression spring.
- The 3-bar truss design aims to minimize the volume of the truss subject to the stress constraints [6, 29, 40]. This problem has two design variables representing the cross-sectional areas of two bars (two identical of three-bar) and three inequality constraints.
- In the tubular column design, the cost of fabrication is to be minimized [6, 29]. This problem has two design variables with two inequality constraints.
- The cylindrical vessel design [6, 17, 24, 29, 34] (with both ends capped with a hemispherical head) is to minimize the total cost of fabrication. The problem has four design variables (two of them are multiples of 0.0625) and four inequality constraints.

The C++ programming language is used in this real-coded algorithm. The computational tests were performed on a PC with a 2.8 GHz Core Duo Processor P9700 and 6 Gb of memory. The size of the population is defined as $p_{\text{size}} = \min\{50, 5n\}$. Table 1 displays the results of the comparative tests. Since the algorithm relies on some random parameters and variables, we solve each problem 30 times. The best of the solutions found in all 30 runs is denoted in the table by ‘sol.’, ‘S.D.’ gives the standard deviation of the obtained objective function values and ‘n.f.e.’ gives the average number of function evaluations from all the runs. The user defined parameters are set as follows: $\theta_{\max} = 10^4$, $\theta_{\min} = 10^{-6}$, $\gamma_{\theta} = \gamma_f = 10^{-8}$, $\alpha_{\min} = 10^{-3}$, $\kappa = 0.8$ and $\varepsilon = 10^{-3}$. The parameter ζ is not fixed over the iterative process. Initially, is set to one and is reduced until it reaches 0.1. The parameters for the termination criteria are $\epsilon_1 = 10^{-4}$, $\epsilon_2 = 10^{-6}$ and $t_{\max} = 200$.

A comparison with some of the results in the literature follows. Two variants, a modified differential evolution (mDE-r) based on competitive ranking [6] and the differential evolution based on adaptive penalty (DE-AP) [34], are used. The other selected methods for this comparison are: feasibility and dominance rules based on a sufficient reduction of constraint violation or objective function values are implemented with a hybrid electromagnetism-like (HEM) algorithm [29]; a hybrid evolutionary algorithm (HEA) implemented with an adaptive constraint-handling technique [40]; a socio-behavioural (SB) model [2]; an improved harmony search (iHS) proposed by Mahdavi et al. [25]; and finally, a filter method implemented in a simulated annealing algorithm context in [17]. From the preliminary results in Table 1, we may conclude that the performance of the AFSFilter algorithm is comparable to the other ones. For the 3-bar truss and tubular problems, the AFSFilter obtained competitive solutions with reduced function evaluations. The proposed method also converges to a solution of the beam problem that is better than those obtained by SB and HEM, and the solution obtained for the speed problem is also better than the one obtained by SB. We remark that further research is required in the AFSFilter in order to improve the convergence to the solutions with high accuracy.

Table 1. Comparative results

Method	Problems						
		Beam	Speed	Spring	3-bar truss	Tubular	Vessel
AFSFilter	sol.	2.382927	2999.151	0.012667	263.8964	26.53351	5946.636 [†]
	S.D.	1.3E-2	2.1E00	7.2E-6	2.5E-3	5.2E-3	5.5E+1
	n.f.e.	38342	65779	23636	5388	9223	45287
SB [2]	sol.	2.4426	3008.08	-	-	-	6171.00
	S.D.	-	-	-	-	-	-
	n.f.e.	19259	19154	-	-	-	12630
mDE-r [6]	sol.	2.380810	2994.320	0.012664	263.8919	26.5311	6059.525
	S.D.	-	-	-	-	-	-
	n.f.e.	30000	35000	15000	10000	10000	30000
FSA [17]	sol.	2.381065	-	0.012665	-	-	5868.765 [†]
	S.D.	-	-	2.2E-8	-	-	2.6E+2
	n.f.e.	56243	-	49531	-	-	108883
iHS [25]	sol.	1.7248 [‡]	-	0.012671	-	-	5849.762 [†]
	S.D.	-	-	-	-	-	-
	n.f.e.	200000	-	30000	-	-	-
HEM [29]	sol.	2.386269	2995.804	0.012667	263.8960	26.53227	6072.232
	S.D.	3.1E-2	1.3E00	8.0E-6	4.9E-5	3.5E-3	5.3E+1
	n.f.e.	28650	51989	9605	17479	25136	20993
DE-AP [34]	sol.	2.38113	-	-	-	-	6059.718
	S.D.	0.0E00	-	-	-	-	0.0E00
	n.f.e.	40000	-	-	-	-	80000
HEA [40]	sol.	2.380957	2994.499	0.012665	263.8958	-	-
	S.D.	1.3E-5	7.0E-2	1.4E-9	4.9E-5	-	-
	n.f.e.	30000	40000	24000	15000	-	-

[†] all variables are considered continuous; - not available
[‡] slightly different problem formulation.

5 Conclusions

In this paper, a filter line search method is implemented in a population-based swarm intelligence algorithm to solve continuous nonlinear constrained global optimization problems. The innovative nature of the work is focused on the integration of the filter methodology, as a constraint-handling technique, into an artificial fish swarm algorithm. Based on a population of current solutions, the AFS algorithm computes trial solutions using search directions and a step size in a way that sufficient progress towards the optimal solution is obtained. Trial solutions are acceptable only if they improve either the constraint violation or the objective function relative to the current solutions, and are acceptable by the filter.

The preliminary numerical results on six common benchmark engineering design optimization problems show that the proposed AFSFilter method is competitive when compared with other stochastic methods for global optimization, thus encouraging the application of this AFS filter-based paradigm to more complex

problems, such as those with discrete variables. We think that the algorithm efficiency can be improved through proper tuning of the algorithm parameters. The impact of some parameters on the performance of the algorithm will be analyzed in the future. An elitist procedure will be implemented in a way that the best point of the population will be maintained regardless being dominated by a trial point at a particular iteration. It has been also observed that accuracy could be improved with an intensification search around the best found solution, at the final stage of the algorithm. This local search aims at improving the final solution at a reduced computational cost and is a common procedure in global optimization methods.

Acknowledgments. The financial support from FEDER COMPETE (Programa Operacional Fatores de Competitividade / Operational Programme Thematic Factors of Competitiveness) and FCT (Fundação para a Ciência e a Tecnologia / Portuguese Foundation for Science and Technology) Projects FCOMP-01-0124-FEDER-022674 and PEst-C/MAT/UI0013/2011 is gratefully acknowledged.

References

1. Aguirre, A.H., Rionda, S.B., Coello Coello, C.A., Lizárraga, G.L., Montes, E.M.: Handling constraints using multiobjective optimization concepts. *International Journal for Numerical Methods in Engineering* 59, 1989–2017 (2004)
2. Akhtar, S., Tai, K., Tay, T.: A socio-behavioural simulation model for engineering design optimization. *Engineering Optimization* 34, 341–354 (2002)
3. Ali, M.M., Golalikhani, M.: An electromagnetism-like method for nonlinearly constrained global optimization. *Computers and Mathematics with Applications* 60, 2279–2285 (2010)
4. Audet, C., Dennis Jr., J.E.: A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization* 14(4), 980–1010 (2004)
5. Azad, M.A.K., Fernandes, E.M.G.P., Rocha, A.M.A.C.: Nonlinear continuous global optimization by modified differential evolution. In: Simos, T.E., et al. (eds.) *International Conference of Numerical Analysis and Applied Mathematics 2010*, vol. 1281, pp. 955–958 (2010)
6. Azad, M. A.K., Fernandes, E.M.G.P.: Modified Differential Evolution Based on Global Competitive Ranking for Engineering Design Optimization Problems. In: Murgante, B., Gervasi, O., Iglesias, A., Tanar, D., Apduhan, B.O. (eds.) *ICCSA 2011, Part III. LNCS*, vol. 6784, pp. 245–260. Springer, Heidelberg (2011)
7. Barbosa, H.J.C., Lemonge, A.C.C.: An adaptive penalty method for genetic algorithms in constrained optimization problems. In: Iba, H. (ed.) *Frontiers in Evolutionary Robotics*, pp. 9–34. I-Tech Education Publ., Austria (2008)
8. Birgin, E.G., Floudas, C.A., Martinez, J.M.: Global minimization using an augmented Lagrangian method with variable lower-level constraints. *Mathematical Programming* 125, 139–162 (2010)
9. Chootinan, P., Chen, A.: Constrained handling in genetic algorithms using a gradient-based repair method. *Computers and Operations Research* 33, 2263–2281 (2006)
10. Coello Coello, C.A.: Use of a self-adaptive penalty approach for engineering optimization problems. *Computers in Industry* 41, 113–127 (2000)

11. Costa, M.F.P., Fernandes, E.M.G.P.: Assessing the potential of interior point barrier filter line search methods: nonmonotone versus monotone approach. *Optimization* 60(10-11), 1251–1268 (2011)
12. Costa, M.F.P., Fernandes, E.M.G.P.: On Minimizing Objective and KKT Error in a Filter Line Search Strategy for an Interior Point Method. In: Murgante, B., Gervasi, O., Iglesias, A., Tanar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part III. LNCS, vol. 6784, pp. 231–244. Springer, Heidelberg (2011)
13. Deb, K.: An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering* 186, 311–338 (2000)
14. Fernandes, E.M.G.P., Martins, T.F.M.C., Rocha, A.M.A.C.: Fish swarm intelligent algorithm for bound constrained global optimization. In: Aguiar, J.V. (ed.) CMMSE 2009, pp. 461–472 (2009)
15. Fletcher, R., Leyffer, S.: Nonlinear programming without a penalty function. *Mathematical Programming* 91, 239–269 (2002)
16. Hedar, A.-R., Fukushima, M.: Heuristic pattern search and its hybridization with simulated annealing for nonlinear global optimization. *Optimization Methods and Software* 19, 291–308 (2004)
17. Hedar, A.-R., Fukushima, M.: Derivative-free filter simulated annealing method for constrained continuous global optimization. *Journal of Global Optimization* 35, 521–549 (2006)
18. Gao, X.Z., Wu, Y., Zenger, K., Huang, X.: A knowledge-based artificial fish-swarm algorithm. In: 13th IEEE International Conference on Computational Science and Engineering, pp. 327–332 (2010)
19. Jiang, M., Mastorakis, N., Yuan, D., Lagunas, M.A.: Image segmentation with improved artificial fish swarm algorithm. In: Mastorakis, N., et al. (eds.) ECC 2008. LNEE, vol. 28, pp. 133–138 (2009)
20. Jiang, M., Wang, Y., Pfletschinger, S., Lagunas, M.A., Yuan, D.: Optimal Multiuser Detection with Artificial Fish Swarm Algorithm. In: Huang, D.-S., et al. (eds.) ICIC 2007. CCIS, vol. 2, pp. 1084–1093. Springer, Heidelberg (2007)
21. Kaelo, P., Ali, M.M.: A numerical study of some modified differential evolution algorithms. *European Journal of Operational Research* 169, 1176–1184 (2006)
22. Karaboga, D., Basturk, B.: Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 789–798. Springer, Heidelberg (2007)
23. Karimi, A., Nobahari, H., Siarry, P.: Continuous ant colony system and tabu search algorithms hybridized for global minimization of continuous multi-minima functions. *Computational Optimization and Applications* 45, 639–661 (2010)
24. Liu, J.-L., Lin, J.-H.: Evolutionary computation of unconstrained and constrained problems using a novel momentum-type particle swarm optimization. *Engineering Optimization* 39, 287–305 (2007)
25. Mahdavi, M., Fesanghary, M., Damangir, E.: An improved harmony search algorithm for solving optimization problems. *Applied Mathematics and Computation* 188, 1567–1579 (2007)
26. Mallipeddi, R., Suganthan, P.N.: Ensemble of constraint handling techniques. *IEEE Transactions on Evolutionary Computation* 14, 561–579 (2010)
27. Petalas, Y.G., Parsopoulos, K.E., Vrahatis, M.N.: Memetic particle swarm optimization. *Annals of Operations Research* 156, 99–127 (2007)
28. Pereira, A.I., Costa, M.F.P., Fernandes, E.M.G.P.: Interior point filter method for semi-infinite programming problems. *Optimization* 60(10-11), 1309–1338 (2011)

29. Rocha, A.M.A.C., Fernandes, E.M.G.P.: Hybridizing the electromagnetism-like algorithm with descent search for solving engineering design problems. *International Journal of Computer Mathematics* 86, 1932–1946 (2009)
30. Rocha, A.M.A.C., Fernandes, E.M.G.P., Martins, T.F.M.C.: Novel Fish Swarm Heuristics for Bound Constrained Global Optimization Problems. In: Murgante, B., Gervasi, O., Iglesias, A., Tanar, D., Apduhan, B.O. (eds.) *ICCSA 2011, Part III*. LNCS, vol. 6784, pp. 185–199. Springer, Heidelberg (2011)
31. Rocha, A.M.A.C., Fernandes, E.M.G.P.: Numerical study of augmented Lagrangian algorithms for constrained global optimization. *Optimization* 60(10–11), 1359–1378 (2011)
32. Rocha, A.M.A.C., Martins, T.F.M.C., Fernandes, E.M.G.P.: An augmented Lagrangian fish swarm based method for global optimization. *Journal of Computational and Applied Mathematics* 235(16), 4611–4620 (2011)
33. Runarsson, T.P., Yao, X.: Stochastic ranking for constrained evolutionary optimization. *IEEE Transaction on Evolutionary Computation* 4, 284–294 (2000)
34. Silva, E.K., Barbosa, H.J.C., Lemonge, A.C.C.: An adaptive constraint handling technique for differential evolution with dynamic use of variants in engineering optimization. *Optimization and Engineering* 12, 31–54 (2011)
35. Socha, K., Dorigo, M.: Ant colony optimization for continuous domains. *European Journal of Operational Research* 185, 1155–1173 (2008)
36. Stanoyevitch, A.: Homogeneous genetic algorithms. *International Journal of Computer Mathematics* 87, 476–490 (2010)
37. Ulbrich, M., Ulbrich, S., Vicente, L.N.: A globally convergent primal-dual interior-point filter method for nonlinear programming. *Mathematical Programming* 100, 379–410 (2004)
38. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* 106, 25–57 (2006)
39. Wang, C.-R., Zhou, C.-L., Ma, J.-W.: An improved artificial fish-swarm algorithm and its application in feed-forward neural networks. In: *Proceedings of the 4th ICMLC*, pp. 2890–2894 (2005)
40. Wang, Y., Cai, Z., Zhou, Y., Fan, Z.: Constrained optimization based on hybrid evolutionary algorithm and adaptive constraint-handling technique. *Structural and Multidisciplinary Optimization* 37(4), 395–413 (2009)
41. Wang, X., Gao, N., Cai, S., Huang, M.: An Artificial Fish Swarm Algorithm Based and ABC Supported QoS Unicast Routing Scheme in NGL. In: Min, G., Di Martino, B., Yang, L.T., Guo, M., Rünger, G. (eds.) *ISPA Workshops 2006*. LNCS, vol. 4331, pp. 205–214. Springer, Heidelberg (2006)
42. Zahara, E., Hu, C.-H.: Solving constrained optimization problems with hybrid particle swarm optimization. *Engineering Optimization* 40(11), 1031–1049 (2008)

Solving Multidimensional 0–1 Knapsack Problem with an Artificial Fish Swarm Algorithm

Md. Abul Kalam Azad², Ana Maria A.C. Rocha^{1,2},
and Edite M.G.P. Fernandes²

¹ Department of Production and Systems

² Algoritmi R&D Centre

University of Minho, 4710-057 Braga, Portugal
{akazad, arocha, emgpf}@dps.uminho.pt

Abstract. The multidimensional 0–1 knapsack problem is a combinatorial optimization problem, which is NP-hard and arises in many fields of optimization. Exact as well as heuristic methods exist for solving this type of problem. Recently, a population-based artificial fish swarm algorithm was proposed and applied in an engineering context. In this paper, we present a binary version of the artificial fish swarm algorithm for solving multidimensional 0–1 knapsack problem. Infeasible solutions are made feasible by a decoding algorithm. We test the presented method with a set of benchmark problems and compare the obtained results with other methods available in literature. The tested method appears to give good results when solving these problems.

Keywords: 0–1 knapsack problem, multiple constraints, artificial fish swarm, decoding algorithm.

1 Introduction

Generally, the multidimensional 0–1 knapsack problem is a combinatorial optimization problem that can be formulated as follows:

$$\begin{aligned} & \text{maximize } z(\mathbf{x}) \equiv \mathbf{c}\mathbf{x} \\ & \text{subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \quad x_j \in \{0, 1\}, \quad j = 1, 2, \dots, n, \end{aligned} \tag{1}$$

where $\mathbf{c} = (c_1, c_2, \dots, c_n)$ is an n -dimensional row vector of profits, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ is an n -dimensional column vector of 0–1 decision variables, $\mathbf{A} = [a_{k,j}]$, $k = 1, 2, \dots, m$, $j = 1, 2, \dots, n$ is an $m \times n$ coefficient matrix of resources, and $\mathbf{b} = (b_1, b_2, \dots, b_m)^T$ is an m -dimensional column vector of resource capacities. It should be noted here that, in a multidimensional 0–1 knapsack problem, each element of \mathbf{c} , \mathbf{A} and \mathbf{b} is assumed to be nonnegative. The goal is to find a subset of n items that yields maximum profit z without exceeding resource capacities \mathbf{b} . The much simpler case with a single constraint ($m = 1$) is known as the single knapsack problem and effective approximate algorithms

have been developed for obtaining its near optimal solutions. The general case corresponding to $m \geq 2$ is known as the multidimensional 0–1 knapsack problem (MKP), which is NP-hard.

Many practical problems can be formulated as the MKP, such as the capital budgeting problem, resource allocation with financial constraints, allocating processors and databases in a distributed computer system, the project selection and cargo loading, cutting stock problems and so on. In the last decades many exact as well as heuristic methods have been proposed to solve MKP. Exact methods include dynamic programming methods [2,16,38], branch-and-bound algorithms [13,15,32], the Fourier-Motzkin elimination based enumeration algorithms [6], asymptotic analysis method [31], statistical analysis method [12], linked LP-relaxations, disjunctive cuts and implicit enumeration [33]. On the other hand metaheuristic methods include simulated annealing [10], tabu search [3,18,34], genetic algorithm [7,9,23], ant colony optimization [24]. Other heuristic methods have been proposed in [1,5,19,20,35]. Li et al. [25] proposed a genetic algorithm based on orthogonal design (OGA). In their method, the authors used a check-and-repair operator to make feasible solutions. Sakawa and Kato [30] proposed a genetic algorithm with double strings (GADS) based on the decoding algorithm. Deep and Bansal [8] proposed a socio-cognitive particle swarm optimization (SCPSO) based on the penalty function. A nice review of different solution methods for solving multidimensional 0–1 knapsack problem is found in [14].

Recently, a population-based artificial fish swarm algorithm that simulates the behavior of the fish swarm inside water was proposed and applied in an engineering context [21,22,36,37]. Rocha et al. [28,29] proposed an augmented Lagrangian fish swarm based method for global optimization problems and a novel fish swarm heuristic for bound constrained global optimization problems. Applying to the optimization problem, generally a ‘fish’ represents an individual point in a population. The fish swarm movements seem randomly defined and yet they are objectively synchronized. Fishes desire to stay close to the swarm, to protect themselves from predators and to look for food, and to avoid collisions within the group. Inspired by these behavior, researchers aim to solve optimization problems in an efficient manner. The behavioral model-based optimization algorithms seek to imitate, as well as to make variations on the swarm behavior in nature, and to create new types of abstract movements. The fish swarm behavior inside water may be summarized as follows [28,29]:

- i) random behavior – in general, fish swims randomly in water looking for food and other companions;
- ii) chasing behavior – when a fish, or a group of fishes, in the swarm discovers food, the others in the neighborhood find the food dangling quickly after it;
- iii) swarming behavior – when swimming, fish naturally assembles in groups which is a living habit in order to guarantee the existence of the swarm and avoid dangers;

- iv) searching behavior – this is a basic biological behavior since fish tends to the food, when fish discovers a region with more food, by vision or sense, it goes directly and quickly to that region;
- v) leaping behavior – when fish stagnates in a region, it leaps to look for food in other regions.

The artificial fish is a fictitious entity of a true fish. Its movements are simulations and interpretations of the above listed fish behavior [22,28,29]. The environment in which the artificial fish moves, searching for the optimum, is the feasible search space of the problem. Based on the artificial fish swarm algorithm for global optimization, we propose a binary version of the artificial fish swarm algorithm for solving multidimensional 0–1 knapsack problem (II).

The organization of this paper is as follows. We briefly describe the artificial fish swarm algorithm in Section 2. In Section 3 the proposed binary artificial fish swarm algorithm is outlined. Section 4 describes the experimental results and finally we draw the conclusions of this study in Section 5.

2 Artificial Fish Swarm Algorithm

In this section we will give a brief description of the artificial fish swarm algorithm (AFSA) proposed by Rocha et al. [29] for box constrained global optimization problems of the form minimize $_{\mathbf{x} \in \Omega} f(\mathbf{x})$. Here $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a nonlinear function that is to be minimized and $\Omega = \{\mathbf{x} \in \mathbb{R}^n : l_j \leq x_j \leq u_j, j = 1, 2, \dots, n\}$ is the search space. l_j and u_j are the lower and upper bounds of x_j and n is the number of variables of the optimization problem.

The artificial fish swarm algorithm uses a population of N individual points (the fishes) $\mathbf{x}^i, i = 1, 2, \dots, N$ to identify promising regions looking for a global solution [36]. \mathbf{x}^i is a floating-point representation that covers the entire search space Ω . The crucial issue of AFSA is the ‘visual scope’ of each individual point \mathbf{x}^i . This represents a closed neighborhood of \mathbf{x}^i with a radius equal to a positive quantity ν defined by

$$\nu = \delta \max_{j \in \{1, 2, \dots, n\}} (u_j - l_j)$$

where $0 < \delta < 1$ is a positive visual parameter. In general, this parameter is maintained fixed over the iterative process. However, experiments show that a slow reduction accelerates the convergence to the solution [11]. Let \mathbf{I}^i be the set of indices of the points inside the ‘visual scope’ of point \mathbf{x}^i , where $i \notin \mathbf{I}^i$ and $\mathbf{I}^i \subset \{1, 2, \dots, N\}$, and let np^i be the number of points in its ‘visual scope’. Depending on the relative positions of the points in the population, three possible situations may occur:

- a) when $np^i = 0$, the ‘visual scope’ is empty, and the point \mathbf{x}^i , with no other points in its neighborhood to follow, moves randomly for a better region;
- b) when the ‘visual scope’ is not crowded, the point \mathbf{x}^i is able either to swarm moving towards the central or to chase moving towards the best point inside the ‘visual scope’;

- c) when the ‘visual scope’ is crowded, the point \mathbf{x}^i has some difficulty in following any particular point, and searches for a better region choosing randomly another point (from the ‘visual scope’) and moves towards it;

The condition that decides when the ‘visual scope’ of \mathbf{x}^i is not crowded is

$$C_f \equiv \frac{np^i}{N} \leq \theta, \tag{2}$$

where C_f is the crowding factor and $\theta \in (0, 1)$ is the crowd parameter. In this situation, the point \mathbf{x}^i has the ability to swarm or to chase. We refer to [28,29,36,37] for some more details.

3 Binary Artificial Fish Swarm Algorithm

In this section, a binary version of the artificial fish swarm algorithm to solve the multidimensional 0–1 knapsack problem (II), herein denoted by b-AFSA, is presented. The outline of the algorithm is briefly described in the following.

3.1 Initialization (Coding)

The first step of designing a b-AFSA for solving 0–1 MKP is to devise a suitable representation scheme of an individual point in a population. Since we consider 0–1 knapsack problem, N individual points are randomly initialized, each represented by a binary 0/1 string of length n [17,26]. For example, for $n = 15$, an individual point \mathbf{x}^i , $i = 1, 2, \dots, N$ randomly initialized at iteration $t = 1$ is shown in Fig. I.

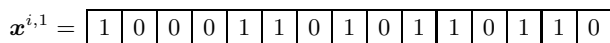


Fig. 1. Individual representation in b-AFSA

3.2 Constraints Handling (Decoding)

The randomly initialized point \mathbf{x}^i may not be feasible since the problem (II) has m constraints.

The widely used approach to deal with constrained optimization problem is based on penalty functions where a penalty term is added to the objective function in order to penalize the constraint violation. This enable us to transform a constrained optimization problem into a sequence of unconstrained subproblems. The penalty function method can be applied to any type of constraints, but the performance of penalty-type method is not always satisfactory due to the choice of an appropriate penalty parameter. Although several ideas have been proposed about how the penalty function is designed and applied to infeasible solutions, it is generally recognized that the smaller the feasible region, the harder it is for

the penalty function methods to generate feasible solutions, as pointed out in the field of nonlinear optimization [30]. For this reason alternative constraints handling techniques have been proposed in the last decades.

There are a number of standard ways of dealing with constraints and infeasible solutions in binary represented population-based solution methods. In b-AFSA, the decoding algorithm proposed by Sakawa and Kato [30] to make infeasible solutions to feasible ones is used. Although the individual point representations are different in GADS and b-AFSA, we modify the decoding algorithm so that it can decode individual points in a population in the same way as [30]. The advantage of this decoding algorithm is that decoding a point \mathbf{x}^i starts from any index and randomly continues until the maximum length of string n is reached to make the point \mathbf{x}^i feasible, aiming to obtain promising solution (and hopefully optimal). Another decoding algorithm which starts from the beginning of index and sequentially continues can be applied but the obtained solution may not be optimal.

3.3 Visual Scope in b-AFSA

Since the Euclidean distance cannot give the actual distance between two points represented by binary 0/1 bits, the Hamming distance H_d is used to calculate the ‘visual scope’, in b-AFSA. The Hamming distance of two points of equal bits length is the number of positions at which the corresponding bits are different. Figure 2 illustrates an example with two binary points. We may observe that the Hamming distance is equal to seven. After calculating the Hamming distance

$$\mathbf{x}^1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{x}^2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$H_d(\mathbf{x}^1, \mathbf{x}^2) = 7$$

Fig. 2. Hamming distance between two binary points in b-AFSA

between all pair of points in the population, the np^i points inside the ‘visual scope’ of \mathbf{x}^i are identified as the points \mathbf{x}^j that satisfy the condition $H_d(\mathbf{x}^i, \mathbf{x}^j) \leq \nu$, for $j \in \{1, \dots, N\}$, $j \neq i$, where

$$\nu = \delta \times n. \quad (3)$$

In (3), n (number of variables) represents the maximum Hamming distance between two binary points in b-AFSA. After computing np^i , the crowding factor C_f of \mathbf{x}^i is calculated using (2).

Depending on the value of C_f , the ‘visual scope’ can be empty, not crowded or crowded. In b-AFSA, the fish (point) behavior that create the trial points are outlined as follows.

Chasing Behavior: If the ‘visual scope’ is not crowded and the point that has the best objective function value inside the ‘visual scope’, denoted by \mathbf{x}^{best} (best $\in I^i$), satisfies $z(\mathbf{x}^{\text{best}}) > z(\mathbf{x}^i)$, the chasing behavior is to be implemented. In b-AFSA, crossover (discussed in Section 3.4) between \mathbf{x}^i and \mathbf{x}^{best} is performed to create the trial point \mathbf{y}^i .

Swarming Behavior: When the ‘visual scope’ is not crowded and $z(\mathbf{x}^{\text{best}}) \leq z(\mathbf{x}^i)$, a random point \mathbf{x}^{rand} (rand $\in I^i$) is selected inside the ‘visual scope’, and if $z(\mathbf{x}^{\text{rand}}) > z(\mathbf{x}^i)$, the point \mathbf{x}^i performs the swarming behavior. In b-AFSA, one position mutation (discussed in Section 3.4) is used to the point \mathbf{x}^i to create the trial point \mathbf{y}^i .

Searching Behavior: The searching behavior is tried in the following situations:

- a) when the ‘visual scope’ is not crowded and neither \mathbf{x}^{best} nor \mathbf{x}^{rand} improve in objective function value;
- b) when the ‘visual scope’ is crowded.

Here, a point \mathbf{x}^{rand} inside the ‘visual scope’ is randomly selected and the point \mathbf{x}^i moves towards it if it is improving in objective function value, i.e. $z(\mathbf{x}^{\text{rand}}) > z(\mathbf{x}^i)$. Otherwise, a random behavior is implemented. In b-AFSA, crossover between \mathbf{x}^i and \mathbf{x}^{rand} is performed to create the trial point \mathbf{y}^i .

Random Behavior: When the ‘visual scope’ is empty or the other fish behavior were not performed, the point \mathbf{x}^i performs the random behavior. This behavior is related with a random movement for a better region. In b-AFSA, the trial point \mathbf{y}^i is created by randomly setting binary 0/1 bits of length n .

3.4 Operators Used in b-AFSA

Like other binary represented population-based solution methods, in b-AFSA crossover and mutation are used in fish behavior to create trial points. Different types of crossover and mutation implemented and tested in b-AFSA are described in the following.

Crossover: In b-AFSA, the crossover is performed in the chasing and searching behavior of the artificial fish swarm algorithm to create the trial points. There are many types of crossover applied to the selected current points of binary represented population-based methods. In this paper, we apply one position and uniform crossover and analyze their performances with respect to some measures of obtained results. With equal length of two current points, in one position crossover, a random index $r_1 \in \{1, 2, \dots, n\}$ is selected and then the bits from r_1 to n are exchanged to each other to create two trial points. In b-AFSA, after performing one position crossover, the best trial point with respect to the objective function value is selected for the new trial point. Figure 3 shows one position crossover.

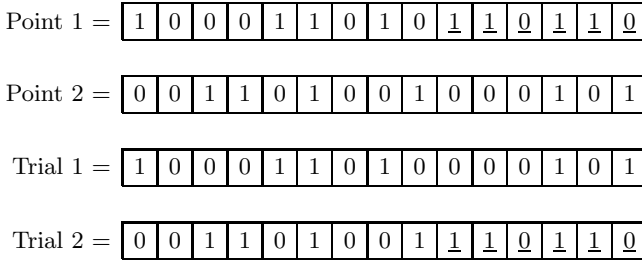


Fig. 3. One position crossover applied in b-AFSA

On the other hand, in uniform crossover two current points create a single trial point as shown in Fig. 4. Each bit in the trial point is created by copying the corresponding bit from one or the other current point with equal probability. If a uniformly distributed random number $\tau_1 \sim U[0, 1]$ is less than or equal to 0.5, the bit is copied from the first current point, otherwise, the bit is copied from the second current point.

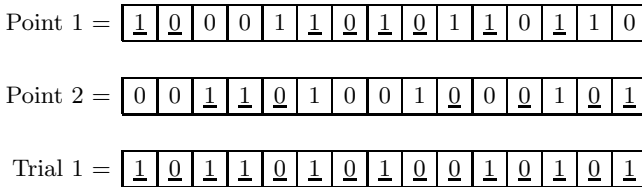


Fig. 4. Uniform crossover applied in b-AFSA

Mutation: The mutation is also performed in the proposed b-AFSA. In the swarming behavior, one position mutation is used to create a trial point. Here, a random index $r_2 \in \{1, 2, \dots, n\}$ is selected and then the bit of selected position is changed from 0 to 1 or vice versa. Figure 5 shows an example of one position mutation.

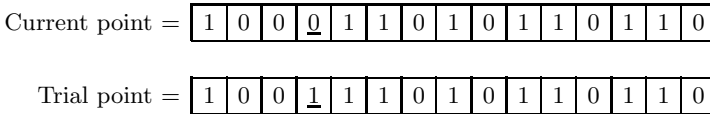


Fig. 5. One position mutation applied in b-AFSA

In the leaping (discussed in Section 3.7) behavior, a mutation with probability p_m is performed that mutates some randomly selected bits to create a trial point, i.e., if a uniformly distributed random number $\tau_2 \sim U[0, 1]$ generated for $j = 1, 2, \dots, n$ is less than or equal to p_m , then the bit of the successful position is changed from 0 to 1 or vice versa, as shown in Fig. 6. The probability of mutation $p_m = 0.01$ is widely used in binary represented solution methods.

Current point =	1	0	0	<u>0</u>	1	1	0	1	0	1	1	0	1	<u>1</u>	0
-----------------	---	---	---	----------	---	---	---	---	---	---	---	---	---	----------	---

Trial point =	1	0	0	<u>1</u>	1	1	0	1	0	1	1	0	1	<u>0</u>	0
---------------	---	---	---	----------	---	---	---	---	---	---	---	---	---	----------	---

Fig. 6. Mutation with probability p_m applied in b-AFSA

3.5 Selection

After creating the N trial points $\mathbf{y}^{i,t+1}$, $i = 1, 2, \dots, N$, the decoding algorithm is performed to make them feasible. In order to decide whether or not they should become members of the population in the next iteration $t + 1$, the trial point $\mathbf{y}^{i,t+1}$ is compared to the current point $\mathbf{x}^{i,t}$ using the following greedy criterion:

$$\mathbf{x}^{i,t+1} = \begin{cases} \mathbf{y}^{i,t+1} & \text{if } z(\mathbf{y}^{i,t+1}) \geq z(\mathbf{x}^{i,t}) \\ \mathbf{x}^{i,t} & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, N. \quad (4)$$

3.6 Termination Condition

Let nfe_{\max} be the maximum number of function evaluations. If nfe and z_{\max} are the number of function evaluations and the maximum objective function value attained at iteration t , and if z_{opt} is the known optimal value, then the proposed b-AFSA terminates if ($nfe > nfe_{\max}$ or $(|z_{\max} - z_{\text{opt}}|) \leq \epsilon$), for a small positive number ϵ .

This termination condition enables the b-AFSA to terminate when it reaches the best solution with a tolerance ϵ , otherwise it continues execution until nfe_{\max} is reached. But if the optimal value of the given problem is unknown, the algorithm may use other termination conditions.

3.7 Leaping

When the best objective function value in the population does not change for a certain number of iterations, the algorithm may have stagnated. The other points of the population will in the subsequent iterations eventually converge to that objective function value. To be able to escape from this region and try to converge to the optimal solution, the b-AFSA performs the leaping behavior, at every L iterations. In the leaping, a point \mathbf{x}^{rand} ($\text{rand} \in \{1, 2, \dots, N\}$) is randomly selected from the current population and the mutation with probability p_m is performed to that point. After mutation, decoding is performed and the new point replaces the point \mathbf{x}^{rand} .

3.8 The b-AFSA

The algorithm of the herein proposed binary version of the artificial fish swarm algorithm for solving (II) is outlined.

Step 1: Set parameter values.

Step 2: Set $t = 1$. Randomly initialize $\mathbf{x}^{i,1}$, $i = 1, 2, \dots, N$.

Step 3: Perform decoding and evaluate z . Identify \mathbf{x}_{\max} and z_{\max} .

Step 4: If termination condition is met, stop.

Step 5: For all $\mathbf{x}^{i,t}$,

Calculate ‘visual scope’ and crowding factor;

Perform fish behavior to create trial point $\mathbf{y}^{i,t+1}$;

Perform decoding to make the trial point feasible.

Step 6: Perform selection according to (4) to create new current points.

Step 7: Evaluate z and identify \mathbf{x}_{\max} and z_{\max} .

Step 8: If $\text{MOD}(t, L) = 0$ perform leaping.

Step 9: Set $t = t + 1$ and go to Step 4.

4 Experimental Results

We code b-AFSA in C and compile with Microsoft Visual Studio 9.0 compiler in a PC having 2.5 GHz Intel Core 2 Duo processor and 4 GB RAM. We set $\delta = 0.5$, $\theta = 0.8$, $p_m = 0.01$ and $\epsilon = 10^{-4}$. For a fair comparison with the solution method presented in [8], we use $N = 40$ and $nfe_{\max} = 1000n$, but one can set other suitable values. In order to perform the leaping behavior in b-AFSA, we also set $L = \max(25, n)$.

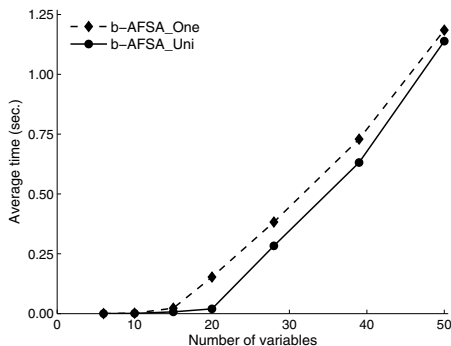
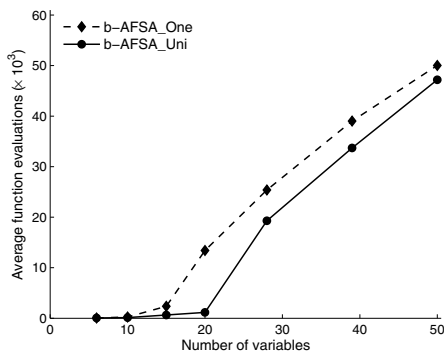
We consider seven benchmark problems from OR-library [4]. These problems are known as Petersen set PT1 to PT7 [27] of multidimensional 0–1 knapsack problems in the optimization community. The number of variables, n , in these problems vary from six to 50, and m (number of constraints) varies from five to 10. The optimum solution of each problem, z_{opt} , is known. So we use the termination condition described in Section 3.6.

Firstly, we compare the performance criteria of the two different variants of b-AFSA namely b-AFSA_One and b-AFSA_Uni on Petersen set PT1–PT7. These variants are based on the two different crossover operators used in the chasing and searching behavior. We implemented the one position crossover in variant b-AFSA_One and the uniform crossover in b-AFSA_Uni. Thirty independent runs were carried out for each problem using each variant. The performance criteria among 30 runs are: z_{\max} ; the average of best objective function values, z_{avg} ; the average computational time (in seconds), ‘AT’; the average number of function evaluations, ‘AFE’; and the number of successful runs, ‘Nsr’. In a run if the algorithm finds the optimal solution (or near optimal with a tolerance) of a test problem, then the run is considered to be a successful run. The comparative results are shown in Table 1. The table shows that the variant b-AFSA_Uni gives better performances than the other in respect to all measures of the performance criteria.

We compare in Fig. 7 and 8, the profiles of ‘AT’ and ‘AFE’ with respect to the number of variables, n , respectively. These figures show that the variant b-AFSA_Uni, with uniform crossover, outperforms the other one in comparison. Hence, this is the variant that will be used for the comparison with other solution

Table 1. Comparative results of b-AFSA_One and b-AFSA_Uni

Prob.	Size (n/m)	z_{opt}	b-AFSA_One					b-AFSA_Uni				
			z_{max}	z_{avg}	AT	AFE	Nsr	z_{max}	z_{avg}	AT	AFE	Nsr
PT1	6/10	3800.0	3800.0	3800.0	0.00	45	30	3800.0	3800.0	0.00	44	30
PT2	10/10	8706.1	8706.1	8706.1	0.00	239	30	8706.1	8706.1	0.00	133	30
PT3	15/10	4015.0	4015.0	4014.7	0.02	2360	29	4015.0	4015.0	0.01	651	30
PT4	20/10	6120.0	6120.0	6103.3	0.15	13407	14	6120.0	6120.0	0.02	1165	30
PT5	28/10	12400.0	12400.0	12261.0	0.38	25388	3	12400.0	12382.3	0.28	19287	10
PT6	39/5	10618.0	10605.0	10440.3	0.73	39018	0	10618.0	10559.2	0.63	33687	5
PT7	50/5	16537.0	16394.0	16143.8	1.18	50021	0	16537.0	16422.5	1.14	47199	2

**Fig. 7.** Profile of average computational time among 30 runs**Fig. 8.** Profile of average number of function evaluations among 30 runs

methods available in literature. For simplicity the variant will be denoted only by b-AFSA.

Secondly, we compare b-AFSA with the algorithms described in [8,25,30] respectively denoted by SCPSO, OGA and GADS. We also code GADS in C and

run with the recommended parameters [30]. We use the data available in the corresponding literature for OGA and SCPSO. In this comparison, the values of N and nfe_{\max} are the same for SCPSO, GADS and b-AFSA, although OGA had different values. All the results are based on 30 runs. See Table 2. In this

Table 2. Comparative results of OGA, SCPSO, GADS and b-AFSA

Prob.	Method	SR	AFE	AFEs _r	AT	AT _s	AE	LE	SDE
PT1	OGA	–	357.00	–	N/A	–	–	0.00	–
	SCPSO	100.00	109.33	109.33	–	–	0.00	0.00	0.00
	GADS	100.00	45.33	45.33	0.00	0.00	0.00	0.00	0.00
	b-AFSA	100.00	44.00	44.00	0.00	0.00	0.00	0.00	0.00
PT2	OGA	–	488.00	–	N/A	–	–	0.00	–
	SCPSO	100.00	446.66	446.66	–	–	0.00	0.00	0.00
	GADS	100.00	144.00	144.00	0.00	0.00	0.00	0.00	0.00
	b-AFSA	100.00	133.33	133.33	0.00	0.00	0.00	0.00	0.00
PT3	OGA	–	4645.00	–	N/A	–	–	0.00	–
	SCPSO	100.00	1736.00	1736.00	–	–	0.00	0.00	0.00
	GADS	100.00	1497.33	1497.33	0.01	0.01	0.00	0.00	0.00
	b-AFSA	100.00	650.70	650.70	0.01	0.01	0.00	0.00	0.00
PT4	OGA	–	7971.00	–	N/A	–	–	0.00	–
	SCPSO	96.67	5226.67	4717.24	–	–	0.33	0.00	1.79
	GADS	100.00	4047.00	4047.00	0.02	0.02	0.00	0.00	0.00
	b-AFSA	100.00	1165.47	1165.47	0.02	0.02	0.00	0.00	0.00
PT5	OGA	–	10059.00	–	N/A	–	–	10.00	–
	SCPSO	46.67	19009.30	8734.29	–	–	8.66	0.00	10.87
	GADS	73.33	15706.67	11236.36	0.09	0.07	2.67	0.00	4.50
	b-AFSA	33.33	19287.33	1816.00	0.28	0.03	17.67	0.00	21.28
PT6	OGA	–	17100.00	–	N/A	–	–	0.00	–
	SCPSO	50.00	27818.70	16637.30	–	–	19.63	0.00	27.67
	GADS	6.67	37220.00	12300.00	0.25	0.08	52.70	0.00	27.70
	b-AFSA	16.67	33686.70	7009.00	0.63	0.15	58.83	0.00	50.00
PT7	OGA	–	22659.00	–	N/A	–	–	13.00	–
	SCPSO	10.00	47005.30	20053.30	–	–	55.40	0.00	48.97
	GADS	0.00	50000.00	*	0.41	*	155.10	93.00	31.80
	b-AFSA	6.67	47198.80	7680.00	1.14	0.22	94.53	0.00	56.43

– Not available, * No success run, N/A not applicable

table, the performance criteria among 30 runs are: the success rate, ‘SR’; the average number of function evaluations, ‘AFEs_r’, and the average computational time, ‘AT_s’, in the successful runs. Performance criteria ‘AE’, ‘LE’ and ‘SDE’ are the average, least and standard deviation of errors, respectively, based on the objective function values. ‘AFE’ and ‘AT’ bear the same previously defined meanings. The values of the performance criteria ‘AT’ and ‘AT_s’ are reported only for GADS and b-AFSA, since they came from the same machine.

The table shows that with respect to all measures of the performance criteria, b-AFSA provides satisfactory results. Based on ‘SR’ and the errors in objective function values, SCPSO performs relatively better than the others although it did not give 100% success rate for problem PT4, when comparing to GADS and the herein presented b-AFSA. GADS performs better than SCPSO and b-AFSA for PT5, but it did not give any optimal solution for PT7. According to ‘AFE’, b-AFSA performs relatively better than the other methods for problem PT1 to PT4, but for PT5 to PT7, OGA gives better performance. Based on ‘AFESr’, b-AFSA outperforms other methods although the success rate decreased for problems PT5 to PT7.

In b-AFSA some extra computational time is required to calculate the ‘visual scope’ of all points in all iterations so the average computational times are a little bit higher for larger dimensional problems. We plot in Fig. 9 the profile of ‘AT’ and ‘ATsr’ of b-AFSA with respect to n , obtained after 30 runs, to show how the presented method performs with respect to computational time for solving multidimensional 0–1 knapsack problems. This figure shows that the average

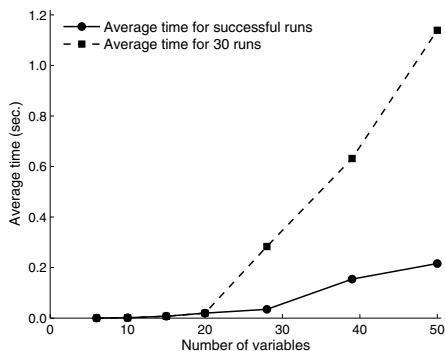


Fig. 9. Profile of average computational time of b-AFSA

computational time among 30 runs for larger dimensional problems increases almost linearly. The average time for successful runs also increases linearly with small slope. It means the b-AFSA can find optimal solution with less computational time although the success rate is low.

We may conclude from the above discussion that the herein presented b-AFSA has a good performance when solving multidimensional 0–1 knapsack problems.

5 Conclusions

In this paper, a binary version of the artificial fish swarm algorithm for solving multidimensional 0–1 knapsack problem has been presented. In this method an individual point in a population is represented by a binary string of 0/1 bits.

The Hamming distance is used to identify the neighborhood points inside the ‘visual scope’ of a reference point using visual radius. Depending on the number of points inside the ‘visual scope’, a point can perform either chasing, swarming, searching or random behavior. Crossover and mutation are implemented in the presented method to create trial points. In order to make points feasible, a decoding algorithm is also implemented. A greedy selection criterion is used to decide whether or not the trial points should become members of the population in the next iteration.

Performance of the herein presented binary version of the artificial fish swarm algorithm was tested on a set of multidimensional 0–1 knapsack problems. At first, a comparison of two variants of the method has been presented and it is found that the variant with uniform crossover is the best. Secondly, a comparison of the b-AFSA with other solution methods available in literature has been presented. It is found that the herein presented method has a good performance when solving a set of multidimensional 0–1 knapsack problems. Future development will focus on the multidimensional integer knapsack problems.

Acknowledgments. The authors thank three anonymous referees for their valuable comments to improve this paper.

The first author acknowledges Ciência 2007 of FCT, Fundação para a Ciência e a Tecnologia (Foundation for Science and Technology), Portugal for the financial support under fellowship grant: C2007-UMINHO-ALGORITMI-04. The second and third authors acknowledge FEDER COMPETE, Programa Operacional Fatores de Competitividade (Operational Programme Thematic Factors of Competitiveness) and FCT for the financial support under project grant: FCOMP-01-0124-FEDER-022674.

References

1. Akçay, Y., Li, H., Xu, S.H.: Greedy algorithm for the general multidimensional knapsack problem. *Ann. Oper. Res.* 150, 17–29 (2007)
2. Balev, S., Yanev, N., Fréville, A., Andonov, R.: A dynamic programming based reduction procedure for the multidimensional 0–1 knapsack problem. *Eur. J. Oper. Res.* 186, 63–76 (2008)
3. Battiti, R., Tecchiolli, G.: Local search with memory: benchmarking RTS. *OR Spektrum* 17, 67–86 (1995)
4. Beasley, J.E.: OR-Library; Distributing test problems by electronic mail. *J. Oper. Res. Soc.* 41, 1069–1072 (1990), <http://people.brunel.ac.uk/~mastjjb/jeb/info.html>
5. Boyer, V., Elkihel, M., Baz, D.E.: Heuristics for the 0–1 multidimensional knapsack problem. *Eur. J. Oper. Res.* 199, 658–664 (2009)
6. Cabot, A.V.: An enumeratuion algorithm for knapsack problems. *Oper. Res.* 18, 306–311 (1970)
7. Chu, P.C., Beasley, J.E.: A genetic algorithm for the multidimensional knapsack problem. *J. Heuristics* 4, 63–86 (1998)

8. Deep, K., Bansal, J.C.: A socio-cognitive particle swarm optimization for multidimensional knapsack problem. In: *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology*, pp. 355–360 (2008)
9. Djannaty, F., Doostdar, S.: A hybrid genetic algorithm for the multidimensional knapsack problem. *Int. J. Contemp. Math. Sci.* 3(9), 443–456 (2008)
10. Drexl, A.: A simulated annealing approach to the multiconstraint zero–one knapsack problem. *Computing* 40, 1–8 (1988)
11. Fernandes, E.M.G.P., Martins, T.F.M.C., Rocha, A.M.A.C.: Fish swarm intelligent algorithm for bound constrained global optimization. In: Aguiar, J.V. (ed.) *CMMSE 2009*, pp. 461–472 (2009)
12. Fontanari, J.F.: A statistical analysis of the knapsack problem. *J. Phys. A: Math. Gen.* 28, 4751–4759 (1995)
13. Fréville, A., Plateau, G.: The 0–1 bidimensional knapsack problem: Towards an efficient high-level primitive tool. *J. Heuristics* 2, 147–167 (1996)
14. Fréville, A.: The multidimensional 0–1 knapsack problem: An overview. *Eur. J. Oper. Res.* 155, 1–21 (2004)
15. Gavish, B., Pirkul, H.: Efficient algorithms for solving multiconstraint zero–one knapsack problems to optimality. *Math. Program.* 31, 78–105 (1985)
16. Gilmore, P.C., Gomory, R.E.: The theory and computation of knapsack functions. *Oper. Res.* 14, 1045–1075 (1966)
17. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading (1989)
18. Hanafi, S., Fréville, A.: An efficient tabu search approach for the 0–1 multidimensional knapsack problem. *Eur. J. Oper. Res.* 106, 659–675 (1998)
19. He, J., Miao, Z., Zhang, Z., Shi, X.: Solving multidimensional 0–1 knapsack problem by tissue P systems with cell division. In: *Proceedings of the Fourth International Conference on Bio-Inspired Computing BIC–TA 2009*, pp. 249–253 (2009)
20. Hill, R.R., Cho, Y.K., Moore, J.T.: Problem reduction heuristic for the 0–1 multidimensional knapsack problem. *Comput. Oper. Res.* 39, 19–26 (2012)
21. Jiang, M., Wang, Y., Pfletschinger, S., Lagunas, M.A., Yuan, D.: Optimal Multiuser Detection with Artificial Fish Swarm Algorithm. In: Huang, D.-S., Heutte, L., Loog, M. (eds.) *ICIC 2007, Part 22. CCIS, vol. 2*, pp. 1084–1093. Springer, Heidelberg (2007)
22. Jiang, M., Mastorakis, N., Yuan, D., Lagunas, M.A.: Image Segmentation with Improved Artificial Fish Swarm Algorithm. In: Mastorakis, N., Mladenov, V., Konstantgyri, V.T. (eds.) *ECC 2008. LNEE, vol. 28*, pp. 133–138. Springer, Heidelberg (2009)
23. Khuri, S., Bäck, T., Heitkötter, J.: The zero/one multiple knapsack problem and genetic algorithm. In: *Proceedings of the 1994 ACM Symposium on Applied Computing*, pp. 188–193 (1994)
24. Kong, M., Tian, P., Kao, Y.: A new ant colony optimization algorithm for the multidimensional knapsack problem. *Comput. Oper. Res.* 35, 2672–2683 (2008)
25. Li, H., Jiao, Y.-C., Zhang, L., Gu, Z.-W.: Genetic Algorithm Based on the Orthogonal Design for Multidimensional Knapsack Problems. In: Jiao, L., Wang, L., Gao, X.-b., Liu, J., Wu, F. (eds.) *ICNC 2006, Part I. LNCS, vol. 4221*, pp. 696–705. Springer, Heidelberg (2006)
26. Michalewicz, Z.: *Genetic Algorithms+Data Structures=Evolution Programs*. Springer, Berlin (1996)
27. Petersen, C.C.: Computational experience with variants of the Balas algorithm applied to the selection of R&D projects. *Manag. Sci.* 13(9), 736–750 (1967)

28. Rocha, A.M.A.C., Martins, T.F.M.C., Fernandes, E.M.G.P.: An augmented Lagrangian fish swarm based method for global optimization. *J. Comput. Appl. Math.* 235, 4611–4620 (2011)
29. Rocha, A.M.A.C., Fernandes, E.M.G.P., Martins, T.F.M.C.: Novel Fish Swarm Heuristics for Bound Constrained Global Optimization Problems. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part III. LNCS, vol. 6784, pp. 185–199. Springer, Heidelberg (2011)
30. Sakawa, M., Kato, K.: Genetic algorithms with double strings for 0–1 programming problems. *Eur. J. Oper. Res.* 144, 581–597 (2003)
31. Schilling, K.E.: The growth of m -constraint random knapsacks. *Eur. J. Oper. Res.* 46, 109–112 (1990)
32. Shih, W.: A branch and bound method for the multiconstraint zero–one knapsack problem. *J. Oper. Res. Soc.* 30, 369–378 (1979)
33. Soyster, A.L., Lev, B., Slivka, W.: Zero–one programming with many variables and few constraints. *Eur. J. Oper. Res.* 2, 195–201 (1978)
34. Vasquez, M., Vimont, Y.: Improved results on the 0–1 multidimensional knapsack problem. *Eur. J. Oper. Res.* 165, 70–81 (2005)
35. Veni, K.K., Balachandar, S.R.: A new heuristic approach for large size zero–one multi knapsack problem using intercept matrix. *Int. J. Comput. Math. Sci.* 4(5), 259–263 (2010)
36. Wang, C.-R., Zhou, C.-L., Ma, J.-W.: An improved artificial fish swarm algorithm and its application in feed-forward neural networks. In: Proceedings of the 4th ICMLC, pp. 2890–2894 (2005)
37. Wang, X., Gao, N., Cai, S., Huang, M.: An Artificial Fish Swarm Algorithm Based and ABC Supported QoS Unicast Routing Scheme in NGL. In: Min, G., Di Martino, B., Yang, L.T., Guo, M., Rünger, G. (eds.) ISPA Workshops 2006. LNCS, vol. 4331, pp. 205–214. Springer, Heidelberg (2006)
38. Weingartner, H.M., Ness, D.N.: Methods for the solution of the multidimensional 0/1 knapsack problem. *Oper. Res.* 15, 83–103 (1967)

Optimization Model of COTS Selection Based on Cohesion and Coupling for Modular Software Systems under Multiple Applications Environment

Pankaj Gupta*, Shilpi Verma, and Mukesh Kumar Mehlawat

Department of Operational Research, University of Delhi, Delhi, India
pgupta@or.du.ac.in, {vermashilpi,mukesh0980}@yahoo.com

Abstract. Due to the rapid growth of development of component based software systems, the optimal commercial-off-the-shelf (COTS) selection has become the key concept of optimization techniques used for the purpose. In this paper, we propose an optimization model that aims to select the best-fit COTS components for a modular software system under multiple applications development task. The proposed model maximizes the functional performance and minimizes the total cost of the software system satisfying the constraints of minimum threshold on intra-modular coupling density and reusability of COTS components. A real-world scenario of developing two financial applications for two small-scale industries is included to illustrate the efficiency of the model.

Keywords: Optimization model, COTS selection, Cohesion and Coupling, Reusability, Modular software system.

1 Introduction

Modern software systems are becoming more and more large-scale, complex and uneasily controlled, resulting in high development cost, low productivity, unmanageable software quality and high risk to move to new technology. Consequently, there is a growing demand of searching for a new, efficient, and cost-effective software development paradigm. One of the most promising solutions today is the component-based software development (CBSD) approach. This approach is based on the idea that software systems can be developed by selecting appropriate commercial-off-the-shelf (COTS) components and then assemble them to fit a specific architectural style for some application(s) domain.

A COTS component can be developed by different developers using different languages and different platforms. In general, a COTS component has three main features: 1) a component is an independent and replaceable part of a system

* The first author acknowledges the research grant received under a scheme for strengthening R & D Doctoral Research Programme of University of Delhi, Delhi, India.

that fulfills a given function; 2) a component works in the context of a well-defined architecture; and 3) a component communicates with other components of the software system through its interfaces [3]. In CBSD, the main focus is how to choose the most appropriate and most suited component from COTS components' market so that it can significantly reduce development cost and time-to-market, and improve maintainability, reliability and overall quality of software system. Several COTS selection methods [5, 8, 10, 11, 13, 16] have been proposed in literature. However, it may be noted that there is no single method which is accepted as a standard COTS selection method. A detailed list of the COTS selection methods has been provided in Mohamed et al. [14].

Alternatively, optimization techniques have been used in the COTS selection process to achieve the different attributes of quality along with the objective of minimizing the cost or keeping cost to a specified budgetary level. Berman and Ashrafi [2] discussed optimization models for reliability of modular software systems. Chi et al. [4] presented a software reliability optimization model. Cortellessa et al. [6] developed an optimization model that supports "build-or-buy" decisions in selecting software components based on cost-reliability trade-off. Jung and Choi [9] introduced two optimization models for the COTS selection in the development of modular software systems considering cost-reliability trade-off. Kwong et al. [12] presented an optimization model for determining the optimal selection of software components for component-based software system development. Neubauer and Stummer [15] presented a two-phase decision support approach based on multiobjective optimization for the COTS selection. Tang et al. [18] presented an optimization model for software component selection under multiple applications development. Zachariah and Rattihalli [19] used goal-programming approach in a multi-criteria optimization model for the COTS selection of modular software systems. Zahedi and Ashrafi [20] discussed software reliability allocation using optimization approach based on structure, utility, price and cost.

All the optimization models discussed above are based on the assumption that in the software development process, COTS components within a set of alternative components exhibit similar functionality. However, in real-world scenario, the functions of the COTS components could be different from each other because they are provided by different vendors. Thus, in order to fulfill the functional requirements of software system using CBSD, the functional contributions of various COTS components must also be considered. It may also be noted that in the development of a modular software system, the criteria of maximizing the cohesion and minimizing the coupling of software modules are commonly used. Coupling is about the measure of interactions among software modules while cohesion is about the measure of interactions among software components which are within a software module. A good software system should possess software modules with high cohesion and low coupling. A highly cohesive module exhibits high reusability and loosely coupled systems enable easy maintenance.

In this paper, we propose a bi-objective optimization model for the COTS selection in the development of a modular software system. The proposed model

simultaneously maximizes the functional performance and minimizes the total cost of a modular software system. The selection of COTS components is constrained using minimum threshold on the intra-modular coupling density of the software, and reusability of COTS components. The proposed research can be considered as a generalization and extension of the optimization models proposed in [12, 18] in terms of providing a systematic framework for the COTS selection that facilitates software development process of a modular software under multiple applications development task.

The rest of the paper is organized as follows. Section 2 describes the criteria used for COTS component selection. In Section 3, mathematical formulation of the optimization model is introduced. Section 4 discusses solution methodology. Section 5 presents numerical illustrations of a real-world scenario inspired from CBSD to test the effectiveness of the proposed model. Finally, we furnish our concluding remarks in Section 6.

2 Criteria for COTS Selection under Multiple Applications Development

In order to select COTS components for modular software systems, the following criteria may be used.

2.1 Functional Performance

The functional capabilities of the COTS components are different for different components. Functionality of the COTS component is nothing but the ability of the component to perform according to the specific needs of the organization. We use functional ratings of the COTS components to the software modules as coefficients in the objective function corresponding to maximizing the functional performance of the modular software system. It may be noted that these ratings are assumed to be provided by the software development team.

2.2 Cost

The cost criterion is used to assess cost related characteristics of the components. In this paper, we consider cost based on procurement and adaptation costs of COTS components. The procurement cost contains licensing arrangement cost, product and technology cost and consulting cost.

2.3 Intra-modular Coupling Density

Abreu and Goulão [1] have proposed quantitative measures of cohesion and coupling. The relationship between cohesion and coupling of modules in the development of modular software system can be measured by using intra-modular coupling density (*ICD*) defined as follows:

$$ICD = \frac{CI_{IN}}{CI_{IN} + CI_{OUT}} \quad (1)$$

where CI_{IN} is the number of class interactions within modules, and CI_{OUT} is the number of interactions between classes of distinct modules. ICD presents the ratio between cohesion and coupling. It is well known that loose coupling and tight cohesion can achieve high maintainability of a software system. Thus, the values of ICD for each of the application would have great influence on the maintainability of the modular software system. Fig. 1 (replicated from [12]) shows the diagrammatic depiction of cohesion and coupling of software modules in the development of a modular software system.

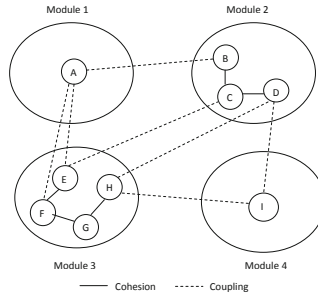


Fig. 1. Cohesion and coupling of software modules in CBSD

2.4 Reusability

Reusability of a component defines the extent to which parts of a software can be reused in other applications. Software reuse means to abstract the general logic from different applications, implement the logic, and be used into more applications with slight or no modification. Thus, reusability of a COTS component means that a component can be reused into different applications rather than one application.

3 COTS Selection Problem Formulation

We consider component selection problem for developing a modular software system under multiple applications development task. The software developer concurrently undertakes N applications which consists of M modules as shown in Fig. 2. Suppose the i th application requires m_i modules, then $M = \sum_{i=1}^N m_i$. Each module must contain at least one COTS component. The COTS components' market contains L components which are distributed among T different sets of alternative COTS components. The COTS components within a set fulfills the same functional requirement of the system and only one component from a given set is selected in each application. The main tasks considered in this paper are how to select software components available in the COTS components' market

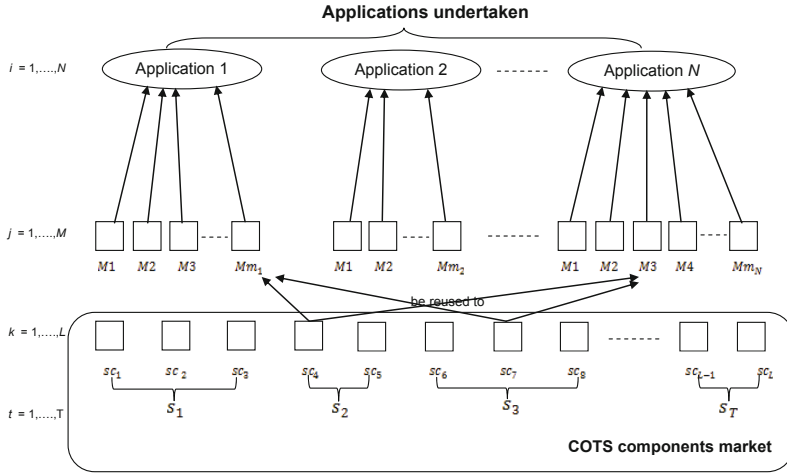


Fig. 2. Three layers of hierarchy of a modular software system with multiple applications

and deploy them into which application in order to maximize the functional requirements of the software system and minimize the total cost of procurement of COTS components and adaptation of components to various modules in all the undertaken applications.

3.1 Notations

The following notations are used to formulate the mathematical model:

N : the number of applications handled concurrently,

M : the number of modules in the given software system,

L : the number of COTS components,

T : the number of sets of alternative COTS components,

sc_k : the k th COTS component, $k = 1, \dots, L$,

s_t : the set of alternative COTS components for the t th functional requirement of the software system, $t = 1, \dots, T$,

c_k^p : the procurement cost of k th COTS component, $k = 1, \dots, L$,

c_{jk}^a : the adaptation cost if the k th COTS component is adapted into the j th module, $j = 1, \dots, M$, $k = 1, \dots, L$,

$r_{kk'}$: the number of interactions between k th and k' th COTS components, $k, k' = 1, \dots, L$,

f_{jk} : the functional rating of k th COTS component to j th module, $f_{jk} \in [0, 1]$, $j = 1, \dots, M$, $k = 1, \dots, L$,

s_{ij} : the binary parameter,

$$s_{ij} = \begin{cases} 1, & \text{if the } j\text{th module belongs to the } i\text{th application,} \\ 0, & \text{otherwise, } i = 1, \dots, N, j = 1, \dots, M, \end{cases}$$

b_{jk} : the binary parameter,

$$b_{jk} = \begin{cases} 1, & \text{if the } k\text{th COTS component can be reused to implement the } \\ & \text{ } j\text{th module,} \\ 0, & \text{otherwise, } j = 1, \dots, M, k = 1, \dots, L, \end{cases}$$

$x_{j,k}$: the binary variable,

$$x_{j,k} = \begin{cases} 1, & \text{if the } k\text{th COTS component is selected to implement the } \\ & \text{ } j\text{th module,} \\ 0, & \text{otherwise, } j = 1, \dots, M, k = 1, \dots, L, \end{cases}$$

y_k : the binary variable,

$$y_k = \begin{cases} 1, & \text{if the } k\text{th COTS component is selected,} \\ 0, & \text{otherwise, } k = 1, \dots, L, \end{cases}$$

H : a threshold value of ICD for each application in order to achieve a given level of maintainability of software system,

N'_k : the number of times k th COTS component can be used in all the applications.

It may be noted that $r_{kk'} = r_{k'k}$ since cohesion and coupling are undirected relations.

3.2 Bi-objective Optimization Model

The proposed bi-objective optimization model of COTS component selection maximizes the functional requirements of the modular software system and minimizes the total development cost of the system which includes the procurement and adaptation costs of COTS components subject to many realistic constraints including a minimum threshold on ICD , reusability constraint, selection of only one COTS component from a set of alternative components for each functional requirement per application and selection of more than one component per module if required.

Suppose the cohesion within the j th module, $(CI_{IN})_j$, is given by

$$(CI_{IN})_j = \sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} x_{j,k} x_{j,k'}.$$

Then, the sum of cohesions within all modules of the i th application, $(CI_{IN})_i$, can be expressed as

$$(CI_{IN})_i = \sum_{j=1}^M s_{ij} (CI_{IN})_j = \sum_{j=1}^M s_{ij} \left(\sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} x_{j,k} x_{j,k'} \right). \quad (2)$$

Further, let all interactions including cohesion and coupling associated with the j th module, $(CA)_j$, is expressed as:

$$(CA)_j = (CI_{IN})_j + (CI_{OUT})_j = \sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} x_{j,k} \left(\sum_{j=1}^M s_{ij} x_{j,k'} \right).$$

Then, all interactions including cohesion and coupling of the i th application, $(CA)_i$, can be expressed as:

$$\begin{aligned} (CA)_i &= (CI_{IN})_i + (CI_{OUT})_i = \sum_{j=1}^M s_{ij} (CA)_j \\ &= \sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} \left(\sum_{j=1}^M s_{ij} x_{j,k} \right) \left(\sum_{j=1}^M s_{ij} x_{j,k'} \right). \end{aligned} \quad (3)$$

Thus, using Eqs. (1), (2) and (3), ICD for the i th application is given by

$$(ICD)_i = \frac{\sum_{j=1}^M s_{ij} \left(\sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} x_{j,k} x_{j,k'} \right)}{\sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} \left(\sum_{j=1}^M s_{ij} x_{j,k} \right) \left(\sum_{j=1}^M s_{ij} x_{j,k'} \right)}.$$

The optimization model is now formulated as follows:

$$(P) \max F = \sum_{j=1}^M \sum_{k=1}^L f_{jk} x_{j,k} \quad (4)$$

$$\min C = \sum_{k=1}^L c_k^p y_k + \sum_{j=1}^M \sum_{k=1}^L c_{jk}^a x_{j,k} \quad (5)$$

Subject to

$$\frac{\sum_{j=1}^M s_{ij} \left(\sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} x_{j,k} x_{j,k'} \right)}{\sum_{k=1}^{L-1} \sum_{k'=k+1}^L r_{kk'} \left(\sum_{j=1}^M s_{ij} x_{j,k} \right) \left(\sum_{j=1}^M s_{ij} x_{j,k'} \right)} \geq H, \quad i = 1, \dots, N \quad (6)$$

$$\sum_{j=1}^M s_{ij} x_{j,k} \leq 1, \quad i = 1, \dots, N, \quad k = 1, \dots, L \quad (7)$$

$$\sum_{k \in s_t} \sum_{j=1}^M s_{ij} x_{j,k} = 1, \quad i = 1, \dots, N, \quad t = 1, \dots, T \quad (8)$$

$$\sum_{k=1}^L x_{j,k} \geq 1, \quad j = 1, \dots, M \quad (9)$$

$$\sum_{j=1}^M x_{j,k} = y_k \cdot N'_k, \quad k = 1, \dots, L \quad (10)$$

$$x_{j,k} \leq b_{jk}, \quad j = 1, \dots, M, \quad k = 1, \dots, L \quad (11)$$

$$x_{j,k}, y_k \in \{0, 1\}, \quad j = 1, \dots, M, \quad k = 1, \dots, L \quad (12)$$

3.3 Explanation on the Formulation of the Model

The objective function (4) maximizes the functional performance of the software system. The first term of the objective function (5) represents the total procurement cost of all selected COTS components and the second term represents the total adaptation cost of all selected components. Hence, the objective (5) minimizes the total cost for accomplishing all the undertaken applications.

Constraint (6) indicates that the *ICD* of each of the applications should be greater than the minimum threshold value set by the decision maker. Constraint (7) expresses that each of the COTS component can be reused into an appli-

cation at most once. If $\sum_{j=1}^M s_{ij}x_{j,k} = 0$, then the k th COTS component is not

selected into the i th application, otherwise, it means the k th component is selected. Constraint (8) denotes that only one component can be selected from a set of alternative COTS components for a particular functional requirement of each application. Constraint (9) denotes that each software module must contain at least one COTS component. Constraint (10) represents the logical relation between y_k and $x_{j,k}$. It indicates that the k th COTS component can be available for implementing a module into an application, and N'_k times if it is selected,

i.e. $\sum_{j=1}^M x_{j,k} = N'_k$ while $y_k = 1$, and $\forall k, x_{j,k} = 0$ when $y_k = 0$. Constraint

(11) denotes that the k th component is selected to implement the j th module if and only if $b_{jk} = 1$. Constraint (12) suggests selection or rejection of COTS components.

4 Solution Approach

The involvement of multiple objectives in the proposed optimization model (P) brings complexity while finding the optimal solution. The multiple objective optimization model as defined above often contains many optimal solutions which are called pareto optimal solutions or efficient solutions. For a minimization model with R objective functions, a feasible solution x^* is said to be pareto optimal if and only if there does not exists another feasible solution x such that $f_i(x^*) \leq f_i(x) \forall i = 1, \dots, R$ and $f_j(x^*) < f_j(x)$ for at least one j .

There are various solution approaches for solving the multiple objective optimization models. Among the most widely used techniques are weighted sum method, ε -constraint method and goal programming method. In this paper, the weighted sum method is used to solve the bi-objective optimization model (P). Let F_{max} and F_{min} be the upper and lower bounds of the objective function

(4), respectively. Let C_{max} and C_{min} denotes the upper and lower bounds of the objective function (5), respectively. The value of F_{max} can be determined by solving the model (FU) defined as follows:

$$(FU) \max \sum_{j=1}^M \sum_{k=1}^L f_{jk} x_{j,k}$$

Subject to
Constraints (6) – (12)

Similarly, the value of F_{min} can be determined by solving the model (FL) defined as follows:

$$(FL) \min \sum_{j=1}^M \sum_{k=1}^L f_{jk} x_{j,k}$$

Subject to
Constraints (6) – (12)

Again, the model (CU) determines the value of C_{max} as follows:

$$(CU) \max \left(\sum_{k=1}^L c_k^p y_k + \sum_{j=1}^M \sum_{k=1}^L c_{jk}^a x_{j,k} \right)$$

Subject to
Constraints (6) – (12)

Also, the model (CL) finds the value of C_{min} as follows:

$$(CL) \min \left(\sum_{k=1}^L c_k^p y_k + \sum_{j=1}^M \sum_{k=1}^L c_{jk}^a x_{j,k} \right)$$

Subject to
Constraints (6) – (12)

After obtaining the upper and lower bounds of the two objective functions, the proposed bi-objective optimization model can be reformulated as the following weighted sum optimization model:

$$(P1) \max (w_f \cdot F' + w_c \cdot C')$$

Subject to
Constraints (6) – (12),
 $w_f + w_c = 1,$
 $w_f \geq 0, w_c \geq 0$

where $F' = \frac{F - F_{min}}{F_{max} - F_{min}}$ and $C' = \frac{C_{max} - C}{C_{max} - C_{min}}$. Also, w_f and w_c are the weights of the objectives (4) and (5), respectively.

Theorem 1 ([7]). *The optimal solution of the weighted sum optimization model (P1) is pareto optimal solution of the bi-objective optimization model (P) if the weighting coefficients w_f and w_c are positive.*

The optimization models (FU), (FL), (CU), (CL) and (P1) are nonlinear optimization models. We use Lingo software [17] to solve them.

5 An Illustrative Example

In order to illustrate the proposed methodology of optimizing the selection of best-fit COTS components for modular software systems under multiple applications development task, a hypothetical small-scale scenario of software development discussed in [12, 18] is presented in this section. Let us consider that a software developer undertakes two financial applications for two different small-size industries, i.e. Garment Industry Financial System (App1) and Pharmaceutical Industry Financial System (App2). App1 consists of three modules: Garment business-related module ($M1$), Garment security module ($M2$) and assistance for Garment Industry ($M3$). Similarly, App2 includes: Pharmaceutical business-related module ($M4$), Pharmaceutical security module ($M5$) and assistance for Pharmaceutical Industry ($M6$). The software system should fulfill six basic functional requirements, namely, Facsimile/Fax (R_1), Encryption (R_2), Credit Card Authorization (R_3), Automatic Updates (R_4), e-Commerce (R_5) and Financial Reporting (R_6). E-commerce and financial reporting functions are provided by business-related module, the functions of encryption and authorization are provided by the security module while functions like fax and software automatic updating are mainly provided by the assistance module. Apart from fulfilling the main functional requirements these modules can meet other functional requirements as well. Let us consider that 12 COTS components are available in COTS components' market which are denoted by sc_1-sc_{12} . Individual functional requirements and their corresponding alternative COTS components, functional ratings of COTS components to the software modules, as well as the procurement cost of COTS components and the adaptation costs of each component to various modules are shown in Table 1.

The functional ratings provides the degree of functional contribution of each COTS component to the various software modules. The function ratings which are fixed based on group decision-making of the software system development team ranges from 0 to 1 where 1 refers to a very high degree of contribution. Table 2 shows the degrees of interaction among the COTS components which are provided by the team's judgement and varies between 0-10, where the degree 10 refers to a very high degree of interaction. The degree of interaction among the COTS components which exhibits similar functionality can be considered as 0.

Suppose the software developer initially set $H = 0.3$ and give equal importance to both the objective functions, i.e. $w_f = w_c = 0.5$. Also, let each COTS component can be used twice in all the applications. The values of F_{max} , F_{min} , C_{max} and C_{min} are calculated by solving the problems (FU), (FL), (CU), (CL), respectively. These values are obtained as 9.65, 6.21, 569 and 509, respectively.

Table 1. Initial parameters for COTS components

Functional Requirements		R_1				R_2	R_3	R_4				R_5		R_6
		sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7	sc_8	sc_9	sc_{10}	sc_{11}	sc_{12}	
	c_k^p	49	66	54	55	62	61	74	74	79	47	40	49	
M1	c_{1k}^a	-	-	-	-	22	-	-	-	-	12	14	16	
	f_{1k}	0	0	0	0	0.35	0	0	0	0	0.98	0.89	0.75	
App1	M2	c_{2k}^a	20	19	18	17	14	13	-	-	-	-	-	
		f_{2k}	0.32	0.22	0.15	0.23	0.94	0.68	0	0	0	0	0	
	M3	c_{3k}^a	18	12	13	11	-	21	17	19	16	-	-	
		f_{3k}	0.51	0.63	0.72	0.57	0	0.45	0.94	0.86	1	0	0	
	M4	c_{4k}^a	-	-	-	-	20	-	-	-	-	12	16	
		f_{4k}	0	0	0	0	0.15	0	0	0	0	0.90	0.83	
App2	M5	c_{5k}^a	17	18	14	15	11	13	-	-	-	-	-	
		f_{5k}	0.35	0.24	0.13	0.21	0.85	0.70	0	0	0	0	0	
	M6	c_{6k}^a	13	9	11	13	-	20	18	14	13	-	-	
		f_{6k}	0.45	0.65	0.49	0.54	0	0.40	0.90	0.65	0.97	0	0	

Table 2. Interactions among COTS components

	sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7	sc_8	sc_9	sc_{10}	sc_{11}	sc_{12}
sc_1	0	0	0	0	0	1	6	8	7	0	0	0
sc_2	0	0	0	0	7	6	8	9	7	0	0	0
sc_3	0	0	0	0	8	7	9	7	6	0	0	0
sc_4	0	0	0	0	4	3	5	6	8	0	0	0
sc_5	0	7	8	4	0	8	0	0	0	7	7	8
sc_6	1	6	7	3	8	0	5	8	7	0	0	0
sc_7	6	8	9	5	0	5	0	0	0	0	0	0
sc_8	8	9	7	6	0	8	0	0	0	0	0	0
sc_9	7	7	6	8	0	7	0	0	0	0	0	0
sc_{10}	0	0	0	0	7	0	0	0	0	0	0	8
sc_{11}	0	0	0	0	7	0	0	0	0	0	0	9
sc_{12}	0	0	0	0	8	0	0	0	0	8	9	0

Using these values and the data given in Tables 1 and 2, we obtain the mathematical model (P1) of COTS selection problem as follows:

$$\begin{aligned} \max = & 0.5(((0.35x_{1,5} + 0.98x_{1,10} + 0.89x_{1,11} + 0.75x_{1,12} + 0.32x_{2,1} + 0.22x_{2,2} + \\ & 0.15x_{2,3} + 0.23x_{2,4} + 0.94x_{2,5} + 0.68x_{2,6} + 0.51x_{3,1} + 0.63x_{3,2} + 0.72x_{3,3} + \\ & 0.57x_{3,4} + 0.45x_{3,6} + 0.94x_{3,7} + 0.86x_{3,8} + x_{3,9} + 0.15x_{4,5} + 0.90x_{4,10} + \\ & 0.83x_{4,11} + 0.60x_{4,12} + 0.35x_{5,1} + 0.24x_{5,2} + 0.13x_{5,3} + 0.21x_{5,4} + 0.85x_{5,5} + \\ & 0.70x_{5,6} + 0.45x_{6,1} + 0.65x_{6,2} + 0.49x_{6,3} + 0.54x_{6,4} + 0.40x_{6,6} + 0.90x_{6,7} + \\ & 0.65x_{6,8} + 0.97x_{6,9}) - 6.21)/(3.44) + 0.5((569 - (49y_1 + 66y_2 + 54y_3 + 55y_4 \\ & + 62y_5 + 61y_6 + 74y_7 + 74y_8 + 79y_9 + 47y_{10} + 40y_{11} + 49y_{12} + 22x_{1,5} + \\ & 12x_{1,10} + 14x_{1,11} + 16x_{1,12} + 20x_{2,1} + 19x_{2,2} + 18x_{2,3} + 17x_{2,4} + 14x_{2,5} + \\ & 13x_{2,6} + 18x_{3,1} + 12x_{3,2} + 13x_{3,3} + 11x_{3,4} + 21x_{3,6} + 17x_{3,7} + 19x_{3,8} + 16x_{3,9} \\ & + 20x_{4,5} + 12x_{4,10} + 16x_{4,11} + 15x_{4,12} + 17x_{5,1} + 18x_{5,2} + 14x_{5,3} + 15x_{5,4} + \\ & 11x_{5,5} + 13x_{5,6} + 13x_{6,1} + 9x_{6,2} + 11x_{6,3} + 13x_{6,4} + 20x_{6,6} + 18x_{6,7} + 14x_{6,8} \\ & + 13x_{6,9}))/60) \end{aligned}$$

Subject to

$$\begin{aligned} & \left(\sum_{j=1}^3 x_{j,1}(x_{j,6} + 6x_{j,7} + 8x_{j,8} + 7x_{j,9}) + \sum_{j=1}^3 x_{j,2}(7x_{j,5} + 6x_{j,6} + 8x_{j,7} + 9x_{j,8} \right. \\ & + 7x_{j,9}) + \sum_{j=1}^3 x_{j,3}(8x_{j,5} + 7x_{j,6} + 9x_{j,7} + 7x_{j,8} + 6x_{j,9}) + \sum_{j=1}^3 x_{j,4}(4x_{j,5} + 3x_{j,6} \\ & + 5x_{j,7} + 6x_{j,8} + 8x_{j,9}) + \sum_{j=1}^3 x_{j,5}(8x_{j,6} + 7x_{j,10} + 7x_{j,11} + 8x_{j,12}) + \sum_{j=1}^3 x_{j,6} \\ & \left. (5x_{j,7} + 8x_{j,8} + 7x_{j,9}) + \sum_{j=1}^3 x_{j,10}(8x_{j,12}) + \sum_{j=1}^3 x_{j,11}(9x_{j,12}) \right) / \\ & \left(\sum_{j=1}^3 x_{j,1} \left(\sum_{j=1}^3 x_{j,6} + 6 \sum_{j=1}^3 x_{j,7} + 8 \sum_{j=1}^3 x_{j,8} + 7 \sum_{j=1}^3 x_{j,9} \right) + \sum_{j=1}^3 x_{j,2} \left(7 \sum_{j=1}^3 x_{j,5} + \right. \right. \\ & 6 \sum_{j=1}^3 x_{j,6} + 8 \sum_{j=1}^3 x_{j,7} + 9 \sum_{j=1}^3 x_{j,8} + 7 \sum_{j=1}^3 x_{j,9} \left. \right) + \sum_{j=1}^3 x_{j,3} \left(8 \sum_{j=1}^3 x_{j,5} + 7 \sum_{j=1}^3 x_{j,6} \right. \\ & + 9 \sum_{j=1}^3 x_{j,7} + 7 \sum_{j=1}^3 x_{j,8} + 6 \sum_{j=1}^3 x_{j,9} \left. \right) + \sum_{j=1}^3 x_{j,4} \left(4 \sum_{j=1}^3 x_{j,5} + 3 \sum_{j=1}^3 x_{j,6} + 5 \sum_{j=1}^3 x_{j,7} \right. \\ & + 6 \sum_{j=1}^3 x_{j,8} + 8 \sum_{j=1}^3 x_{j,9} \left. \right) + \sum_{j=1}^3 x_{j,5} \left(8 \sum_{j=1}^3 x_{j,6} + 7 \sum_{j=1}^3 x_{j,10} + 7 \sum_{j=1}^3 x_{j,11} + \right. \\ & 8 \sum_{j=1}^3 x_{j,12} \left. \right) + \sum_{j=1}^3 x_{j,6} \left(5 \sum_{j=1}^3 x_{j,7} + 8 \sum_{j=1}^3 x_{j,8} + 7 \sum_{j=1}^3 x_{j,9} \right) + \sum_{j=1}^3 x_{j,10} \left(8 \sum_{j=1}^3 x_{j,12} \right) \\ & + \sum_{j=1}^3 x_{j,11} \left(9 \sum_{j=1}^3 x_{j,12} \right) \geq 0.3, \\ & \left(\sum_{j=4}^6 x_{j,1}(x_{j,6} + 6x_{j,7} + 8x_{j,8} + 7x_{j,9}) + \sum_{j=4}^6 x_{j,2}(7x_{j,5} + 6x_{j,6} + 8x_{j,7} + 9x_{j,8} \right. \\ & + 7x_{j,9}) + \sum_{j=4}^6 x_{j,3}(8x_{j,5} + 7x_{j,6} + 9x_{j,7} + 7x_{j,8} + 6x_{j,9}) + \sum_{j=4}^6 x_{j,4}(4x_{j,5} + 3x_{j,6} \\ & + 5x_{j,7} + 6x_{j,8} + 8x_{j,9}) + \sum_{j=4}^6 x_{j,5}(8x_{j,6} + 7x_{j,10} + 7x_{j,11} + 8x_{j,12}) + \sum_{j=4}^6 x_{j,6} \\ & \left. (5x_{j,7} + 8x_{j,8} + 7x_{j,9}) + \sum_{j=4}^6 x_{j,10}(8x_{j,12}) + \sum_{j=4}^6 x_{j,11}(9x_{j,12}) \right) / \\ & \left(\sum_{j=4}^6 x_{j,1} \left(\sum_{j=4}^6 x_{j,6} + 6 \sum_{j=4}^6 x_{j,7} + 8 \sum_{j=4}^6 x_{j,8} + 7 \sum_{j=4}^6 x_{j,9} \right) + \sum_{j=4}^6 x_{j,2} \left(7 \sum_{j=4}^6 x_{j,5} + \right. \right. \end{aligned}$$

$$\begin{aligned}
 & 6 \sum_{j=4}^6 x_{j,6} + 8 \sum_{j=4}^6 x_{j,7} + 9 \sum_{j=4}^6 x_{j,8} + 7 \sum_{j=4}^6 x_{j,9} \Big) + \sum_{j=4}^6 x_{j,3} \left(8 \sum_{j=4}^6 x_{j,5} + 7 \sum_{j=4}^6 x_{j,6} \right. \\
 & \left. + 9 \sum_{j=4}^6 x_{j,7} + 7 \sum_{j=4}^6 x_{j,8} + 6 \sum_{j=4}^6 x_{j,9} \right) + \sum_{j=4}^6 x_{j,4} \left(4 \sum_{j=4}^6 x_{j,5} + 3 \sum_{j=4}^6 x_{j,6} + 5 \sum_{j=4}^6 x_{j,7} \right. \\
 & \left. + 6 \sum_{j=4}^6 x_{j,8} + 8 \sum_{j=4}^6 x_{j,9} \right) + \sum_{j=4}^6 x_{j,5} \left(8 \sum_{j=4}^6 x_{j,6} + 7 \sum_{j=4}^6 x_{j,10} + 7 \sum_{j=4}^6 x_{j,11} + \right. \\
 & \left. 8 \sum_{j=4}^6 x_{j,12} \right) + \sum_{j=4}^6 x_{j,6} \left(5 \sum_{j=4}^6 x_{j,7} + 8 \sum_{j=4}^6 x_{j,8} + 7 \sum_{j=4}^6 x_{j,9} \right) + \sum_{j=4}^6 x_{j,10} \left(8 \sum_{j=4}^6 x_{j,12} \right) \\
 & + \sum_{j=4}^6 x_{j,11} \left(9 \sum_{j=4}^6 x_{j,12} \right) \geq 0.3, \\
 & \sum_{j=1}^3 x_{j,k} \leq 1 \quad \forall k = 1, \dots, 12, \quad \sum_{j=4}^6 x_{j,k} \leq 1 \quad \forall k = 1, \dots, 12, \\
 & \sum_{j=1}^3 (x_{j,1} + x_{j,2} + x_{j,3} + x_{j,4}) = 1, \quad \sum_{j=1}^3 x_{j,5} = 1, \quad \sum_{j=1}^3 x_{j,6} = 1, \\
 & \sum_{j=1}^3 (x_{j,7} + x_{j,8} + x_{j,9}) = 1, \quad \sum_{j=1}^3 (x_{j,10} + x_{j,11}) = 1, \quad \sum_{j=1}^3 x_{j,12} = 1, \\
 & \sum_{j=4}^6 (x_{j,1} + x_{j,2} + x_{j,3} + x_{j,4}) = 1, \quad \sum_{j=4}^6 x_{j,5} = 1, \quad \sum_{j=4}^6 x_{j,6} = 1, \\
 & \sum_{j=4}^6 (x_{j,7} + x_{j,8} + x_{j,9}) = 1, \quad \sum_{j=4}^6 (x_{j,10} + x_{j,11}) = 1, \quad \sum_{j=4}^6 x_{j,12} = 1, \\
 & \sum_{k=1}^{12} x_{j,k} \geq 1 \quad \forall j = 1, \dots, 6, \\
 & \sum_{j=1}^6 x_{j,k} = 2y_k \quad \forall k = 1, \dots, 12, \\
 & x_{1,1} \leq 0, x_{1,2} \leq 0, x_{1,3} \leq 0, x_{1,4} \leq 0, x_{1,5} \leq 1, x_{1,6} \leq 0, x_{1,7} \leq 0, x_{1,8} \leq 0, \\
 & x_{1,9} \leq 0, x_{1,10} \leq 1, x_{1,11} \leq 1, x_{1,12} \leq 1, x_{2,1} \leq 1, x_{2,2} \leq 1, x_{2,3} \leq 1, x_{2,4} \leq 1, \\
 & x_{2,5} \leq 1, x_{2,6} \leq 1, x_{2,7} \leq 0, x_{2,8} \leq 0, x_{2,9} \leq 0, x_{2,10} \leq 0, x_{2,11} \leq 0, x_{2,12} \leq 0, \\
 & x_{3,1} \leq 1, x_{3,2} \leq 1, x_{3,3} \leq 1, x_{3,4} \leq 1, x_{3,5} \leq 0, x_{3,6} \leq 1, x_{3,7} \leq 1, x_{3,8} \leq 1, \\
 & x_{3,9} \leq 1, x_{3,10} \leq 0, x_{3,11} \leq 0, x_{3,12} \leq 0, x_{4,1} \leq 0, x_{4,2} \leq 0, x_{4,3} \leq 0, x_{4,4} \leq 0, \\
 & x_{4,5} \leq 1, x_{4,6} \leq 0, x_{4,7} \leq 0, x_{4,8} \leq 0, x_{4,9} \leq 0, x_{4,10} \leq 1, x_{4,11} \leq 1, x_{4,12} \leq 1, \\
 & x_{5,1} \leq 1, x_{5,2} \leq 1, x_{5,3} \leq 1, x_{5,4} \leq 1, x_{5,5} \leq 1, x_{5,6} \leq 1, x_{5,7} \leq 0, x_{5,8} \leq 0, \\
 & x_{5,9} \leq 0, x_{5,10} \leq 0, x_{5,11} \leq 0, x_{5,12} \leq 0, x_{6,1} \leq 1, x_{6,2} \leq 1, x_{6,3} \leq 1, x_{6,4} \leq 1, \\
 & x_{6,5} \leq 0, x_{6,6} \leq 1, x_{6,7} \leq 1, x_{6,8} \leq 1, x_{6,9} \leq 1, x_{6,10} \leq 0, x_{6,11} \leq 0, x_{6,12} \leq 0, \\
 & x_{j,k}, y_k \in \{0, 1\}, j = 1, \dots, 6, k = 1, \dots, 12
 \end{aligned}$$

By solving the above optimization model using Lingo software, we obtain the optimal values of F and C as 9.58 and 511, respectively. The COTS components sc_{10} and sc_{12} selected for modules $M1$ and $M4$ provides the functional requirements of e-commerce and financial reporting as desired. Modules $M2$ and $M5$ gets COTS components sc_5 and sc_6 which fulfills the functional requirements of encryption and credit card authorization. Modules $M3$ and $M6$ gets COTS components sc_3 and sc_9 which contributes toward the functional requirements of fax and automatic updates.

Further, we perform sensitivity analysis with respect to changes in the minimum threshold value of ICD for each application in order to increase maintainability of the software system. For different values of H the results obtained are listed in Table 3. From Table 3 it is clear that if we increase the minimum threshold value of H , i.e. increase the maintainability of the software system, it has adverse effect on the two objective functions. Also, if the software development team desires that the ICD level should be at least 0.35 then second and third solutions can be considered for implementation. Besides, considering the objective values the development team can also select any one of the obtained solutions based on various criteria such as their preferences or customers' expectations. Suppose, a customer has a limited budget of 512 then only first two solutions can be considered for implementation.

Table 3. COTS selection corresponding to $w_f = 0.5, w_c = 0.5$

H -threshold	F	C	Software components selected		
			$(M1, M4)$	$(M2, M5)$	$(M3, M6)$
0.3	9.58	511	sc_{10}, sc_{12}	sc_5, sc_6	sc_3, sc_9
0.4	9.48	512	sc_{10}, sc_{12}	sc_5, sc_6	sc_4, sc_9
0.5	9.33	513	sc_{10}, sc_{12}	sc_5, sc_6	sc_1, sc_9

Next, by varying the weights of the two objective functions we can obtain different solutions for a fixed value of ICD . For example, suppose the development team gives more importance to cost as compared to functional performance of the system by setting $w_f = 0.1$ and $w_c = 0.9$. Then, for same threshold value of H the cost of the system decreases but at the same time the functional performance also decreases. The computational result shown in Table 3 corresponding to $H = 0.3$ when compared with the computational results presented in Table 4 justifies the claim. Again, if the team gives more importance to functional performance as compared to cost then functional performance increases but cost also increases. The cost-functional performance efficient frontier is shown in Fig. 3.

Table 4. COTS selection corresponding to $H = 0.3$

w_f	w_c	F	C	Software components selected		
				$(M1, M4)$	$(M2, M5)$	$(M3, M6)$
0.1	0.9	8.96	509	sc_{11}, sc_{12}	sc_5, sc_6	sc_3, sc_8
0.2	0.8	9.42	510	sc_{11}, sc_{12}	sc_5, sc_6	sc_3, sc_9
0.3	0.7	9.58	511	sc_{10}, sc_{12}	sc_5, sc_6	sc_3, sc_9
0.9	0.1	9.65	520	sc_{10}, sc_{12}	sc_5, sc_6	sc_2, sc_9

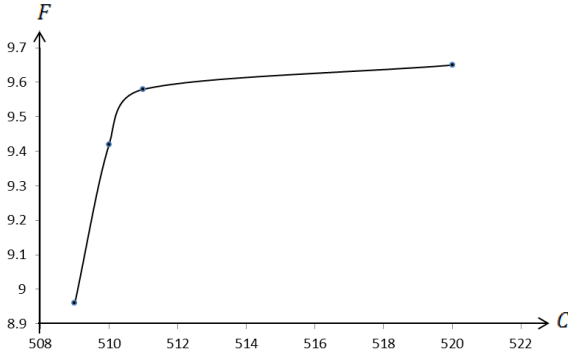


Fig. 3. Cost-functional performance efficient frontier

6 Conclusions

In this paper, we have introduced a bi-objective optimization model which maximizes the functional requirements and minimizes the total cost of the modular software system. Compared with the previous studies, the proposed model additionally considered components' reusability and cohesion and coupling of software modules simultaneously under multiple applications environment. The model is used to assist software developers in selecting best-fit COTS components when multiple applications are undertaken concurrently. The effectiveness of the model is demonstrated through numerical examples constructed corresponding to a real-world scenario of multiple applications environment. The model sensitivity have been shown with respect to changes in the minimum threshold value of the intra-modular coupling density for each application and also by varying the weight parameters of the two objective functions reflecting the preferences of the development team. The proposed methodology involves some subjective judgements from software developer team such as the determination of the scores of interaction and the functional ratings which may be considered as limitations of the study. Fuzzy set theory could be used as an alternative to deal with the fuzziness caused by subjective judgements which will be addressed in the near future as an extension of this paper.

References

1. Abreu, F.B., Goulão, M.: Coupling and cohesion as modularization drivers: are we being over-persuaded? In: Proceedings of the Fifth European Conference on Software Maintenance and Reengineering, IEEE Computer Society, Washington, DC, USA (2001)
2. Berman, O., Ashrafi, N.: Optimization models for reliability of modular software systems. *IEEE Transactions on Software Engineering* 19(11), 1119–1123 (1993)
3. Brown, A.W., Wallnau, K.C.: The current state of CBSE. *IEEE Software* 15(5), 37–46 (1998)

4. Chi, D.-H., Lin, H.-H., Kuo, W.: Software reliability and redundancy optimization. In: Proceedings of the Annual Reliability and Maintainability Symposium. IEEE, pp. 41–45 (1989)
5. Chung, L., Cooper, K., Courtney, S.: COTS-Aware requirements engineering: The CARE process. In: Proceedings of the 2nd International Workshop on Requirements Engineering for COTS Components (RECOTS 2004), Kyoto, Japan, September 7 (2004)
6. Cortellessa, V., Marinelli, F., Potena, P.: An optimization framework for “build-or-buy” decisions in software architecture. *Computers & Operations Research* 35, 3090–3106 (2008)
7. Ehrgott, M.: *Multicriteria optimization*, 2nd edn. Springer, New York (2005)
8. Grau, G., Carvallo, J.P., Franch, X., Quer, C.: DesCOTS: A software system for selecting COTS components. In: Proceedings of the 30th IEEE Euromicro Conference (EUROMICRO 2004). IEEE (2004)
9. Jung, H.-W., Choi, B.: Optimization models for quality and cost of modular software systems. *European Journal of Operational Research* 112, 613–619 (1999)
10. Kontio, J., Chen, S.-F., Limperos, K., Tesoriero, R., Caldiera, G., Deutsch, M.: A COTS selection method and experiences of its use. In: Twentieth Annual Software Engineering Workshop, NASA Goddard Space Flight Center, Greenbelt, Maryland (November 1995)
11. Kotonya, G., Hutchinson, J.: Viewpoints for Specifying Component-Based Systems. In: Crnkovic, I., et al. (eds.) CBSE 2004. LNCS, vol. 3054, pp. 114–121. Springer, Heidelberg (2004)
12. Kwong, C.K., Mu, L.F., Tang, J.F., Luo, X.G.: Optimization of software components selection for component-based software system development. *Computers & Industrial Engineering* 58, 618–624 (2010)
13. Leung, K.R.P.H., Leung, H.K.N.: On the efficiency of domain-based COTS product selection method. *Information and Software Technology* 44(12), 703–715 (2002)
14. Mohamed, A., Ruhe, G., Eberlein, A.: COTS selection: past, present, and future. In: Proceedings of the 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS 2007). IEEE (2007)
15. Neubauer, T., Stummer, C.: Interactive decision support for multiobjective COTS selection. In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS 2007). IEEE (2007)
16. Rolland, C.: Requirement engineering for COTS based systems. *Information and Software Technology* 41(14), 985–990 (1999)
17. Schrage, L.: *Optimization Modeling with LINGO*, 5th edn. Lindo Systems Inc., Chicago (2003)
18. Tang, J.F., Mu, L.F., Kwong, C.K., Luo, X.G.: An optimization model for software component selection under multiple applications development. *European Journal of Operational Research* 212(2), 301–311 (2011)
19. Zachariah, B., Rattihalli, R.N.: A multicriteria optimization model for quality of modular software systems. *Asia-Pacific Journal of Operational Research* 24(6), 797–811 (2007)
20. Zahedi, F., Ashrafi, N.: Software reliability allocation based on structure, utility, price and cost. *IEEE Transactions on Software Engineering* 17(4), 345–356 (1991)

A Derivative-Free Filter Driven Multistart Technique for Global Optimization

Florbela P. Fernandes^{1,3}, M. Fernanda P. Costa^{2,3},
and Edite M.G.P. Fernandes⁴

¹ Polytechnic Institute of Bragança, ESTiG, 5301-857 Bragança, Portugal
fflor@ipb.pt

² Department of Mathematics and Applications, University of Minho, 4800-058
Guimarães, Portugal
mfc@mct.uminho.pt

³ Mathematics R&D Centre

⁴ Algoritmi R&D Centre,
University of Minho, 4710-057 Braga, Portugal
emgpf@dps.uminho.pt

Abstract. A stochastic global optimization method based on a multistart strategy and a derivative-free filter local search for general constrained optimization is presented and analyzed. In the local search procedure, approximate descent directions for the constraint violation or the objective function are used to progress towards the optimal solution. The algorithm is able to locate all the local minima, and consequently, the global minimum of a multi-modal objective function. The performance of the multistart method is analyzed with a set of benchmark problems and a comparison is made with other methods.

Keywords: Global optimization, Multistart, Descent Direction, Filter Method.

1 Introduction

Global optimization problems arise in many engineering applications. Owing to the existence of multiple minima, it is a challenging task to solve a multilocal optimization problem and to identify all the global minima.

The purpose of this paper is to present a technique for solving constrained global optimization problems based on a multistart method that uses a filter methodology to handle the constraints of the problem. The problem to be addressed is of the following type

$$\begin{aligned} & \min f(x) \\ & \text{subject to } g_j(x) \leq 0, \quad j = 1, \dots, m \\ & \quad \quad \quad l_i \leq x_i \leq u_i, \quad i = 1, \dots, n \end{aligned} \tag{1}$$

where, at least one of the functions $f, g_j : \mathbb{R}^n \rightarrow \mathbb{R}$ is nonlinear and $F = \{x \in \mathbb{R}^n : l_i \leq x_i \leq u_i, i = 1, \dots, n, g_j(x) \leq 0, j = 1, \dots, m\}$ is the feasible region.

Problems with general nonlinear equality constraints can be reformulated in the above form by introducing $h(x) = 0$ as inequality constraints $|h(x)| - \tau \leq 0$, where τ is a small positive relaxation parameter. Since this kind of problems may have many global and local (non-global) optimal solutions (convexity is not assumed), it is important to develop a methodology that is able to explore the entire search space, find all the (local) minima guaranteeing, in some way, that convergence to a previously found minimum is avoided, and identify the global ones.

The two major classes of methods for solving problem (II) globally are the deterministic and the stochastic one. One of the most known stochastic algorithms is the multistart. In the last decade some research has been focused on this type of methods [1, 6, 9–11]; see also [7] and the references therein included. The underlying idea of this method is to sample uniformly a point from the search region and to perform a local search, starting from this point, to obtain an optimal (local) solution, using a local technique. This is repeated until the stop conditions are met. One of the advantages of multistart is that it has the potential of finding all local minima; although, it has the drawback of locating the same solution more than once.

Here, we are specially interested in developing a simple to implement and efficient method for the identification of at least one global optimal solution of problem (II) that is based on a multistart paradigm. A multistart strategy is chosen due to its simplicity and previously observed practical good performance.

The herein proposed method does not compute or approximate any derivatives or penalty parameters. Our proposal for the local search relies on a procedure, namely, the approximate descent direction (ADD) method, which is a derivative-free procedure with high ability of producing a descent direction. The ADD method is combined with a (line search) filter method to generate trial solutions that might be acceptable if they improve the constraint violation or the objective function. Hence, the progress towards a solution that is feasible and optimal is carried out by a filter method. This is a recent strategy that has shown to be highly competitive with penalty function methods [2–4].

This paper is organized as follows. In Section 2, the algorithm based on the multistart strategy and on the filter methodology is presented. In Section 3, we report the results of our numerical experiments with a set of benchmark problems. In the last section, conclusions are summarized and recommendations for future work are given.

2 The Filter Driven Multistart Method

This section describes a multistart approximate descent direction filter-based approach, hereafter denoted by MADDF, that relies on a derivative-free local search procedure to converge to the local solutions of the problem. The exploration feature of the method is carried out by a multistart strategy that aims at generating points randomly spread all over the search space. Exploitation of promising regions are made by a simple local search approach. A derivative-free technique

that computes approximate descent directions, for either the constraint violation or the objective function, is implemented with reduced computational costs. To measure progress towards an optimal solution a filter methodology, as outlined in [4], is integrated into the local search procedure. The filter methodology appears naturally from the observation that an optimal solution of the problem (1) minimizes both constraint violation and objective function [2–5].

2.1 A Multistart Strategy

The basic multistart algorithm starts by randomly generating a point x from the search space $S \subset \mathbb{R}^n$, and a local search procedure is applied from x to converge to a local minimum y . We will denote the implementation of the local procedure to provide the minimum y by $y = \mathcal{L}(x)$. Subsequently, another point is randomly generated from the search space and the local search is again applied to give another local minimum. This process is repeated until a stopping rule is satisfied. The pseudo-code of this procedure is presented below in Algorithm 1.

Algorithm 1. Basic multistart algorithm

```

1: Set  $k = 1$ ;
2: Randomly generate  $x$  from  $S$ ;
3: Compute  $y_1 = \mathcal{L}(x)$ ;
4: while the stopping rule is not satisfied do
5:   Randomly generate  $x$  from  $S$ ;
6:   Compute  $y = \mathcal{L}(x)$ ;
7:   if  $y \notin \{y_i, i = 1, \dots, k\}$  then
8:      $k = k + 1$ ;
9:     Set  $y_k = y$ ;
10:  end if
11: end while

```

Unfortunately, this multistart strategy has a drawback since the same local minimum may be found over and over again. To prevent the repetitive invoking of the local search procedure, converging to previously found local minima, clustering techniques have been incorporated into the multistart strategy. To guarantee that a local minimum is found only once, the concept of region of attraction of a local minimum is introduced.

Definition 1. *The region of attraction of a local minimum associated with a local search procedure \mathcal{L} is defined as:*

$$A_i \equiv \{x \in \mathcal{S}, y_i = \mathcal{L}(x)\}, \quad (2)$$

where $y_i = \mathcal{L}(x)$ is the minimizer obtained when the local search procedure \mathcal{L} is started at point x .

This concept is very important because it guarantees that the local search applied to any point x from the region of attraction A_i will converge eventually to the same minimizer y_i . Thus, after y_i has been found there is no point in starting the local search from any other point in that region of attraction.

Let N be the number of local minima in \mathcal{S} . From the previous definition it follows that

$$\mathcal{S} = \bigcup_{i=1}^N A_i \text{ and } A_i \cap A_j = \emptyset, \text{ for } i \neq j. \quad (3)$$

A multistart method that uses the concept of region of attraction proceeds as follows: it starts by randomly generating a point from \mathcal{S} , and applies a local search to obtain the first minimum y_1 with the region of attraction A_1 . Afterwards, other points are randomly generated from \mathcal{S} until a point is found that does not belong to A_1 . Next, the local search is performed and a new minimizer y_2 is obtained, with the region of attraction A_2 . The next point from which a local search will start does not belong to $A_1 \cup A_2$. This procedure continues until a stopping rule is satisfied. The corresponding multistart algorithm is presented in the Algorithm 2.

Algorithm 2. Multistart Clustering algorithm

```

1: Set  $k = 1$ ;
2: Randomly generate  $x$  from  $\mathcal{S}$ ;
3: Compute  $y_1 = \mathcal{L}(x)$  and the corresponding  $A_1$ ;
4: while the stopping rule is not satisfied do
5:   Randomly generate  $x$  from  $\mathcal{S}$ ;
6:   if  $x \notin \cup_{i=1}^k A_i$  then
7:     Compute  $y = \mathcal{L}(x)$ ;
8:      $k = k + 1$ ;
9:     Set  $y_k = y$  and compute the corresponding  $A_k$ ;
10:  end if
11: end while

```

Theoretically, this algorithm invokes the local search procedure only N times, where N is the number of existing minima of (1). In practice, the regions of attraction A_k of the minima found so far are not easy to compute. A simple stochastic procedure is used to estimate the probability, p , that a randomly generated point will not belong to a specific set, which is the union of a certain number of regions of attraction, i.e., $p = P[x \notin \cup_{i=1}^k A_i]$. Using this reasoning, the new steps of the regions of attraction based multistart algorithm are described in Algorithm 3.

The probability p is estimated as follows [11]. Let the maximum attractive radius of the minimizer y_i be defined by:

$$R_i = \max_j \left\{ \left\| x_i^{(j)} - y_i \right\| \right\}, \quad (4)$$

Algorithm 3. Ideal Multistart algorithm

```

1: Set  $k = 1$ ;
2: Randomly generate  $x$  from  $S$ ;
3: Compute  $y_1 = \mathcal{L}(x)$  and the corresponding  $A_1$ ;
4: while the stopping rule is not satisfied do
5:   Randomly generate  $x$  from  $S$ ;
6:   Compute  $p = P[x \notin \cup_{i=1}^k A_i]$ ;
7:   Let  $\zeta$  be a uniform distributed number in  $(0, 1)$ ;
8:   if  $\zeta < p$  then
9:     Compute  $y = \mathcal{L}(x)$ ;
10:    if  $y \notin \{y_i, i = 1, \dots, k\}$  then
11:       $k = k + 1$ ;
12:      Set  $y_k = y$  and compute the corresponding  $A_k$ ;
13:    end if
14:  end if
15: end while

```

where $x_i^{(j)}$ are the generated points which led to the minimizer y_i . Given a randomly generated point x , let $z = \frac{\|x - y_i\|}{R_i}$. Clearly, if $z \leq 1$ then x is likely to be inside the region of attraction of y_i . On the other hand, if the direction from x to y_i is ascent then x is likely to be outside the region of attraction of y_i . Based on a suggestion presented in [11], an estimate of the probability that $x \notin A_i$ is herein computed by:

$$p(x \notin A_i) = \begin{cases} 1, & \text{if } z > 1 \text{ or the direction from } x \text{ to } y_i \text{ is ascent} \\ \varrho \phi(z, l), & \text{otherwise} \end{cases} \quad (5)$$

where $0 \leq \varrho \leq 1$ is a factor that depends on the directional derivative of f along the direction from x to y_i , l is the number of times y_i has been identified/recovered so far and the function $\phi(z, l)$ satisfies the properties:

$$\lim_{z \rightarrow 0} \phi(z, l) \rightarrow 0, \quad \lim_{z \rightarrow 1} \phi(z, l) \rightarrow 1, \quad \lim_{l \rightarrow \infty} \phi(z, l) \rightarrow 0 \quad \text{and} \quad 0 < \phi(z, l) < 1.$$

In the Ideal Multistart method [11], Voglis and Lagaris propose the

$$\phi(z, l) = z \exp(-l^2(z-1)^2) \quad \text{for all } z \in (0, 1). \quad (6)$$

Since the Algorithm 3 has the potential of finding all local minima and one global solution is to be required, each solution is compared with the previously identified solutions and the one with the most extreme value is always saved.

2.2 The Derivative-Free Filter Local Procedure

The local search procedure is an iterative method that is applied to a randomly generated point x and provides a trial point y that is an approximate minimizer of problem (II). Our proposal for the local search \mathcal{L} is an Approximate Descent

Direction Filter (ADDF) method. The point y is computed based on a direction d and a step size $\alpha \in (0, 1]$ in such a way that

$$y = x + \alpha d. \quad (7)$$

The procedure that decides which step size is accepted to generate an acceptable approximate minimizer is a filter method. The herein proposed multistart method uses the filter set concept [4] that has the ability to explore both feasible and infeasible regions. This technique incorporates the concept of nondominance, present in the field of multiobjective optimization, to build a filter that is able to accept a trial point if it improves either the objective function or the constraint violation, relative to the current point. Filter-based algorithms treat the optimization problem as a biobjective problem aiming to minimize both the objective function and the nonnegative constraint violation function. In this way, the previous constrained problem (II) is reformulated as a biobjective problem involving the original objective function f and the constraint violation function θ , as follows:

$$\min_{x \in S} (f(x), \theta(x)) \quad (8)$$

where for $\beta \in \{1, 2\}$

$$\theta(x) = \sum_{i=1}^m \left(\max_i \{0, g_i(x)\} \right)^\beta + \sum_{i=1}^n \left(\left(\max_i \{0, x_i - u_i\} \right)^\beta + \left(\max_i \{0, l_i - x_i\} \right)^\beta \right). \quad (9)$$

After a search direction d has been computed, a step size α is determined by a backtracking line search technique. A decreasing sequence of α values is tried until a set of acceptance conditions are satisfied. The trial point y , in (7), is acceptable if sufficient progress in θ or in f is verified, relative to the current point x , as shown:

$$\theta(y) \leq (1 - \gamma_\theta) \theta(x) \text{ or } f(y) \leq f(x) - \gamma_f \theta(x) \quad (10)$$

where $\gamma_\theta, \gamma_f \in (0, 1)$. However, when x is (almost) feasible, i.e., in practice when $\theta(x) \leq \theta_{\min}$, the trial point y has to satisfy only the condition

$$f(y) \leq f(x) - \gamma_f \theta(x) \quad (11)$$

to be acceptable, where $0 < \theta_{\min} \ll 1$. To prevent cycling between points that improve either θ or f , at each iteration, the algorithm maintains the filter \mathcal{F} which is a set of pairs (θ, f) that are prohibited for a successful trial point. During the backtracking line search procedure, the y is acceptable only if $(\theta(y), f(y)) \notin \mathcal{F}$. If the stopping conditions are not satisfied (see (14) ahead), $x \leftarrow y$ and this procedure is repeated.

The filter is initialized with pairs (θ, f) that satisfy $\theta \geq \theta_{\max}$, where $\theta_{\max} > 0$ is the upper bound on θ . Furthermore, whenever y is accepted because condition (10) is satisfied, the filter is updated by the formula

$$\mathcal{F} = \mathcal{F} \cup \{(\theta, f) \in \mathbb{R}^2 : \theta > (1 - \gamma_\theta)\theta(x) \text{ and } f > f(x) - \gamma_f\theta(x)\}.$$

When it is not possible to find a point y with a step size $\alpha > \alpha_{\min}$ ($0 < \alpha_{\min} \ll 1$) that satisfy one of the conditions (10) or (11), a restoration phase is invoked. In this phase, the algorithm recovers the best point in the filter, herein denoted by $x_{\mathcal{F}}^{best}$, and a new trial point is determined according to the strategy based on equation (7).

The algorithm implements the ADD method [5] to compute the direction d , required in (7). This strategy has a high ability of producing a descent direction for a specific function. The ADD method is a derivative-free procedure which uses several points around a given point $x \in \mathbb{R}^n$ to generate an approximate descent direction for a function ψ at x [5]. More specifically, the ADD method chooses r exploring points close to x , in order to generate an approximate descent direction $d \in \mathbb{R}^n$ for ψ at x . Hence, the direction $d = \frac{v}{\|v\|}$ is computed at x , after generating r points $\{a_i\}_{i=1}^r$ close to x , as shown:

$$v = \sum_{i=1}^r w_i e_i \quad (12)$$

where

$$w_i = \frac{\Delta\psi_i}{\sum_{j=1}^r |\Delta\psi_j|}, \quad \Delta\psi_i = \psi(a_i) - \psi(x), \quad i = 1, \dots, r \quad (13)$$

$$e_i = -\frac{a_i - x}{\|a_i - x\|} \quad i = 1, \dots, r.$$

In the ADDF context, the ADD method generates the search direction d , at a given point x , according to the following rules:

- If x is feasible (in practice, if $\theta(x) < \theta_{tol}$), the ADD method computes an approximate descent direction d for the objective function f at x and then $\psi = f$ in (13);
- If x is infeasible, the ADD method is used to compute an approximate descent direction d for the constraint violation function θ , at x , and in this case $\psi = \theta$.

To judge the success of the ADDF algorithm, the three below presented conditions are applied simultaneously, i.e., if

$$\begin{aligned} |f(y) - f(x)| \leq 10^{-4} |f(y)| + 10^{-6} \wedge |\theta(y) - \theta(x)| \leq 10^{-4} \theta(y) + 10^{-6} \\ \wedge \|y - x\| \leq 10^{-4} \|y\| + 10^{-6} \end{aligned} \quad (14)$$

hold, the local search procedure stops with a successful approximate local minimizer of problem (1). The proposed algorithm for the local procedure is presented in Algorithm 4.

Algorithm 4. ADDF algorithm

Require: x (sampled in multistart); Set $x_{\mathcal{F}}^{best} = x$ and $\tilde{x} = x$;

```

1: Initialize the filter;
2: while the stopping conditions are not satisfied do
3:   Set  $x = \tilde{x}$ ;
4:   Use ADD to compute  $v$  by (12);
5:   Set  $\alpha = 1$ ;
6:   Compute  $y$  using (7);
7:   while new trial  $y$  is not acceptable do
8:     Check acceptability of trial point, using (10) and (11);
9:     if acceptable by the filter then
10:      Update the filter if appropriate;
11:      Set  $\tilde{x} = y$ ; Update  $x_{\mathcal{F}}^{best}$ ;
12:     else
13:       Set  $\alpha = \alpha/2$ ;
14:       if  $\alpha < \alpha_{\min}$  then
15:         Set  $\alpha = 1$ ; Set  $x = x_{\mathcal{F}}^{best}$ ;
16:         Invoke restoration phase;
17:       end if
18:       Compute  $y$  using (7);
19:     end if
20:   end while
21: end while

```

2.3 Stopping Rule

Good stopping rules to identify multiple optimal solutions should combine reliability and economy. A reliable rule is one that stops only when all minima have been identified with certainty. An economical rule is one that invokes the local search the least number of times to verify that all minima have been found. A lot of research about stopping rules has been carried out in the past (see [6] and the references therein included). There are three established rules that have been successfully used [6].

We choose to use the following stopping condition [6]. If s denotes the number of recovered local minima after having performed t local search procedures, then the estimate of the fraction of the uncovered space is given by

$$P(s) = \frac{s(s+1)}{t(t-1)} \quad (15)$$

and the stopping rule is

$$P(s) \leq \epsilon \quad (16)$$

with ϵ being a small positive number.

3 Experimental Results

The MADDF method was coded in MatLab and the results were obtained in a PC with an Intel(R) Core(TM)2 Duo CPU P7370 2.00GHz processor and 3 GB of memory.

To perform some comparisons between other methods, it is necessary to set the MADDF parameters. The parameter τ used to reformulate equality into inequality constraints was set to $\tau = 10^{-5}$. Since derivatives are not provided to the algorithm, the factor ϱ is estimated and set to 0.05. The closer the direction $(y - x)$ is to the greatest decrease of f , the smaller is ϱ . The power factor used in equation (9) was set to 2 and to generate the approximate descent directions we set $r = 2$ and $r_{ADD} = 10^{-3}$ (the radius of the neighborhood in which the exploring points are generated), as suggested in [5]. In ADDF method, $\gamma_\theta = \gamma_f = 10^{-5}$, $\alpha_{\min} = 10^{-6}$, $\theta_{tol} = 10^{-5}$, $\theta_{\min} = 10^{-3} \max\{1, 1.25\theta(x_{initial})\}$, $\theta_{\max} = \max\{1, 1.25\theta(x_{initial})\}$, where $x_{initial}$ is the initial point in the local search.

In this section, we report the performance of the MADDF algorithm on 14 well-known test problems, which are shown in the Appendix of this paper, in an effort to make the article as self-contained as possible. The MADDF code was applied 30 times to solve each problem.

In the first set of experiments, summarized in Table 1, the stopping rule (16) with $\epsilon = 0.06$ is used. Table 1 summarizes the MADDF results obtained for each

Table 1. Numerical results obtained with MADDF and FSA [5]

Prob.	f_{OPT}	Method	Best	Average	Worst	S.D.	Av. f.eval.
g3	-1	MADDF	-1.0000968	-0.9998019	-0.9993183	0.000208	45466
		in [5]	-1.0000015	-0.9991874	-0.9915186	0.001653	314938
g6	-6961.81388	MADDF	-6961.23915	-6957.99845	-6954.65040	1.92544	15544
		in [5]	-6961.81388	-6961.81388	-6961.81388	0.000000	44538
g8	-0.095825	MADDF	-0.095825	-0.095825	-0.095825	0.000000	4999
		in [5]	-0.095825	-0.095825	-0.095825	0.000000	56476
g9	680.630057	MADDF	681.08698	683.31319	685.49488	1.38392	38099
		in [5]	680.63008	680.63642	680.69832	0.014517	324596
g11	0.75	MADDF	0.749980	0.750204	0.751048	0.000295	139622
		in [5]	0.749999	0.749999	0.749999	0.000000	23722

test problem as well as the best known objective function value for each problem (f_{OPT}). In order to show more details concerning the quality of the obtained solution, the best ('Best'), the average ('Average'), the worst ('Worst'), as well as the standard deviation ('S.D.') of the obtained objective function values are also reported in Table 1. The average number of function evaluations required to converge to the solution ('Av. f.eval.') is also reported. In this table, the results for each problem using the Filter Simulated Annealing Method (FSA) proposed in [5] are also reported.

Problems g3 and g8 were originally maximization problems. They were rewritten as minimization problems. As it can be seen, for all five problems, MADDF method finds the global minimum. The quality of the solution is good. The worst results are obtained with problems g6 and g9. The average number of function evaluations is much smaller than the one reported by FSA method, for all test problems, except g11. In [5], a comparison with four evolutionary algorithms (EA) was made. These EA methods need a higher number of function evaluations than FSA and, consequently, our proposed method. Hence, the MADDF is better than the EA methods used in [5] as far as the number of function evaluations is concerned. These four EA-based methods are: Homomorphous Mappings (HM) method, Stochastic Ranking (SR) method, Adaptive Segregational Constraint Handling EA (ASCHEA) method and Simple Multimembered Evolution Strategy (SMES) method. In Table 2, the results of the proposed MADDF method are repeated, in order to compare them with those of the EA methods. We may observe that the MADDF method is competitive with the EA methods relative to the quality of the solution.

Table 2. Numerical results obtained with MADDF and EA methods [5]

Prob.	Method	Best	Average	Worst
g3	MADDF	-1.0000968	-0.9998019	-0.9993183
	HM	-0.9997	-0.9989	-0.9978
	SR	-1.000	-1.000	-1.000
	ASCHEA	-1	-0.99989	N.A
	SMES	-1.001038	-1.000989	-1.000579
g6	MADDF	-6961.23915	-6957.99845	-6954.65040
	HM	-6952.1	-6342.6	-5473.9
	SR	-6961.814	-6875.940	-6350.262
	ASCHEA	-6961.81	-6961.81	N.A
	SMES	-6961.813965	-6961.283984	-6961.481934
g8	MADDF	-0.095825	-0.095825	-0.095825
	HM	-0.0958250	-0.0891568	-0.0291438
	SR	-0.095825	-0.095825	-0.095825
	ASCHEA	-0.09582	-0.09582	N.A
	SMES	-0.095826	-0.095826	-0.095826
g9	MADDF	681.08698	683.31319	685.49488
	HM	680.91	681.16	683.18
	SR	680.630	680.656	680.763
	ASCHEA	680.630	680.641	N.A
	SMES	680.631592	680.643410	680.719299
g11	MADDF	0.749980	0.750204	0.751048
	HM	0.75	0.75	0.75
	SR	0.750	0.750	0.750
	ASCHEA	0.75	0.75	N.A
	SMES	0.749090	0.749358	0.749830

To establish other comparisons with other stochastic global methods, we applied the following conditions that appear in [8] to the next set of nine problems and the results are shown in the next two tables. Two conditions to judge the success of the run were applied. First,

$$|f(x^{best}) - f_{OPT}| \leq 10^{-4} |f_{OPT}| \quad (17)$$

where $f(x^{best})$ is the best solution found so far and f_{OPT} is the known optimal solution available in the literature, is used instead of the stopping rule (16).

In practice, when solving any benchmark problem whose global optimal solution is known, the Algorithm 3 is stopped as soon as a sufficiently accurate solution is found, according to the condition in (17). We remark that in multistart clustering methods based on the region of attraction, the stopping rule of the algorithm is crucial to promote convergence to all local optimal solutions (cf. [6]). In the presented algorithm, the likelihood of choosing a point that does not belong to the regions of attraction of previously identified optimal solutions is very high, although convergence to a local minimum that has not been located before is not guaranteed. Convergence to an optimal solution more than once may happen. So far, during the herein presented experiments this situation has occurred although not frequently.

Table 3 contains the average number of function evaluation obtained after the 30 runs. A comparison is made with the results reported in [8] - two artificial fish swarm based methods (AFS and m-AFS) and an electromagnetism-like mechanism algorithm (EM).

Table 3. Average number of function evaluations, using (17)

Prob.	f_{OPT}	MADDF	AFS	m-AFS	EM
BR	0.39789	493	550	475	315
CB6	-1.03160	660	331	247	233
GP	3.00000	787	676	417	420
H3	-3.86278	6022	2930	1891	1114
H6	-3.32237	5001	7091	2580	2341
S5	-10.1532	2396	3928	1183	3368
S7	-10.4029	2655	4033	1103	1782
S10	-10.5364	3514	2069	1586	5620
SBT	-186.731	938	472	523	358

From the table we may conclude that the performance of the proposed MADDF is similar to the AFS algorithm, in terms of efficiency (number of function evaluations), while m-AFS and EM are slightly better than MADDF.

The results shown in Table 4 were obtained using the following stopping condition,

$$|f(x^{best}) - f_{OPT}| \leq 10^{-3} \quad (18)$$

instead of (16). This set of experiments is compared with the results obtained by AFS, m-AFS, two particle swarm algorithms, PSO-RPB and PSO-HS, and a differential evolution method, DE, available in [8].

Table 4. Average number of function evaluations, using (18)

Prob.	f_{OPT}	MADDF	AFS	m-AFS	PSO-RPB	PSO-HS	DE
BR	0.39789	506	651	438	2652	2018	1305
CB6	-1.03160	660	246	245	2561	2390	1127
GP	3.00000	1063	562	485	2817	1698	884
H3	-3.86278	5845	1573	1142	3564	2948	1238
H6	-3.32237	7559	7861	2845	8420	8675	7053
S5	-10.1532	2929	3773	1150	6641	6030	5824
S7	-10.4029	4428	2761	1240	6860	6078	5346
S10	-10.5364	4489	2721	1190	6747	5602	4822
SBT	-186.731	1867	659	516	4206	6216	2430

As it can be seen, the MADDF method has a similar performance to AFS method, outperforms the two variants of the particle swarm optimization and the differential evolution methods, although is less efficient than m-AFS.

4 Conclusions and Future Work

We present a multistart technique based on a derivative-free filter method to solve constrained global optimization problems. The multistart strategy relies on the concept of region of attraction to prevent the repetitive use of the local search procedure in order to avoid convergence to previously found local minima. Our proposal for the local search computes approximate descent directions combined with a (line search) filter method to generate a sequence of approximate solutions that improve either the constraint violation or the objective function value.

A set of 14 well-known test problems was used and the results obtained are very promising. In all problems we could reach the global minimum and the performance of the algorithm, in terms of number of function evaluations and the quality of the solution is quite satisfactory.

In the future, we aim to extend MADDF method to multilocal programming, so that all global as well as local (non-global) minimizers are obtained. This is an interesting and promising area of research due to their real applications in the chemical engineering field.

Acknowledgments. The authors wish to thank three anonymous referees for their valuable comments and suggestions to improve the paper.

This work was financed by FEDER funds through COMPETE-Programa Operacional Fatores de Competitividade and by portuguese funds through FCT-Fundação para a Ciência e a Tecnologia within projects PEst-C/MAT/UI0013/2011 and FCOMP- 01-0124-FEDER-022674.

References

1. Ali, M.M., Gabere, M.N.: A simulated annealing driven multi-start algorithm for bound constrained global optimization. *Journal of Computational and Applied Mathematics* 233, 2661–2674 (2010)
2. Audet, C., Dennis Jr., J.E.: A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization* 14(4), 980–1010 (2004)
3. Costa, M.F.P., Fernandes, E.M.G.P.: Assessing the potential of interior point barrier filter line search methods: nonmonotone versus monotone approach. *Optimization* 60(10-11), 1251–1268 (2011)
4. Fletcher, R., Leyffer, S.: Nonlinear programming without a penalty function. *Mathematical Programming* 91, 239–269 (2002)
5. Hedar, A.R., Fukushima, M.: Derivative-Free Filter Simulated Annealing Method for Constrained Continuous Global Optimization. *Journal of Global Optimization* 35, 521–549 (2006)
6. Lagaris, I.E., Tsoulos, I.G.: Stopping rules for box-constrained stochastic global optimization. *Applied Mathematics and Computation* 197, 622–632 (2008)
7. Marti, R.: Multi-start methods. In: Glover, F., Kochenberger, G. (eds.) *Handbook of Metaheuristics*, pp. 355–368. Kluwer, Dordrecht (2003)
8. Rocha, A.M.A.C., Fernandes, E.M.G.P.: Mutation-Based Artificial Fish Swarm Algorithm for Bound Constrained Global Optimization. In: *Numerical Analysis and Applied Mathematics ICNAAM 2011*. AIP Conf. Proc., vol. 1389, pp. 751–754 (2011)
9. Tsoulos, I.G., Lagaris, I.E.: MinFinder: Locating all the local minima of a function. *Computer Physics Communications* 174, 166–179 (2006)
10. Tu, W., Mayne, R.W.: Studies of multi-start clustering for global optimization. *International Journal for Numerical Methods in Engineering* 53(9), 2239–2252 (2002)
11. Voglis, C., Lagaris, I.E.: Towards "Ideal Multistart". A stochastic approach for locating the minima of a continuous function inside a bounded domain. *Applied Mathematics and Computation* 213, 1404–1415 (2009)

Appendix - Test Problems

– Branin (BR)

$$\min f(x) \equiv (x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10 \left(1 - \frac{1}{8\pi}\right) \cos(x_1) + 10$$

subject to $-5 \leq x_1 \leq 10$
 $0 \leq x_2 \leq 15$

– Camel (CB6)

$$\min f(x) \equiv \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 - 4(1 - x_2^2)x_2^2$$

subject to $-2 \leq x_i \leq 2, i = 1, 2$

– Goldstein and Price (GP)

$$\min f(x) \equiv (1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)) \times$$

$$\times (30 + (2x_1 - 3x_2)^2(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2))$$

subject to $-2 \leq x_i \leq 2, i = 1, 2$

– **Hartman3 (H3)**

$$\min f(x) \equiv - \sum_{i=1}^4 c_i \exp \left(- \sum_{j=1}^3 a_{ij} (x_j - p_{ij})^2 \right)$$

subject to $0 \leq x_i \leq 1, i = 1, 2, 3$
with

$$a = \begin{bmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix}, c = \begin{bmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{bmatrix} \text{ and } p = \begin{bmatrix} 0.3689 & 0.117 & 0.2673 \\ 0.4699 & 0.4387 & 0.747 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.03815 & 0.5743 & 0.8828 \end{bmatrix}$$

– **Hartman6 (H6)**

$$\min f(x) \equiv - \sum_{i=1}^4 c_i \exp \left(- \sum_{j=1}^6 a_{ij} (x_j - p_{ij})^2 \right)$$

subject to $0 \leq x_i \leq 1, i = 1, \dots, 6$

with $a = \begin{bmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{bmatrix},$

$$c = \begin{bmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{bmatrix} \text{ and } p = \begin{bmatrix} 0.1312 & 0.1696 & 0.5569 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4135 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6650 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{bmatrix}$$

– **Shekel-5 (S5)**

$$\min f(x) \equiv - \sum_{i=1}^5 \frac{1}{(x - a_i)(x - a_i)^T + c_i}$$

subject to $0 \leq x_i \leq 10, i = 1, \dots, 4$
with

$$a = \begin{bmatrix} 4 & 4 & 4 & 4 \\ 1 & 1 & 1 & 1 \\ 8 & 8 & 8 & 8 \\ 6 & 6 & 6 & 6 \\ 3 & 7 & 3 & 7 \end{bmatrix} \text{ and } c = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.4 \\ 0.4 \end{bmatrix}$$

– **Shekel-7 (S7)**

$$\min f(x) \equiv - \sum_{i=1}^7 \frac{1}{(x - a_i)(x - a_i)^T + c_i}$$

subject to $0 \leq x_i \leq 10, i = 1, \dots, 4$
with

$$a = \begin{bmatrix} 4 & 4 & 4 & 4 \\ 1 & 1 & 1 & 1 \\ 8 & 8 & 8 & 8 \\ 6 & 6 & 6 & 6 \\ 3 & 7 & 3 & 7 \\ 2 & 9 & 2 & 9 \\ 5 & 3 & 5 & 3 \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.4 \\ 0.4 \\ 0.6 \\ 0.3 \end{bmatrix}$$

– **Shekel-10 (S10)**

$$\min f(x) \equiv - \sum_{i=1}^{10} \frac{1}{(x - a_i)(x - a_i)^T + c_i}$$

subject to $0 \leq x_i \leq 10, i = 1, \dots, 4$
with

$$a = \begin{bmatrix} 4 & 4 & 4 & 4 \\ 1 & 1 & 1 & 1 \\ 8 & 8 & 8 & 8 \\ 6 & 6 & 6 & 6 \\ 3 & 7 & 3 & 7 \\ 2 & 9 & 2 & 9 \\ 5 & 5 & 3 & 3 \\ 8 & 1 & 8 & 1 \\ 6 & 2 & 6 & 2 \\ 7 & 3.6 & 7 & 3.6 \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.4 \\ 0.4 \\ 0.6 \\ 0.3 \\ 0.7 \\ 0.5 \\ 0.5 \end{bmatrix}$$

– **Shubert (SBT)**

$$\min f(x) \equiv \left(\sum_{i=1}^5 i \cos((i+1)x_1 + i) \right) \left(\sum_{i=1}^5 i \cos((i+1)x_2 + i) \right)$$

subject to $-10 \leq x_i \leq 10, i = 1, 2$

– **Problem G3**

$$\min f(x) \equiv -(\sqrt{n})^n \prod_{i=1}^n x_i$$

subject to $\sum_{i=1}^n x_i^2 - 1 = 0$
 $0 \leq x_i \leq 1, i = 1, 2$

– **Problem G6**

$$\min f(x) \equiv (x_1 - 10)^3 + (x_2 - 20)^3$$

subject to $-(x_1 - 5)^2 - (x_2 - 5)^2 + 100 \leq 0$
 $(x_1 - 6)^2 + (x_2 - 5)^2 - 82.81 \leq 0$
 $13 \leq x_1 \leq 100$
 $0 \leq x_2 \leq 100,$

– **Problem G8**

$$\begin{aligned} \min f(x) &\equiv -\frac{\sin^3(2\pi x_1)\sin(2\pi x_2)}{x_1^3(x_1+x_2)} \\ \text{subject to } &x_1^2 - x_2 + 1 \leq 0 \\ &1 - x_1 + (x_2 - 4)^2 \leq 0 \\ &0 \leq x_i \leq 10, i = 1, 2 \end{aligned}$$

– **Problem G9**

$$\begin{aligned} \min f(x) &\equiv (x_1 - 10)^2 + 5(x_2 - 12)^2 + x_3^4 + 3(x_4 - 11)^2 + \dots \\ &\quad + 10x_5^6 + 7x_6^2 + x_7^4 - 4x_6x_7 - 10x_6 - 8x_7 \\ \text{subject to } &v_1 + 3v_2^2 + x_3 + 4x_4^2 + 5x_5 - 127 \leq 0 \\ &7x_1 + 3x_2 + 10x_3^2 + x_4 - x_5 - 282 \leq 0 \\ &23x_1 + v_2 + 6x_6^2 - 8x_7 - 196 \leq 0 \\ &2v_1 + v_2 - 3x_1x_2 + 2x_3^2 + 5x_6 - 11x_7 \leq 0 \\ &-10 \leq x_i \leq 10, i = 1, \dots, 7 \end{aligned}$$

$$\text{with } v_1 = 2x_1^2; v_2 = x_2^2$$

– **Problem G11**

$$\begin{aligned} \min f(x) &\equiv -x_1^2 + (x_2 - 1)^2 \\ \text{subject to } &x_2 - x_1^2 - 1 = 0 \\ &-1 \leq x_i \leq 1, i = 1, 2 \end{aligned}$$

On Lower Bounds Using Additively Separable Terms in Interval B&B*

José L. Berenguel¹, Leocadio G. Casado², I. García³, Eligius M.T. Hendrix³,
and F. Messine⁴

¹ TIC 146: Supercomputing-Algorithms Research Group, University of Almería,
Agrifood Campus of International Excellence (ceiA3), 04120, Spain
jlberenguel@gmail.com

² Department of Computer Architecture and Electronics, University of Almería,
Agrifood Campus of International Excellence (ceiA3), 04120, Spain
leo@ual.es

³ Department of Computer Architecture, University of Málaga, Campus de Teatinos,
29017, Spain
igarcia@ual.es, Eligius@uma.es

⁴ University of Toulouse, ENSEEIHT-IRIT UMR-CNRS-5505, 2 rue Camichel, 31000
Toulouse, France
Frederic.Messine@n7.fr

Abstract. Interval Branch-and-Bound (B&B) algorithms are powerful methods which aim for guaranteed solutions of Global Optimisation problems. Lower bounds for a function in a given interval can be obtained directly with Interval Arithmetic. The use of lower bounds based on Taylor forms show a faster convergence to the minimum with decreasing size of the search interval. Our research focuses on one dimensional functions that can be decomposed into several terms (sub-functions). The question is whether using this characteristic leads to sharper bounds when based on bounds of the sub-functions. This paper deals with functions that are an addition of two sub-functions, also called additively separable functions. The use of the separability is investigated for the so-called Baumann form and Lower Bound Value Form (LBVF). It is proven that using the separability in the LBVF form may lead to a combination of linear minorants that are sharper than the original one. Numerical experiments confirm this improving behaviour and also show that not all separable methods do always provide sharper lower bounds.

Keywords: Branch-and-Bound, Interval methods, Separable functions.

* This work has been funded by grants from the Spanish Ministry of Science and Innovation (TIN2008-01117), and Junta de Andalucía (P11-TIC-7176), in part financed by the European Regional Development Fund (ERDF). Eligius M.T. Hendrix is a fellow of the Spanish “Ramón y Cajal” contract program, co-financed by the European Social Fund.

1 Introduction

Interval Branch-and-Bound methods aim for guaranteed solutions of Global Optimisation problems. Consider the one dimensional generic interval constrained global optimisation problem, which is to find

$$f^* = \min_{x \in S} f(x) \quad (1)$$

where $S \in \mathbb{I}$ is the search region and \mathbb{I} stands for the set of all one-dimensional closed real intervals.

Definition 1. *Function $f : S \subset \mathbb{R} \rightarrow \mathbb{R}$ is additively separable, if it can be written as*

$$f(x) = \sum_{j=1}^p f_j(x), \quad x \in S. \quad (2)$$

We have

$$\min_{x \in S} f(x) \geq \sum_{j=1}^p \min_S f_j(x). \quad (3)$$

Let \underline{F}_j be a lower bound of f_j over S . Then we have

$$\min_{x \in S} f(x) \geq \sum_{j=1}^p \underline{F}_j. \quad (4)$$

To create a lower bound \underline{F} of f over interval X in an interval B&B framework, can be done in several ways. Sharper bounds are better, i.e. higher values of \underline{F} lead to more efficient performance of the B&B algorithm. Considering functions that have an additively separable structure (2), our research question is: for which cases

$$\underline{F} \leq \sum_{j=1}^p \underline{F}_j? \quad (5)$$

Alternatively, the question is to find ways to combine minorants on the separable terms, such that we get sharper bounds.

Example 1. Consider function $f(x) = f_1(x) + f_2(x) = (x + 1)^2 + (x - 1)^2$ on the interval $X = [-2, 2]$. The minima of the sub-functions is 0, whereas the minimum of f itself is $f(0) = 2$. Figure 1 illustrates this idea and also draws lower bounds of all functions based on the so-called Baumann point that will be explained in the following section. The lower bound of f , based on the red dashed minorant is $\underline{F} = -14$. The sub-functions have a lower bound of $\underline{F}_j = -6$ such that $\underline{F}_1 + \underline{F}_2 = -12$, illustrating question (5).

Apart from studying the consequences of an additively separable Baumann lower bound (see Example 1), we also focus on a so-called additively separable Lower Boundary Value Form. We present a new lower bounding method for additive separable inclusion functions that produces sharper lower bounds.

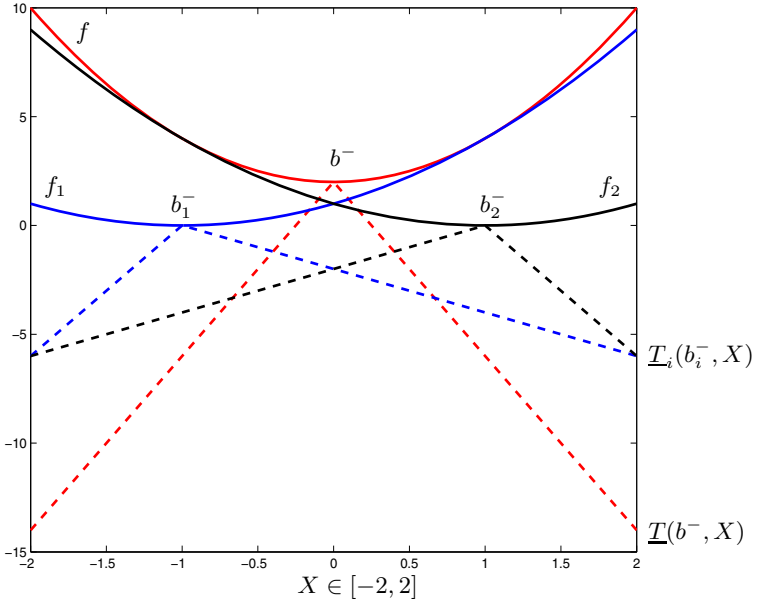


Fig. 1. Quadratic illustration of (3) and (5)

The remainder of this paper is organised as follows. Section 2 introduces relevant Interval Arithmetic properties and general lower bounding concepts. Section 3 discusses several ways to combine lower bounds of sub-functions for an additively separable function. Finally, a numerical illustration and conclusions are presented in Sections 4 and 5, respectively.

2 Properties of Interval Inclusion Functions

Algorithms based on Interval Arithmetic (IA) have several ingredients. We start with the generic ideas and then focus on how the mathematical characteristics can be refined for separable functions. IA has been widely studied in the last forty years [5,6]. We mention several relevant IA properties and definitions.

Definition 2. Let $f(X)$ be the range of f on X . An interval function $F : \mathbb{I}^n \rightarrow \mathbb{I}$ is an inclusion function, if $f(X) \subseteq F(X) = [\underline{F}(X), \overline{F}(X)]$.

An interesting property of IA is that it generates isotone inclusion functions.

Definition 3. Inclusion Isotonicity. Inclusion function $F : \mathbb{I}^n \rightarrow \mathbb{I}$ of f is inclusion isotone, if $\forall (X, Y) \in \mathbb{I}, X \subseteq Y \Rightarrow F(X) \subseteq F(Y)$.

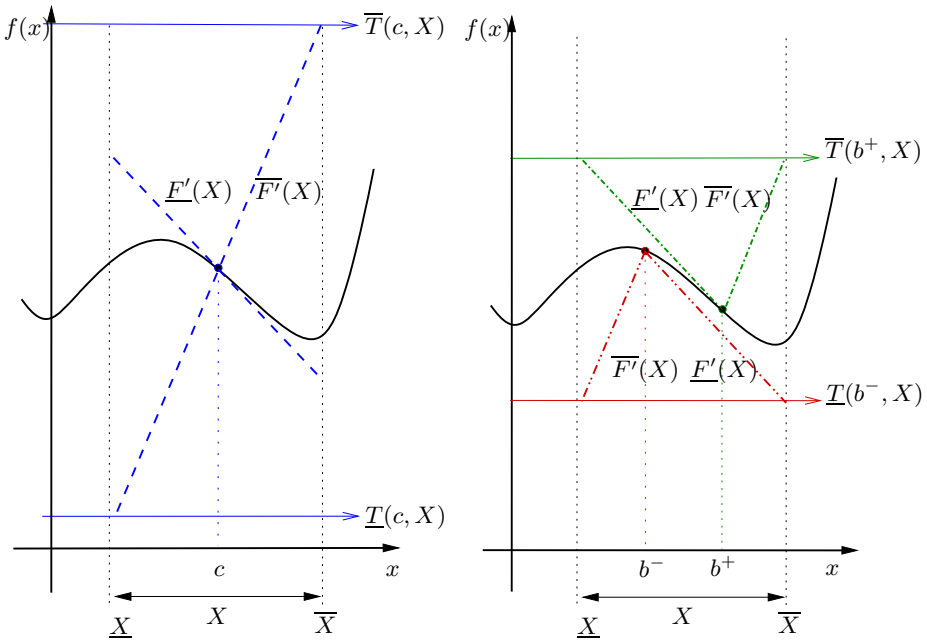


Fig. 2. Center (left) and Baumann (right) forms

Besides the standard IA bounding, called “natural interval extension”, one can obtain an inclusion function F of f using the inclusion function F' of f' . Consider the first order Taylor expression

$$T(c, X) := f(c) + (X - c)F'(X), \tag{6}$$

where $c \in X$. Notice that this expression is mainly of interest if the function is not monotonous on X , so at least $0 \in F'(X)$. By taking for c the middle $m = \frac{X + \bar{X}}{2}$ of the interval, we have what is called a center form of the inclusion, as illustrated at the left graph in Fig. 2. In [1], Baumann proves that taking $c = b^-$ in the Taylor expression, leads to the best lower bound, where:

$$b^- = \begin{cases} \frac{X\bar{F}'(X) - \bar{X}F'(X)}{F'(X) - F'(X)}, & 0 \in F'(X) \\ \bar{X} & , \bar{F}'(X) \leq 0 \\ X & , F'(X) \geq 0 \end{cases}$$

So,

$$f(X) \geq \underline{T}(b^-, X). \tag{7}$$

To obtain the best upper bound, the point of b^- should be reflected in midpoint $m(X)$ leading to:

$$b^+ = m(X) + (m(X) - b^-).$$

The choice of the Baumann points, to get upper and lower bounds, is illustrated in Fig. 2 at the right. These points can also be used to reduce the search space as shown by 9. As illustrated in Fig. 1, b^- is the minimum point for quadratic functions. Figure 1 shows the possibility of having different points b^- for each sub-function and to combine the obtained lower bounds rather than using b^- of the composite function itself.

Another way to compose derivative based linear minorants is the so-called Lower Boundary Value Form (LBVF), ([7] p. 60 and [24]) that uses the evaluation of the end-points of the interval. Consider the most left point of X . Function

$$\varphi l(x) = \underline{F}(\underline{X}) + \underline{F}'(X)(x - \underline{X}), \tag{8}$$

provides an affine minorant. Similarly, the right most point of X provides

$$\varphi r(x) = \underline{F}(\overline{X}) - \overline{F}'(X)(\overline{X} - x) = \underline{F}(\overline{X}) + \overline{F}'(X)(x - \overline{X}). \tag{9}$$

The values $\varphi l(\overline{X})$ and $\varphi r(\underline{X})$ are lower bounds of $f(X)$ over X . A sharper lower bound can be obtained when $0 \in F'(X)$ by combining (8) and (9) in lower bounding function

$$\varphi m(x) = \max\{\varphi l(x), \varphi r(x)\}. \tag{10}$$

The Lower Boundary Value Form $\underline{\varphi m}(X)$ follows from finding y for which (8) and (9) are equal:

$$y = \frac{\underline{F}(\underline{X}) - \underline{F}(\overline{X})}{w(F'(X))} + \frac{\overline{X} \cdot \overline{F}'(X) - \underline{X} \cdot \underline{F}'(X)}{w(F'(X))}, \tag{11}$$

with $w(X) = \overline{X} - \underline{X}$ the width of interval X . Evaluation of (8) for y provides

$$\underline{\varphi m}(X) = \varphi m(y) = \frac{\underline{F}(\underline{X})\overline{F}'(X) - \underline{F}(\overline{X})\underline{F}'(X)}{w(F'(X))} + \frac{w(X)\overline{F}'(X)\underline{F}'(X)}{w(F'(X))}. \tag{12}$$

Similarly, linear majorants of $f(X)$ can be constructed. Using the left most point of X gives

$$\xi l(x) = \overline{F}(\underline{X}) + \overline{F}'(X)(x - \underline{X}), \tag{13}$$

and the right most point of X provides

$$\xi r(x) = \overline{F}(\overline{X}) - \underline{F}'(X)(\overline{X} - x) = \overline{F}(\overline{X}) + \underline{F}'(X)(x - \overline{X}). \tag{14}$$

The values $\overline{\xi l}(\overline{X})$ and $\overline{\xi r}(\underline{X})$ are upper bounds of $f(X)$. A sharper majorant can be obtained by combining both functions.

$$\xi m(x) = \min\{\xi l(x), \xi r(x)\}. \tag{15}$$

The Upper Boundary Value Form follows from equating (13) and (14)

$$z = \frac{\overline{F}(\overline{X}) - \overline{F}(\underline{X})}{w(F'(X))} + \frac{\underline{X}\overline{F}'(X) - \overline{X}\underline{F}'(X)}{w(F'(X))}, \tag{16}$$

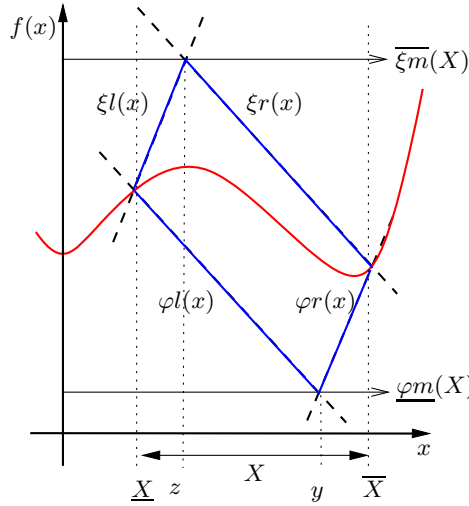


Fig. 3. Graphical illustration of one-dimensional Lower and Upper Boundary Value Form

and substitution of z in (15):

$$\overline{\xi m}(X) = \xi m(z) = \frac{\overline{F}(\overline{X})\overline{F}'(X) - \overline{F}(\underline{X})\underline{F}'(X)}{w(F'(X))} - \frac{w(X)\overline{F}'(X)\underline{F}'(X)}{w(F'(X))}. \quad (17)$$

The bounding functions $\varphi m(x)$ and $\xi m(x)$ are illustrated in Fig. 3. Summarizing:

$$f(X) \geq \underline{\varphi m}(X), \quad 0 \in F'(X), \quad (18)$$

$$f(X) \leq \overline{\xi m}(X), \quad 0 \in F'(X). \quad (19)$$

The Baumann minorant and LBVF minorant $\varphi m(x)$ can be combined to construct sharper lower bounds as presented in [10].

3 Additively Separate Lower Bounds

Consider the case where a one-dimensional function f is additively separable in two sub-functions, i.e.

$$f(x) = f_1(x) + f_2(x).$$

If both sub-functions are monotonous in the same direction (either $\underline{F}'_j(X) > 0$, or $\overline{F}'_j(X) < 0$), then the function f is also monotonous in X . If both sub-functions are monotonous in different directions, w.l.o.g. $\underline{F}'_1(X) > 0$ and $\overline{F}'_2(X) < 0$, it may happen that $0 \in F'(X)$. Moreover, one of the sub-functions may be monotonous and the other not. In general, when the complete function is monotonous, i.e. $0 \notin F'(X)$, the function value in one of the extreme points is the minimum on X . So, a lower bound is given by $\min\{\underline{F}(\underline{X}), \underline{F}(\overline{X})\}$. Separability may improve the bound only in the case $0 \in F'(X)$.

3.1 Additively Separable Baumann Lower Bound

An additively separable Baumann form bound $ASB(X)$ can be constructed in a straightforward way evaluating the Taylor expression (6) for the two sub-functions in their Baumann point and adding the resulting lower bounds,

$$f(X) \geq ASB(X) = \underline{T}_1(b_1^-, X) + \underline{T}_2(b_2^-, X). \tag{20}$$

The question is whether (20) always provides a sharper lower bound than using the standard Baumann form (7). We investigate this experimentally.

3.2 Additively Separable Lower Bound Value Form (ASLBV)

Following the same reasoning as done for the separable Baumann form, leads for the separable LBVF to

$$f(X) \geq ASLBV(X) = \underline{\varphi}m_1(X) + \underline{\varphi}m_2(X), \quad 0 \in F'_1(X), \quad 0 \in F'_2(X). \tag{21}$$

However, this way of reasoning is only of interest if both sub-functions are not monotonous in the evaluated interval. Again, if F is monotonous on X , using separability does not make sense. In the other case, one can combine $\min\{f_j(\underline{X}), f_j(\overline{X})\}$ with (12). The question is whether this procedure gives a better lower bound than using (18). We investigate this experimentally.

3.3 Bound $ASLB\varphi$ Based on New Minorant φ

We focus further on the LBVF minorants of both sub-functions in order to obtain a sharper lower bound than $\underline{\varphi}m(X)$ without worrying about the monotonicity of the sub-functions for a given interval. Notice again, that we are only interested in the case where the composite function f is not monotonous, $0 \in F'(X)$. Consider the addition of the separate minorant terms

$$\varphi(x) = \varphi m_1(x) + \varphi m_2(x),$$

where φm_i is defined by (10). First of all, notice that φ is a piecewise linear minorant function and the maximum of four different affine terms:

$$\varphi(x) = \max \left\{ \begin{array}{l} \varphi l(x) := \varphi l_1(x) + \varphi l_2(x) \\ \varphi a(x) := \varphi l_1(x) + \varphi r_2(x) \\ \varphi b(x) := \varphi r_1(x) + \varphi l_2(x) \\ \varphi r(x) := \varphi r_1(x) + \varphi r_2(x) \end{array} \right\}. \tag{22}$$

Then one can see that $\varphi(x)$ is a sharper minorant than $\varphi m(x)$.

Theorem 1. *Let $\forall x \in X, f(x) = f_1(x) + f_2(x)$ and $\varphi l, \varphi r$ and φm be defined by (8), (9) and (10). $\forall x \in X, \varphi m_1(x) + \varphi m_2(x) \geq \varphi m(x)$.*

Proof

Given equivalence (22), we have that

$$\varphi m(x) = \max\{\varphi l(x), \varphi r(x)\} \leq \max\{\varphi l(x), \varphi a(x), \varphi b(x), \varphi r(x)\} = \varphi(x).$$

■

Given the result of the theorem, we are interested in the minimum $\min_{x \in X} \varphi(x)$. In order to find the minimum, we consider several properties of the piecewise linear minorant function $\varphi(x)$.

The maximum of φ can typically be found at the boundary.

Proposition 1. *Let $0 \in F'(X)$ and φ be defined by (22). Then*

$$\varphi(\underline{X}) = \varphi l(\underline{X}) \geq \max\{\varphi a(\underline{X}), \varphi b(\underline{X}), \varphi r(\underline{X})\}$$

and

$$\varphi(\overline{X}) = \varphi r(\overline{X}) \geq \max\{\varphi a(\overline{X}), \varphi b(\overline{X}), \varphi l(\overline{X})\}.$$

Proof

The inequalities follow from writing explicitly the terms in (22) and considering $\varphi l_j(\underline{X}) \geq \varphi r_j(\underline{X})$ and $\varphi r_j(\overline{X}) \geq \varphi l_j(\overline{X})$. ■

Proposition 2. *Let $0 \in F'(X)$ and φ be defined by (22). Function φ is a convex minorant of f on X .*

Proof

This follows from φ being a maximum of convex functions, according to (22). ■

Due to φ being piecewise linear, it has only one minimum with either a unique minimum point or infinitely many minimum points. The first order conditions for a minimum point of a piecewise linear function state that it should have a subgradient of 0. Points with more than one derivative (subgradients) are the intersection points of the affine minorants. Writing out the terms shows that these intersection points are typically y_1 and y_2 (11):

- $\varphi l(x) = \varphi a(x) \rightarrow \varphi l_2(x) = \varphi r_2(x)$ with solution y_2 ,
- $\varphi l(x) = \varphi b(x) \rightarrow \varphi l_1(x) = \varphi r_1(x)$ with solution y_1 ,
- $\varphi r(x) = \varphi a(x) \rightarrow \varphi r_1(x) = \varphi l_1(x)$ with solution y_1 ,
- $\varphi r(x) = \varphi b(x) \rightarrow \varphi r_2(x) = \varphi l_2(x)$ with solution y_2 ,
- $\varphi l(x) = \varphi r(x)$ with solution y (11), being the minimum point of φm , is not a unique minimum point of φ due to Theorem 1, $\varphi(x) \geq \varphi m(x)$ is a sharper minorant
- $\varphi a(x) = \varphi b(x)$ can be shown to produce the following intersection point

$$\hat{x} = \frac{\underline{X}(F'_1 - F'_2) + \overline{X}(\overline{F}'_2 - \overline{F}'_1) + \underline{F}_1(\overline{X}) - \underline{F}_1(\underline{X}) + \underline{F}_2(\underline{X}) - \underline{F}_2(\overline{X})}{w(F'_2) - w(F'_1)}.$$

(23)

We will prove, that \hat{x} is not a candidate for minimum point of φ .

The following proposition shows that one can focus on the interval between y_1 and y_2 .

Proposition 3. *Let $0 \in F'(X)$, φ be defined by (22) and y_j by (11). Then $\min_{x \in X} \varphi(x)$ is attained in the interval $[\min\{y_1, y_2\}, \max\{y_1, y_2\}]$.*

Proof

Following the argumentation in Proposition 1 and having excluded y to be a unique minimum point, we have

- $\forall x \in [\underline{X}, y_2] \varphi l(x) \geq \varphi a(x)$
- $\forall x \in [\underline{X}, y_1] \varphi l(x) \geq \varphi b(x)$

and $\varphi' \leq 0$, so that no unique minimum is attained in $[\underline{X}, \min\{y_1, y_2\}]$. From symmetry with respect to φr follows no unique minimum is attained in $(\max\{y_1, y_2\}, \bar{X}]$ ■

Now we can characterise the exact minimum of $\varphi(x)$ on X , which helps to generate a sharper lower bound than minorant φm .

Theorem 2. *Let $0 \in F'(X)$, φ be defined by (22) and y_j by (11). Then $\min_{x \in X} \varphi(x)$ is attained in $\{y_1, y_2\}$.*

Proof

Proposition 3 limits the minimum to interval $[\min\{y_1, y_2\}, \max\{y_1, y_2\}]$. Consider $y_1 \leq y_2$.

On $[y_1, y_2]$ we have $\varphi r_1(x) \geq \varphi l_1(x)$ and $\varphi l_2(x) \geq \varphi r_2(x)$, such that $\varphi(x) = \varphi b(x) \geq \varphi a(x)$. So, $\varphi(x)$ is affine on $[y_1, y_2]$, attaining its minimum in one of the endpoints.

Consider the case $y_2 \leq y_1$.

On $[y_2, y_1]$ we have $\varphi l_1(x) \geq \varphi r_1(x)$ and $\varphi r_2(x) \geq \varphi l_2(x)$, such that $\varphi(x) = \varphi a(x) \geq \varphi b(x)$. So, $\varphi(x)$ is affine on $[y_2, y_1]$, attaining its minimum in one of the endpoints. ■

Notice that the reasoning in the proof also shows that point \hat{x} , where φa and φb intersect, cannot be a unique minimum point. The theory provides us with a new lower bound $ASLB\varphi$ defined by

$$f(X) \geq ASLB\varphi(X) = \underline{\varphi}(X) = \min\{\varphi(y_1), \varphi(y_2)\}. \tag{24}$$

4 Numerical Illustration

To measure the effectiveness of the different lower bounds, we use the following experiment. An interval B&B global optimization algorithm is run on a test bed of 18 one dimensional functions. We first describe the set of test function in Sect. 4.1. The used Algorithm is presented in Sect. 4.2. The performance indicators and their experimental values are provided in Sect. 4.3.

4.1 Test Problems

Table 1 describes the test functions: the first column is a number to index the functions, f_1 and f_2 are the additive terms of function $f = f_1 + f_2$ and last column provides a reference to literature where the function is described. The search space for all functions is taken as $S = [0.5, 20]$ apart from instances N. 6 and N. 11 where $S = [-10, 10]$.

Table 1. Test problems

N.	f_1	f_2	Ref.
1	e^{-3x}	$-\sin^3 x$	[8]
2	xe^{-x^2}	$\sin xe^{-x^2}$	[8]
3	$\sin x + \ln x$	$\sin \frac{10x}{3} - 0.84x$	[8]
4	$x^4 - 10x^3$	$35x^2 - 50x + 24$	[8]
5	$24x^4 - 142x^3$	$303x^2 - 276x + 93$	[8]
6	$2x^2$	$-\frac{3}{100}e - (200(x - 0.0675))^2$	[8]
7	$\frac{x^2}{20}$	$-\cos x + 2$	[8]
8	x^2	$-\cos(18x)$	[8]
9	$x^4 - 12x^3$	$47x^2 - 60x - 20e^{-x}$	[8]
10	$x^6 + 250$	$-15x^4 + 27x^2$	[8]
11	$\sin^2(1 + \frac{x-1}{4})$	$(\frac{x-1}{4})^2$	[8]
12	$(x - x^2)^2$	$(x - 1)^2$	[8]
13	$-\sum_{k=1}^5 k \sin [(k + 1)x + k] + 3$	$(3x - 1.4) \sin(18x) + 1.7$	[3]
14	$-x + \sin(3x) + 1$	$(3x - 1.4) \sin(18x) + 1.7$	[3]
15	$-\sum_{k=1}^5 k \sin [(k + 1)x + k] + 3$	$\sum_{k=0}^5 k \cos [(k + 1)x + k] + 12$	[3]
16	$-\sum_{k=1}^5 k \sin [(k + 1)x + k] + 3$	$\cos x - \sin 5x + 1$	[3]
17	$-x + \sin(3x) + 1$	$\sum_{k=0}^5 k \cos [(k + 1)x + k] + 12$	[3]
18	$-x + \sin(3x) + 1$	$\sum_{k=1}^5 -\cos[(k + 1)x] + 4$	[3]

An interval branch-and-bound global optimization algorithm was used to evaluate the presented additively separable lower bounds computation methods.

4.2 Interval B&B Algorithm

Algorithm 1 shows the use of the basic B&B rules. The selection rule selects the interval with the best lower bound of $f(X)$ (line 4) to be processed next. We use bisection as division rule (line 9). An interval is stored in the final list when its width is smaller than a determined accuracy. In all the experiments we use as termination rule $width(X) < 10^{-6}$ (line 13). An upper bound of the global minimum f^U is computed evaluating the midpoint of selected intervals (line 7). For each interval, a lower bound of f is calculated (line 11). Actually, that is where the different described lower bounds come in. Intervals with $\underline{F}(X) > f^U$ are rejected (lines 8 and 12). The monotonicity test is also applied to reject or reduce intervals when $0 \notin F'(X)$. Therefore, Algorithm 1 only rejects an interval based on lower bounds ($\underline{F}(X) > f^U$) when f is not monotonous on the interval.

The lower bounds used in the rejection are based on natural interval extension $\underline{F}(X)$. For each non-rejected interval, we can then measure the effectiveness of one of the other bounds.

4.3 Experimental Results

To measure the effectiveness of the new bounds, we use the natural interval extension as benchmark. For each non-rejected interval, we measure whether it would have been rejected if one of the other lower bounds would have been used. The following lower bounds are evaluated.

- Baumann calculated as $\underline{T}(b^-, X)$ in (7).
- Additively separable Baumann ASB , see (20).
- Lower Boundary Value Form $LBVF$ calculated as $\underline{\varphi m}(X)$ in (12).
- Additively separable Lower Boundary value $ASLBV$, see (21).
- $ASLB\varphi(X)$, see (24).

The results in Table 2 concern the measurement of the following: the number of the test function, number of iterations of Algorithm 1, number of intervals rejected by $\underline{F}(X) > f^U$, number of intervals rejected by monotonicity test MT, and number of intervals that might have been eliminated when using the other lower bounds. $ASLBV$ is computed only if both sub-functions are not monotonous, while $ASLB\varphi(X)$ is computed whether subfunctions are monotonous or not.

Algorithm 1. General interval B&B algorithm.

Funct $IBB(S, f)$

1. Set the working list $L := \{S\}$ and the final list $Q := \emptyset$
 2. Set $f^U = \overline{F}(m(S))$
 3. **while** ($L \neq \emptyset$)
 4. Select an interval X from L *Selection rule*
 5. Compute $\overline{F}(m(X))$
 6. **if** $\overline{F}(m(X)) < f^U$
 7. $f^U = \overline{F}(m(X))$
 8. Remove all $X \in L \cup Q$ with $\underline{F}(X) > f^U$ *Cut-off test*
 9. Divide X into subintervals $X^j, j = 1, \dots, k$ *Division rule*
 10. **for** $j = 1$ to k
 11. Compute a lower bound $lb(X)$ of $f(X^j)$ *Bounding rule*
 12. **if** X^j cannot be eliminated *Elimination rule*
 13. **if** X^j satisfies the termination criterion *Termination rule*
 14. Store X^j in Q
 15. **else**
 16. Store X^j in L
 17. **return** Q, f^U
-

Table 2. Additional rejection power of lower bounds

N.	rejected by			Possible additional rejections				
	Iter	$\underline{F}(X) > f^U$	MT	Baumann	ASB	LBVF	ASLBV	ASLB ϕ
1	25	16	9	0	0	0	0	0
2	114	110	5	0	0	1	1	2
3	27	5	22	0	0	0	0	0
4	1272	118	1093	878	57	1006	1006	1006
5	1529	313	1153	1244	156	1340	1340	1340
6	25	13	12	0	0	0	0	0
7	27	4	23	0	0	0	0	0
8	25	8	17	0	0	0	0	0
9	202	82	117	94	33	121	121	121
10	105	7	94	15	0	38	38	38
11	34	5	29	0	0	0	0	0
12	26	21	4	0	0	0	0	0
13	32	15	17	0	0	0	0	0
14	25	12	13	0	0	0	0	0
15	146	74	69	8	2	17	17	17
16	129	67	60	2	2	7	7	7
17	39	20	18	1	1	1	1	1
18	27	8	19	0	0	0	0	0

Table 3. Lower bound improvement

N.	ASB		ASLBV		ASLB ϕ	
	+	-	+	-	+	-
1	24	0	0	0	1	0
2	20	0	11	0	19	0
3	10	19	1	4	20	0
4	9	1324	2	2	782	0
5	10	1582	0	0	1102	0
6	16	10	1	7	17	0
7	6	21	0	0	16	0
8	17	14	0	0	22	0
9	10	197	3	1	174	0
10	9	100	0	0	79	0
11	6	28	1	1	20	0
12	34	2	1	14	21	0
13	25	11	6	5	29	0
14	17	14	2	7	23	0
15	76	73	19	41	127	0
16	69	64	20	37	111	0
17	24	21	7	9	34	0
18	14	16	3	7	21	0

The LBVF based methods show the best additional rejection power for this test bed. Using the separability, only for function number 2, $ASLB\varphi(X)$ might reject only one additional interval. An interesting result is that for the Baumann form, using the separability, appears to be a bad idea. The additional rejection power over the standard natural interval extension is worse.

We now zoom in on the improvement of considering a bound from a separable perspective. To do so, we compare the Baumann and LBVF lower bound with the variants developed for separable functions. Table 3 show the number of intervals where the separable based lower bound is better (+) and worse (-). As expected from earlier results, considering Baumann from a separable perspective via ASB gives on average worse lower bounds. Using the straightforward idea of separating the LBVF into two forms as in $ASLBV$ neither shows a lot of gain. However, when refining towards $ASLB\varphi$, one can observe bound improvements for all test functions.

5 Conclusions

Interval Branch-and-Bound (B&B) algorithms are powerful methods which aim for guaranteed solutions of Global Optimisation problems. Our research question is whether when, using the separable structure of functions, one can derive sharper bounds based on bounds of the sub-functions. Several ways were discussed to extend the so-called Baumann form and Lower Bound Value Form (LBVF) for separable functions. The separable variant for the Baumann lower bound is usually worse than the original one.

For one of the variants called $ASLBV\varphi$, it is proven that the corresponding minorant is sharper than the standard one for LBVF. Numerical experiments confirm this improving behaviour. Unfortunately, $ASLBV\varphi$, compared with LBVF, does not reduce the number of iterations carried out by an interval global optimization algorithm.

Future investigation could focus on the question how to extend the $ASLBV\varphi$ lower bound for n-dimensional functions. Another question is the derivation of specific interval based bounds for multiplicative terms.

References

1. Baumann, E.: Optimal centered forms. BIT 28(1), 80–87 (1988), doi:10.1007/BF01934696
2. Casado, L., García, I., Martínez, J., Sergeyev, Y.D.: New interval analysis support functions using gradient information in a global minimization algorithm. Journal of Global Optimization 25(4), 345–362 (2003), doi:10.1023/A:1022512411995
3. Casado, L., García, I., Sergeyev, Y.: Interval algorithms for finding the minimal root in a set of multiextremal one-dimensional nondifferentiable functions. SIAM Journal on Scientific Computing 24(2), 359–376 (2002), doi:10.1137/S1064827599357590

4. Hansen, P., Lagouanelle, J.L., Messine, F.: Comparison between baumann and admissible simplex forms in interval analysis. *Journal of Global Optimization* 37, 215–228 (2007), doi:10.1007/s10898-006-9045-9
5. Moore, R.: *Interval analysis*. Prentice-Hall, New Jersey (1966)
6. Moore, R., Kearfott, R., Cloud, M.: *Introduction to Interval analysis*. SIAM, Philadelphia (2009)
7. Neumaier, A.: *Interval Methods for Systems of Equations*. Cambridge University Press, Cambridge (1990)
8. Ratz, D.: A nonsmooth global optimization technique using slopes - the one-dimensional case. *Journal of Global Optimization* 14, 365–393 (1999), doi:10.1023/A:1008391326993
9. Tóth, B., Casado, L.: Multi-dimensional pruning from baumann point in an interval global optimization algorithm. *Journal of Global Optimization* 38(2), 215–236 (2007), doi:10.1007/s10898-006-9072-6
10. Vinkó, T., Lagouanelle, J.L., Csendes, T.: A new inclusion function for optimization: Kite – the one dimensional case. *Journal of Global Optimization* 30, 435–456 (2004), doi:10.1007/s10898-004-8430-5

A Genetic Algorithm for the Job Shop on an ASRS Warehouse

José Figueiredo, José A. Oliveira, Luis Dias, and Guilherme A.B. Pereira

Centre ALGORITMI, University of Minho,
4700-057 Braga, Portugal

josefilipefig@gmail.com, {zan,lsd,gui}@dps.uminho.pt

Abstract. This paper describes the application of a metaheuristic to a real problem that arises within the domain of loads' dispatch inside an automatic warehouse. The truck load operations on an automated storage and retrieval system warehouse could be modeled as a job shop scheduling problem with recirculation. The genetic algorithm is based on random key representation, that is very easy to implement and it allows the use of conventional genetic operators for combinatorial optimization problems. This genetic algorithm includes specific knowledge of the problem to improve its efficiency. A constructive algorithm based in Giffler-Thompson's algorithm is used to generate non delay plans. The constructive algorithm reads the chromosome and decides which operation is scheduled next. This option increases the efficiency of the genetic algorithm. The algorithm was tested using some instances of the real problem and computational results are presented.

Keywords: Genetic Algorithm, Random Keys, Job Shop, Recirculation, ASRS, Warehouses.

1 Introduction

Automatic storage equipments must be efficient in order to justify the investment they imply and also to provide an alternative to conventional storage systems. The efficiency of an automatic storage system depends, among other factors, on the plan for the loading operations of the trucks. In the AS/RS (Automated Storage and Retrieval System) warehouses, where a large number of truckloads are performed on a daily basis, it is necessary to plan and execute accurately the loading procedures in order to fulfill the delivery deadlines.

This paper describes the application of a metaheuristic to a real problem that arises within the domain of loads' dispatch inside an automatic warehouse. An effective and efficient genetic algorithm is presented to sequence the pallets' retrieval aiming to maximize the warehouse throughput and fulfill the delivery deadlines.

The paper is organized in the following way: next section describes the automatic warehouse type of operations; the third section presents the model adopted, including some remarks about its application and some extensions to the model are also presented; the fourth section is dedicated to the characterization of the solution's

methodology adopted; the fifth section presents computational results of the developed algorithm; finally the conclusions about the work are discussed.

2 The Storage System

Eleven aisles of pallets racks compose the main body of the warehouse, with capacity for forty thousand pallets. There is an automatic stacker crane (also S/R machine, operating in dual command mode) in each aisle to move the pallets from their storage position to the collector at the top of the aisle. Next, several forklift trucks move the pallets and place them inside the trucks. The warehouse has 13 docking bays to load the trucks. Fig. 1 presents a scheme of the warehouse.

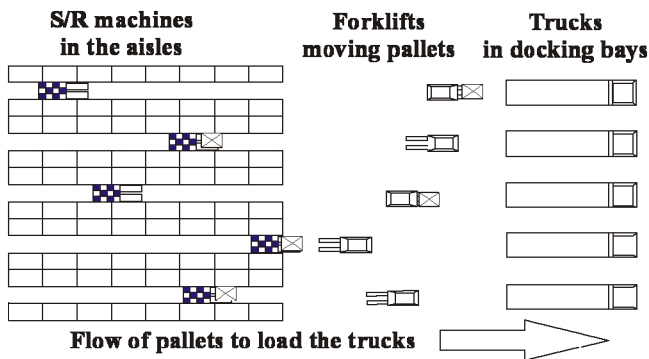


Fig. 1. Scheme of the warehouse

There are about a hundred loads to dispatch per day (truck loads - also called "trips"). The strategy adopted to program the trucks' loading consists in defining a partition of the whole load into disjoint subsets of loads, which are processed simultaneously, called batches (blocks). This strategy guarantees that every load of a batch is finished before starting to prepare the loads for the next batch. The batches' dimensioning respects the imposed limits, in order to fulfill the delivery deadlines. The number of docking bays limits the maximum number of loads in a batch. A standard workday can originate plans with 15 - 20 batches, with 6 - 13 loads each. Fig. 2 shows an example with 3 batches. In the first one, trips 1, 2 and 3 are prepared. In the second batch trucks 4 to 8 are loaded, and the third batch loads the last four trucks - 9 to 12.

The time spent to process the batches depends on the number of loads of the batch, and it is calculated based on the nominal values of the system's throughput. It is assumed that the storage system guarantees a nominal flow of pallets' dispatch at a constant rate, and therefore it considers that the duration of the batch is proportional to the total number of the batches' pallets.

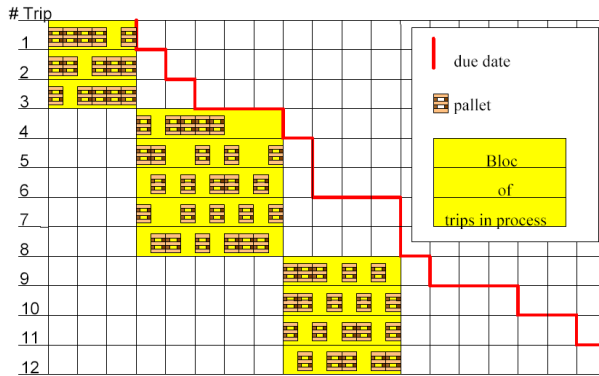


Fig. 2. Strategy of load trucks by batches

We shall demonstrate that time to process the batch depends also on the sequence for retrieving the pallets, which is to say, on the S/R machines' operations scheduling. Due to precedence constraints between pallets, the order in which the pallets of the batches are retrieved by the S/R machines may lead to different processing times. The problem of establishing the best pallets' retrieving operations sequence in each aisle is also an optimization problem. We will see that this problem can be modeled as a Job Shop Scheduling Problem (JSSP).

2.1 Processing Loads

The preparation of a truck's load consists on retrieving the pallets' set, which will be transported in the truck. In one load, the truck can carry pallets for one or more clients. The set of pallets that constitutes a truck's load is determined and known in advance.

The assembly sequence of the pallets in the truck considers the order in which the clients will be visited. The last pallets placed in the truck shall be the first to unload and are destined to the first client that will be visited. Therefore, a load can be seen as a set of pallets with a fixed sequence for its formation, previously established and that may not be altered, from which precedence relations arise between pallets of the same load.

The pallets selected for a specific load can come from any aisle, and the choice obeys to specific criteria [1-2]. The S/R machines' work is programmed and consists on retrieving all the pallets that are selected from their aisle. It is assumed that it is possible to program the S/R machines' activity in different ways. This means that it is possible to establish different sequences for retrieving the pallets in the same aisle.

2.2 Organizing the Process and Processing Time

In Oliveira [2,3] there is a detailed description of the process to retrieve the pallets - from the aisle to the truck. Part of the movement is done by forklifts that are

controlled by Radio Frequency. In his work, Oliveira [2,3] assumes identical processing times to transport pallets independently of the location of the aisle and the truck.

In this work, a new model to consider different processing times is presented, and it takes into account the location of the truck in the docking bays and the aisle where the pallets are retrieved. This model represents better the real problem, but also turns it more difficult to solve the associated job shop problem.

3 Job Shop Scheduling Problem

Some Combinatorial Optimization Problems are very hard to solve and therefore require the use of heuristic procedures. One of them is the Job Shop Scheduling Problem. The use of exact methods to solve the JSSP is limited to the instances of small size. According to Zhang et al. [4] the Branch and Bound methods do not solve instances larger than 250 operations in a reasonable time. As stated in Liu et al. [5] in practical manufacturing environments the scale of job shop scheduling problems could be much larger. They exemplify that with some big textile factories, where the number of jobs may be up to 1,000.

The heuristic methods have become very popular and have gained much success in solving job shop scheduling problems. In the last twenty years a huge quantity of papers has been published, presenting several metaheuristic methods. From Simulated Annealing [6] to Particle Swarm Optimization [7], there are several variants of the same method class. Very popular between the researchers are the Evolutionary Algorithms [3,8-15]. In 1996, Vaessens et al. [16] stated a goal for the Job Shop Problem: to achieve an average error of less than two percent within 1,000 seconds total computation time. In this work the authors presented the Genetic Algorithms as the less effective metaheuristic to solve the JSSP. A possibility to increase the efficiency and the effectiveness is an algorithm that includes specific knowledge of the problem. Several works include some specific local search for the JSSP that is based on the critical path on a disjunctive graph to model the JSSP. For a long period, the Nowicki and Smutnicki's tabu search method [17] was seen as the most effective and efficient method for JSSP. In 2005, the authors presented a new version of a tabu search for the JSSP [18].

The JSSP represents several real production planning situations and for that reason it is a very important problem. The JSSP is an important practical problem in the fields of production management and manufacturing engineering. The applications of JSSP can be found in production planning, project resource management, distributed or parallel computing, and many other related fields. According to Lin et al. [19], a large number of small to medium companies still operate as job shops.

Throughout the years in the vast bibliography of the JSSP, variations to the classic model have been presented, allowing the study of specific real cases. For instance, Kimbrel and Sviridrenko [20] present the particular high-multiplicity JSSP that arises in the integrated circuit fabrication. As Yang et al. [21] say, the JSSP is a hard combinatorial optimization problem and computationally challenging. As they

pointed out, “efficient methods for arranging production and scheduling are very important for increasing production efficiency, reducing cost and improving product quality”.

In the JSSP each job is formed by a set of operations that has to be processed in a set of machines. Each job has a technological definition that determines a specific order to process the job’s operations, and it is also necessary to guarantee that there is no overlap, in time, in the processing of operations in the same machine; and also that there is no overlap, in time, in the processing of operations of the same job. The objective of the JSSP is to conclude the processing of all jobs as soon as possible, that is, to minimize the *makespan*.

The classical JSSP model considers a set of n jobs, and a set of m machines. Each job consists of a set of m operations (one operation on each machine), among which precedence relations exist and that defines a single order of processing. Each operation is processed in one machine only, during p time units. The instant when the job is concluded is C . All operations are concluded at C_{max} (called *makespan*).

For the convenience of the representation, operations are numbered consecutively from 1 to $N=n.m$, in which N is the total number of operations. The classic model considers that all the jobs are processed once in every machine, and the total number of operations is $n.m$. In a more general model, a job can have a number of operations different from m (number of machines). The case in which a job is processed more than once in the same machine, is called a job shop model with recirculation [14]. Scheduling problems occur wherever a number of tasks have to be performed with limited resources.

The computational and practical significance of the JSSP have motivated the attention of researchers for the last several decades. Yang et al. [21] enumerate the existing approaches for the JSSP that include exact methods such as branch-and-bound and dynamic programming, approximate and heuristic methods such as dispatching priority rules, shifting bottleneck approach, and Lagrangian relaxation. The authors associate the development of artificial intelligence techniques with the rise in many metaheuristic methods that have been applied to the JSSP, such as simulated annealing, tabu search, genetic algorithm, ant colony optimization, particle swarm optimization, and artificial immune system.

Over the years a lot of research has been made into this problem, particularly with genetic algorithms. An important issue in the genetic algorithms is the efficiency. Oliveira et al. [22] presents the inclusion of a “new” initial population that a generation procedure takes into account at the instance of the problem. The aim of this procedure is to improve the efficiency and the effectiveness of the genetic algorithm. The proposed genetic algorithm is based on random keys and the authors point out the easiness to model complex systems with this type of representation. Attending this advantage, this type of genetic algorithm is chosen to represent a job shop with recirculation to model the load of a truck in an automatic warehouse.

The JSSP is modeled mathematically. Blazewicz et al. [23] and Jain and Meeran [24] refer to the development of different formulations for this problem. The first one arises in 1959. One of the most used was presented by Adams et al. in 1988 [25], and is based in the disjunctive graph. We address details of this formulation in [23,25].

Roy and Sussman [26] model the JSSP using a disjunctive graph. The set of nodes is formed by all N operations and by a start node s and an end node e . The set of arcs is formed by two subsets. The conjunctive arcs set C , and the disjunctive arcs set D . For each pair of operations that are processed in the same machine (belonging to a different job) there exists one disjunctive arc. This arc is a non-oriented arc. If an operation i is performed before operation j , the arc is oriented from i to j .

The sequencing based on the disjunctive graph consists (for all machines) in the definition of a processing order between all operations that are processed by that same machine. A schedule is valid if the resulting oriented graph is acyclic. The longest path length is also designated by a critical path of the acyclic graph and is equal to the value of *makespan*.

A lot of research has been focused on obtaining and improving solutions for the JSSP. For a review and comparison, we refer the reader to Cheng et al. [8], Vaessens et al. [16], Blazewicz et al. [23] and to Jain and Meeran [24].

The solutions (schedules) for the JSSP can be classified in 3 sets: semi-actives, actives and non-delayed, according Fig. 3. In relation to the optimal solution of the problem (minimization of C_{max}), it is known that it is an active schedule but not necessarily a non-delayed schedule. In this work, the solutions obtained from genetic algorithm belong to Non Delay set. This choice was made in order to obtain solutions in a narrower space. Despite the optimal solution could not belong to this set, the quality of the solutions is reasonable and the computation time is shorter.

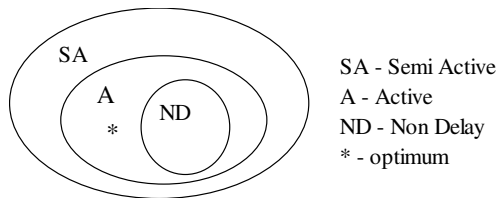


Fig. 3. Type of schedules

4 Methodology

In this work we adopted a method based on genetic algorithms. This technique's simplicity to model more complex problems and its easy integration with other optimization methods were factors that were considered for its choice. The algorithm proposed was conceived to solve the classical JSSP, but it is possible to use the same method to solve other variants of the JSSP such as the case of JSSP with recirculation.

One of the features that differentiate conventional genetic algorithms is the fact that the algorithm does not deal directly with the problem's solutions, but with a solution representation, the chromosome. The algorithm manipulations are done over the representation and not directly over the solution [27].

Traditionally, genetic algorithms used bit string chromosomes. These chromosomes consisted of only '0s' and '1s'. Modern genetic algorithms more often use problem-specific chromosomes with evidence that the use of real or integer value chromosomes often outperformed bit string chromosomes.

The permutation code was adequate to permutation problems. In this kind of representation, the chromosome is a literal of the operations sequence on the machines. For the classical JSSP case, Oliveira [3] presents a chromosome that is composed by m sub-chromosomes, one for each machine, each one composed by n genes, one for each operation. The i gene of the sub-chromosome corresponds to the operation processed in i place in the corresponding machine. The allele identifies the operation's index in the disjunctive graph.

Nevertheless, in this work, the random key code presented by Bean [28] is used. As Gonçalves et al. [13] state, the important feature of random keys is that all offspring formed by crossover are feasible solutions, when it is used as a constructive procedure based on the available operations to schedule and the priority is given by the random key allele. Another advantage of the random key representation is the possibility of using the conventional genetic operators. This characteristic allows the use of the genetic algorithm with other optimization problems, adapting only a few routines related with the problem. Equally easy becomes the hybridization with other heuristics when a genetic algorithm with random keys is used.

A chromosome represents a solution to the problem and is encoded as a vector of random keys (random numbers). In this work, according to Cheng et al. [8], the problem representation is indeed a mix from priority rule-based representation and random keys representation.

4.1 Constructive Algorithm

The solutions represented by chromosome are decoded by an algorithm, which is based on Giffler and Thompson's algorithm [29]. While the Giffler and Thompson's algorithm can generate all the active plans, the constructive algorithm only generates the plan in agreement with the chromosome. As advantages of this strategy, we pointed out minor dimension of solution space, and the fact that it does not produce impossible or disinteresting solutions from the optimization point of view. On the other hand, since the dimensions between the representation space and the solution space are very different, this option can represent a problem because two chromosomes can represent the same solution.

The constructive algorithm has N stages and in each stage an operation is scheduled. To assist the algorithm's presentation, consider the following notation existing in stage t :

- P_t - the partial schedule of the $(t-1)$ scheduled operations;
- S_t - the set of operations schedulable at stage t , i.e. all the operations that must precede those in S_t are in P_t ;
- σ_k - the earliest time that operation o_k in S_t could be started;

M^* - the selected machine where $\sigma^* = \min_{o_k \in S_k} \{\sigma_k\}$;

S_t^* - the conflict set formed by $o_j \in S_t$ processed in M^* and $\sigma_j = \sigma^*$.

o_j^* - the selected operation to be scheduled at stage t

The constructive algorithm of solutions is presented in a format similar to the one used by Cheng et al. [8] to present the Giffler and Thompson algorithm [29].

Algorithm 1. Constructive Algorithm

-
- Step 1* Let $t=1$ with P_1 being null. S_1 will be the set of all operations with no predecessors; in other words those that are first in their job.
- Step 2* Find $\sigma^* = \min_{o_k \in S_t} \{\sigma_k\}$ and identify M^* . If there is a choice for M^* , choose arbitrarily. Form S_t^* .
- Step 3* Select operation o_j^* in S_t^* with greatest allele value.
- Step 4* Move to next stage by
- (1) adding o_j^* to P_t , so creating P_{t+1} ;
 - (2) deleting o_j^* from S_t and creating S_{t+1} by adding to S_t the operation that directly follows o_j^* in its job (unless o_j completes its job);
 - (3) incrementing t by 1.
- Step 5* If there are any operations left unscheduled ($t < N$), go to *Step 2*. Otherwise, stop.
-

In Step 3 instead using a priority dispatching rule, the information given by the chromosome is used. If the maximum allele value is equal for two or more operations, one is chosen randomly.

4.2 The Genetic Algorithm Structure

The genetic algorithm has a very simple structure and can be represented in the Algorithm 2. It begins with population generation and her evaluation. Attending to the fitness of the chromosomes the individuals are selected to be parents. The crossover is applied and it generates a new temporary population that is also evaluated. Comparing the fitness of the new elements and of their progenitors the former population is updated.

The constructive algorithm of solutions is presented in a format similar to the one used by Cheng et al. [8] to present the Giffler and Thompson algorithm [29].

Algorithm 2. Genetic Algorithm

```

Step 1  begin
Step 2   $P \leftarrow \text{GenerateInitialPopulation}()$ 
Step 3  Evaluate( $P$ )
Step 4  while termination conditions not meet do
    (1)  $P' \leftarrow \text{Recombine}(P)$  //UX
    (2)  $P'' \leftarrow \text{Mutate}(P')$ 
    (3) Evaluate( $P''$ )
    (4)  $P \leftarrow \text{Select}(P \cup P'')$ 
Step 5  end while

```

The Uniform Crossover (UX) is used in this work. This genetic operator uses a new sequence of random numbers and swaps both progenitors' alleles if the random key is greater than a prefixed value. Table 1 illustrates the UX's application on two parents (prnt1, prnt2), and swaps alleles if the random key is greater or equal than 0.75. The genes 3 and 4 are changed and it originates two descendants (dscndt1, dscndt2). Descendant 1 is similar to parent 1, because it has about 75% of genes of this parent.

Table 1. The UX crossover

i	1	2	3	4	5	6	7	8	9	10
prnt1	0.89	0.48	0.24	0.03	0.41	0.11	0.24	0.12	0.33	0.30
prnt2	0.83	0.41	0.40	0.04	0.29	0.35	0.38	0.01	0.42	0.32
randkey	0.64	0.72	0.75	0.83	0.26	0.56	0.28	0.31	0.09	0.11
dscndt1	0.89	0.48	0.40	0.04	0.41	0.11	0.24	0.12	0.33	0.30
dscndt2	0.83	0.41	0.24	0.03	0.29	0.35	0.38	0.01	0.42	0.32

5 Computational Experiments

Computational experiments were carried out with some representative instances of the real problem. The representative instances of the real problem were generated randomly and they concern a job-shop problem with recirculation. The dimension of these instances corresponds to the defined maximum dimension of the real problem. Instances jr13_1, jr13_2 are constituted by 13 jobs (docking bay number) and 35 operations per job (one load's dimension), which adds up to a total of 455 operations and 11 machines (number of aisles).

The 35 operations of the major part of the jobs are processed in 5 machines. This situation corresponds to the real problem. Usually, the load of a truck (35 pallets) comes from the 5 aisles closer to the docking bay. The recirculation situation happens when a job is processed more than once in the same machine. In the real problem this corresponds to collecting several pallets from the same aisle. The 35 operations of one job are distributed randomly by a group of machines according to the percentages defined in Table 2.

Table 2. Distribution of the operations (%)

Jobs	Machines											
	1	2	3	4	5	6	7	8	9	10	11	
1	50	30	20									
2	25	25	25	25								
3	25	25	25	25								
4		25	25	25	25							
5		20	20	20	20	20						
6			20	20	20	20	20					
7				20	20	20	20	20				
8					20	20	20	20	20			
9						20	20	20	20	20		
10							25	25	25	25		
11								25	25	25	25	
12								25	25	25	25	
13									20	30	50	

Table 3 presents the results of computational experiences.

Table 3. Experimental results

Name	Pop	J	Op	Best	Aver.
jr_1_1	20	1	35	83	83
	100	1	35	83	83
jr_3_1	20	3	105	90	90
	100	3	105	90	90
jr_4_1	20	4	140	93	93
	100	4	140	93	93
jr_4_2	20	4	140	94	94,73
	100	4	140	94	94,4
jr_6_1	20	6	210	110	111,6
	100	6	210	109	110,3
jr_6_2	20	6	210	98	100,7
	100	6	210	99	100,1
jr_8_1	20	8	280	148	148,5
	100	8	280	147	148,1
jr_8_2	20	8	280	141	141,8
	100	8	280	141	141
jr_10_1	20	10	350	138	140,4
	100	10	350	137	139,5
jr_10_2	20	10	350	149	150,2
	100	10	350	149	150,4
jr_13_1	20	13	455	197	201,1
	100	13	455	197	199,9
jr_13_2	20	13	455	166	170,4
	100	13	455	166	168,8

The algorithm was implemented in C++ and the code was compiled using GNU Compiler Collection (GCC) version 4.6.1. The tests were run on a computer Intel i3, with 4 MB of RAM, on the Ubuntu 11.10 in Windows7 using VirtualBox.

The experiments were performed using two different populations - a small one with 20 individuals, and a larger one with 100 individuals. The third column of Table 3 indicates the number of jobs, and column Op. indicates the number of operations. This table presents the best value obtained from 15 runs of each configuration, and the average value of 15 runs. For all experiments 500 iterations were performed.

Table 4 shows the best fitness obtained in the 15 runs. A value in bold indicates that this value is the best fitness obtained for the instance.

Table 4. Best fitness

Name	Pop	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
jr_1_1	20	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83
	100	83	83	83	83	83	83	83	83	83	83	83	83	83	83	83
jr_3_1	20	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
	100	90	90	90	90	90	90	90	90	90	90	90	90	90	90	90
jr_4_1	20	93	93	93	93	93	93	93	93	93	93	93	93	93	93	93
	100	93	93	93	93	93	93	93	93	93	93	93	93	93	93	93
jr_4_2	20	94	95	94	94	95	94	95	96	96	95	95	94	95	94	95
	100	95	95	95	94	94	94	95	94	94	95	94	94	94	94	95
jr_6_1	20	110	113	112	112	111	111	111	110	113	112	112	111	111	111	113
	100	111	110	111	110	109	110	112	111	110	111	110	109	109	112	110
jr_6_2	20	98	102	102	100	101	102	101	98	102	102	100	101	100	101	101
	100	100	101	100	100	101	100	100	100	101	100	100	101	99	99	99
jr_8_1	20	148	149	148	150	149	148	148	148	149	148	148	148	149	149	148
	100	148	148	147	148	148	148	148	148	148	149	148	148	148	149	149
jr_8_2	20	142	142	141	142	141	141	143	142	142	142	142	143	141	141	142
	100	141	141	141	141	141	141	141	141	141	141	141	141	141	141	141
jr_10_1	20	141	140	141	140	139	142	141	141	140	141	140	139	142	138	141
	100	140	140	137	140	139	141	140	140	140	137	140	139	141	140	139
jr_10_2	20	149	149	150	150	150	151	151	149	149	150	150	150	151	152	152
	100	151	151	149	150	151	150	150	151	151	149	150	151	150	151	151
jr_13_1	20	202	201	204	198	200	203	202	202	201	201	204	197	198	201	202
	100	198	201	201	201	198	199	201	200	200	201	200	202	197	201	199
jr_13_2	20	173	171	170	170	173	171	169	172	170	172	172	169	169	169	166
	100	169	167	170	168	169	166	167	171	170	168	169	170	168	169	171

The results are promising, while achieving better solutions in early stages of the optimization process. Also the algorithm proved to be robust and consistent, with similar performances in all experiments considered.

With larger populations it is possible to achieve better results since the beginning of the optimization process, although requiring extra CPU time. Despite the code is not yet optimized for calculations speed, a run for the largest instance (jr_13_2), performing 500 iterations takes about 85 seconds with a population size of 100, and about 18 seconds with a population size of 20.

Fig. 4 shows these two runs - population size 20 (thin line) and population size 100 (thick line). Fig. 4 also illustrates the advantage of using a larger population (better solutions since the beginning), although this option takes more CPU time. With a population five times bigger the algorithm returns a better result within just 20% of the number of iterations.

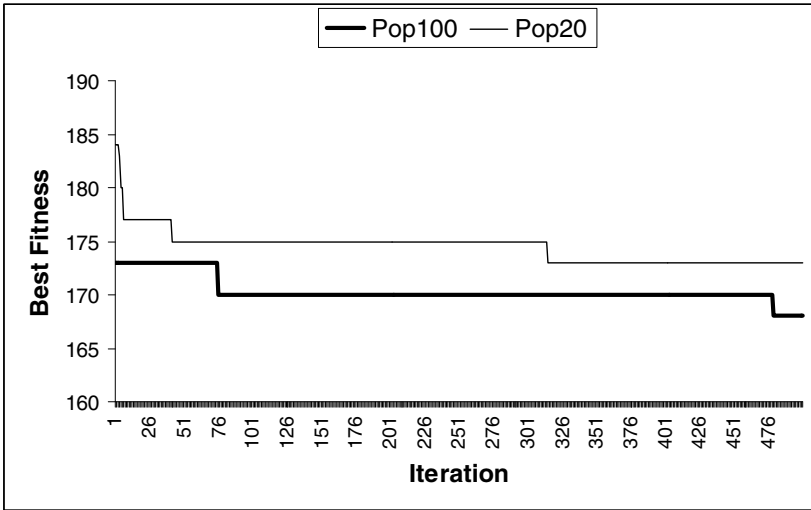


Fig. 4. Population effect

6 Conclusions

This paper presents a model for scheduling load operations in an automatic warehouse and a genetic algorithm to solve the JSSP with recirculation. The schedules are built using information given by the genetic algorithm to sequence the operations.

The algorithm incorporates a constructive algorithm to build the solution from the chromosome, and it always generates feasible plans, and in particular non delay schedules. This constructive algorithm allows the inclusion of specific knowledge of the problem and makes the algorithm very efficient. This algorithm allows, with little effort, the resolution of several daily problems that may occur in the warehouse.

The algorithm was tested with success on instances of equal dimension of the real problem. The computation time suggests that it is an efficient algorithm, allowing its integration in a decision support system to evaluate different alternatives in useful time. The main advantage in generating non delay plans is the fact that all schedules are valid in terms of programming the S/R machines, without generating deadlocks.

Further work consists to implement a constructive algorithm to generate active plans. Since a small population produce similar results than a larger population, it is possible to expect better results for some instances using active plans instead of non delay plans, without increasing the computational time.

Since the random keys representation has no Lemarkin property it is our intention to develop a variation of random keys representation to avoid this weakness.

Acknowledgments. This work was funded by the “Programa Operacional Fatores de Competitividade – COMPETE” and by the FCT – Fundação para a Ciência e Tecnologia in the scope of the project: FCOMP-01-0124-FEDER-022674.

References

1. Hausman, W.H., Schwarz, L.B., Graves, S.C.: Optimal storage assignment in automatic warehousing systems. *Management Science* 22, 629–638 (1976)
2. Oliveira, J.A.: *Aplicação de Modelos e Algoritmos de Investigação Operacional ao Planeamento de Operações em Armazéns*. Ph.D. Thesis, Universidade do Minho, Braga, Portugal (2001)
3. Oliveira, J.A.: Scheduling the truckload operations in automatic warehouses. *European Journal of Operational Research* 179, 723–735 (2007)
4. Zhang, C., Li, P., Guan, Z., Rao, Y.: A tabu search algorithm with a new neighborhood structure for the job shop scheduling problem. *Computers & Operations Research* 53, 313–320 (2007)
5. Liu, M., Hao, J., Wu, C.: A prediction based iterative decomposition algorithm for scheduling large-scale job shops. *Mathematical and Computer Modelling* 47, 411–421 (2008)
6. Laarhoven, P.J., Aarts, E., Lenstra, J.: Job shop scheduling by simulated annealing. *Operations Research* 40, 113–125 (1992)
7. Lian, Z., Gu, X., Jiao, B.: A similar particle swarm optimization algorithm for job-shop scheduling to minimize makespan. *Applied Mathematics and Computation* 183, 1008–1017 (2006)
8. Cheng, R., Gen, M., Tsujimura, Y.: A tutorial survey of job-shop scheduling problems using genetic algorithms - I. representation. *Computers & Industrial Engineering* 30, 983–997 (1996)
9. Davis, L.: Job-shop scheduling with genetic algorithm. In: *1st International Conference on Genetic Algorithms and their Applications*, pp. 136–140. Lawrence Erlbaum, Pittsburgh (1985)
10. Della Croce, F., Tadei, R., Volta, G.: A genetic algorithm for the job shop problem. *Computers & Operations Research* 22, 15–24 (1995)
11. Fang, H.L., Ross, P., Corne, D.: A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems. In: *5th International Conference on Genetic Algorithms*, pp. 375–382. M. Kaufmann Publishers (1993)
12. Gao, J., Sun, L., Gen, M.: A hybrid genetic and variable neighborhood descent algorithm for flexible job shop scheduling problems. *Computers & Operations Research* 35, 2892–2907 (2008)
13. Gonçalves, J.F., Mendes, J.J., Resende, M.G.C.: A hybrid genetic algorithm for the job shop scheduling problem. *European Journal of Operational Research* 167, 77–95 (2005)
14. Oliveira, J.A.: A genetic algorithm with a quasi-local search for the job shop problem with recirculation. In: *Applied Soft Computing Technologies: The Challenge of Complexity*, pp. 221–234. Springer, Heidelberg (2006)
15. Park, B.J., Choi, H.R., Kim, H.S.: A hybrid genetic algorithm for the job shop scheduling problems. *Computers & Industrial Engineering* 45, 597–613 (2003)
16. Vaessens, R., Aarts, E., Lenstra, J.K.: Job Shop Scheduling by local search. *INFORMS Journal on Computing* 8, 302–317 (1996)
17. Nowicki, E., Smutnicki, C.: A fast taboo search algorithm for the job-shop problem. *Management Science* 42, 797–813 (1996)
18. Nowicki, E., Smutnicki, C.: An advanced tabu search algorithm for the job shop problem. *Journal of Scheduling* 8, 145–159 (2005)

19. Lin, T.L., Horng, S.J., Kao, T.W., Chen, Y.H., Run, R.S., Chen, R.J., Lai, J.L., Kuo, I.H.: An efficient job-shop scheduling algorithm based on particle swarm optimization. *Expert Systems with Applications* 37, 2629–2636 (2009)
20. Kimbrel, T., Sviridenko, M.: High-multiplicity cyclic job shop scheduling. *Operations Research Letters* 36, 574–578 (2008)
21. Yang, J., Sun, L., Lee, H.P., Qian, Y., Liang, Y.: Clonal Selection Based Memetic Algorithm for Job Shop Scheduling Problems. *Journal of Bionic Engineering* 5, 111–119 (2008)
22. Oliveira, J.A., Dias, L., Pereira, G.: Solving the Job Shop Problem with a random keys genetic algorithm with instance parameters. In: *Proceedings of 2nd International Conference on Engineering Optimization (EngOpt 2010)*, CDRom, Lisbon, Portugal (2010)
23. Blazewicz, J., Domschke, W., Pesch, E.: The job shop scheduling problem: Conventional and new solution techniques. *European Journal of Operational Research* 93, 1–33 (1996)
24. Jain, A.S., Meeran, S.: A state-of-the-art review of job-shop scheduling techniques. *European Journal of Operations Research* 113, 390–434 (1999)
25. Adams, J., Balas, E., Zawack, D.: The shifting bottleneck procedure for Job Shop Scheduling. *Management Science* 34, 391–401 (1988)
26. Roy, B., Sussmann, B.: Les problemes d'ordonnancement avec contraintes disjonctives. *Note DS, No. 9 Bis, SEMA, Paris* (1964)
27. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading (1989)
28. Bean, J.C.: Genetics and random keys for sequencing and optimization. *ORSA Journal on Computing* 6, 154–160 (1994)
29. Giffler, B., Thompson, G.L.: Algorithms for solving production scheduling problems. *Operations Research* 8, 487–503 (1960)

On Solving the Profit Maximization of Small Cogeneration Systems

Ana C.M. Ferreira¹, Ana Maria A.C. Rocha², Senhorinha F.C.F. Teixeira¹,
Manuel L. Nunes¹, and Luís B. Martins³

¹ CITEPE R&D Centre

{acferreira,st,lnunes}@dps.uminho.pt

² Algoritmi R&D Centre

arocha@dps.uminho.pt

³ Mechanical & Materials Technologies R&D Centre

lmartins@dem.uminho.pt

University of Minho, Portugal

Abstract. Cogeneration is a high-efficiency technology that has been adapted to small and micro scale applications. In this work, the development and test of a numerical optimization model is carried out in order to implement an analysis that will lead to the optimal design of a small cogeneration system. The main idea is the integration of technical and economic aspects in the design of decentralized energy production considering the requirements for energy consumption for the building sector. The nonlinear optimization model was solved in MatLab® environment using two local optimization methods: the Box and the SQP method. The optimal solution provided a positive annual worth and disclosed reasonable values for the decision variables of the thermo-economic model. Both methods converged for the same solution, demonstrating the validity of the implemented approach. This study confirmed that the use of numerical optimization models is of utmost importance in the assessment of energy systems sustainability.

Keywords: cogeneration model, thermoeconomics, numerical optimization.

1 Introduction

The growth of energy consumption, mainly in the building sector, where power and thermal energy are needed, has increased the use of the cogeneration systems. In addition, distributed generation in small and micro cogeneration systems is of increasing interest in the energy market [15,16]. Combined heat and power (CHP) systems have proved to be one of the major options to achieve primary energy savings, minimize grid network investment costs and reduce the distribution losses. The environmental advantages with the reduction of pollutant gas emission are also significant. The cogeneration systems have a great potential since the simultaneous production of electrical and thermal energies leads to overall system efficiency up to 85-90% [1].

In Europe, the Energy Performance of Buildings Directive (EPBD) [5] and its recent recast [7] opened new opportunities for small-scale systems applied to the buildings sector. The directive obliges that, at the building design stage, the economic feasibility of high-efficiency alternative systems such as cogeneration is taken into account. Furthermore, the Cogeneration Directive 2004/8/EC [6], which came into force in 2006, largely promoted the energy efficiency and the improvement of energy supply security. Its purpose is, basically, the creation of a framework for the development of high efficiency cogeneration, based on the useful heat demand of the customer and the Primary Energy Savings (PES). A few EU member States, like Portugal, created subsidized grid-selling tariffs, called Feed-In Tariffs (FIT). These tariffs typically represent a premium comparatively to the buying-back prices, and so, all the produced electricity can be sold to the grid and only a match for the thermal energy is needed [13].

The decision on a particular technology, its design and optimization in order to fulfill the energy demands involves a comprehensive plant study where the technical, social and economic aspects must be included.

Optimization is an important tool in the process of designing, implementing and testing algorithms for solving a large variety of real problems. The selection of the best algorithm to implement is deeply related to the objective function, the number of variables and the constraints that give significance to the physical problem and the smoothness of the functions (differentiable or not differentiable functions) [17]. In recent years, different methods to optimize CHP plants have been proposed. All the power plants modeling techniques require the definition of an objective function and the constraints. The objective function is usually formulated in terms of cost of purchased energy and revenues from power sales and the constraints may represent energy balances, physical operating characteristics, simple upper and lower bounds.

Some authors apply classical optimization techniques such Linear Programming (LP) that can be complemented with mixed integer programming for discrete optimization [10]. Rodriguez-Toral in [18] developed an equation-oriented mathematical model for the optimization of heat and power systems using the sequential quadratic programming method and tested three optimization problems for CHP systems. Different approaches like the metaheuristics based or population methods are also applied. The main population-based metaheuristics include: the Genetic Algorithms (GA) and the Evolutionary Algorithms (EA) [19]. Most of computational optimization methods have focused on solving a single-optimization objective function. However, and due to the complexity of the problems, some authors have proposed multi-objective algorithms: the algorithms that combine the relative weights of all objectives in a single mathematical function and the Pareto-based optimization methods [2].

The main objective of this study is the application of two different numerical optimization methods, the Box Method [4] and the Sequential Quadratic Programming (SQP) [14], in order to model a small cogeneration system based on a micro gas turbine (regenerative Joule-Brayton thermodynamic cycle). Using Natural Gas (NG) as fuel, the system layout includes an internal air pre-heater

to increase the thermal efficiency and an external water heater that recuperates the enthalpy of the exhaust gases.

The mathematical model of the energy system is a complex nonlinear objective function with nonlinear constrains. Most of those constraints account for the physical and thermodynamic limitations of system operation. The objective function was defined as the maximization of Annual Worth of the plant operation and six decision variables were taken into account. This non-linear optimization model was implemented in MatLab® environment.

The organization of the paper is as follows. In Section 2, the physical description of the cogeneration system is presented. Section 3 describes the mathematical formulation developed to model the energy system. The optimization methods that are used to solve the constrained optimization problem are briefly introduced in Section 4. Section 5 shows the numerical results of this preliminary study and Section 6 presents the conclusions and ideas for future work.

2 Physical Description of the Cogeneration System

This study intends to model a small-scale cogeneration system able to deliver 125.5 kW of thermal power to fulfill the base heating load of a medium-size building. The central component of the system is the micro gas turbine that operates under a thermodynamic cycle known as the Joule-Brayton.

A physical description of this cogeneration system follows. The filtered air passes through the Compressor (C) and then through an Internal air Pre-Heater (IPH) before entering the combustion chamber (CC). Natural Gas is fed into the CC where it burns and the high temperature gases are then expanded in the turbine (T). The exhaust gases are firstly used to pre-heat the incoming air in the IPH and secondly for the production of hot water in the external water heat exchanger (WHE), before leaving to the atmosphere. The system components are presented in Fig. 1.

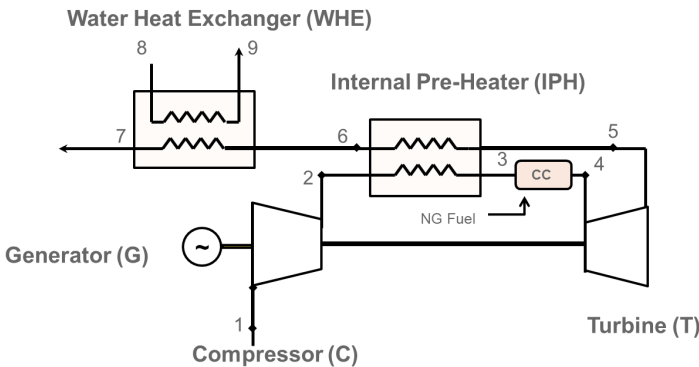


Fig. 1. Micro turbine-based CHP system

In the compressor, the air at entrance is assumed to be at the atmospheric pressure (1.013bar) at 293 K of temperature. All the components were considered adiabatic and a degree of irreversibility was assumed for both the compressor and turbine. The compressed air is heated at the IPH in order to increase the system efficiency. The second heat exchanger, the WHE, transfers the remaining energy from the exhaust gases, where a mass flow of 0.46 kg/s of water is heated from 288 to 353 K. The air and exhaust gases are treated as perfect gases considering constant specific heats (1.004 kJ/kg.K and 1.17 kJ/kg.K for the air and combustion gases, respectively).

3 Mathematical Model

For the mathematical model development of cogeneration systems, a set of standard thermodynamic relationships describing the temperatures and pressures are necessary for each system component, as fully presented in [12]. However, some constraints in terms of temperature should be considered and must be included in the mathematical model formulation.

As stated before, this optimization study takes into account technical and economic aspects. The integration of these two factors is achieved by defining cost equations for each component as a function of the physical variables. The thermodynamic model which computes the temperatures, pressures, flow rates and/or energy flows is solved and the purchase cost equation for each component of CHP system C_i ($i = C, CC, T, IPH, WHE$) can be calculated in order to evaluate the overall cost of the system for a specific range of combined power production.

The cost equations were determined based on data from micro-turbines available in the market in order to better approximate the costs of each plant component based on actual data of CapstoneTM65. The development of the cost equations, the cost coefficients of each system component as well as the dimensionless constants assumed in the model are described in detail by [12].

Along with the investment costs, a complete model for the process optimization has been developed as a nonlinear constrained optimization problem. The objective function of the problem is defined as the maximization of the Annual Worth (AW) of the small-scale CHP system (see (I)), making the balance between the incomes and the costs from CHP system operation

$$\max AW = R_{\text{sell}} + C_{\text{avoided}} - C_{\text{inv}} - C_{\text{op}} \quad (1)$$

where the income covers the revenue from selling electricity to the grid (R_{sell}) and the avoided cost of heat generation by a conventional boiler (C_{avoided}). The costs include the annual system investment cost (C_{inv}) and the operational costs involved in the production of electricity and heat using the CHP system (C_{op}).

The annual income from selling electricity to the grid (R_{sell}) is computed by the cumulative amount of electricity delivered to the grid (E_{sell}), considering the time of system operability, and the electricity-selling price (P_{sell}), as given in (2)

$$R_{\text{sell}} = E_{\text{sell}} P_{\text{sell}}. \quad (2)$$

The electricity-selling price was taken as a guaranteed and fixed (FIT) of 0.12€/kWh, in the case presented in this study. The term C_{avoided} described in (3), represents the avoided cost of NG that would be consumed by a conventional system (a boiler) to produce the same amount of useful thermal energy (H_{CHP})

$$C_{\text{avoided}} = P_{\text{fuel}} \frac{H_{\text{CHP}}}{\eta_b} \tag{3}$$

where, P_{fuel} ($P_{\text{fuel}} = 10\text{€}/\text{GJ}$) is the fuel price and η_b is the efficiency of the conventional boiler.

The system investment cost (C_{inv}) is calculated according to the annualized capital cost. Annualizing the initial investment cost corresponds to the spreading of the initial cost across the lifetime of a system, while accounting for the time value of the money. The initial capital cost is annualized as if it were being paid off a loan at a particular interest of discount rate over the lifetime of the option. The Capital Recovery Factor (CRF) and can be expressed as in (4)

$$\text{CRF} = (P \rightarrow A, i_e, n) = \frac{i_e(1 + i_e)^n}{(1 + i_e)^n - 1} \tag{4}$$

where A is the annuity (a series of equal amount cash transactions); P is the present value of the initial cost; i_e is the effective rate of return, and n is the number of years of the lifetime operation. In the present work, the system lifetime was considered equal to 10 years. In thermoeconomic optimization studies, i_e is calculated as: nominal rate of return (interest rate) minus inflation rate plus owners risk factor and correction for the method of compounding [9]. The effective rate of return i_e , herein considered was 7%, in a CRF of 0.142. Thus, the annual system investment cost, C_{inv} , becomes as in (5)

$$C_{\text{inv}} = \sum_i C_i * \text{CRF} \tag{5}$$

where C_i ($i = C, CC, T, IPH, WH$) is the purchase cost of each component of the CHP system, as mentioned above.

The total operational costs, C_{op} , results from the sum of maintenance and the fuel costs, as given in (6)

$$C_{\text{op}} = P_{\text{fuel}} \dot{m}_{\text{fuel}} \text{LHV}t + \phi C_{\text{inv}} \tag{6}$$

The fuel cost is calculated through the cumulative fuel consumption during the working period of the CHP system ($t = 4000\text{h}/\text{year}$) and considering the fuel price per energy unit, on Lower Heating Value (LHV) basis, and \dot{m}_{fuel} is the fuel mass flow rate. The maintenance costs were assumed as a percentage of the annual investment costs, equal to 15% of C_{inv} .

The decision variables were selected for the optimization based on their physical meaning and importance in these cogeneration systems. The chosen decision variables are: the compressor pressure ratio; the isentropic efficiency of the air

compressor; the isentropic efficiency of the gas turbine; the air temperature at the internal pre-heater; the temperature of the combustion gases at the turbine inlet and the electrical production. The latter decision variable takes into account the heat to power ratio (λ), which defines the relationship between the amount of useful heat (considered with a fixed value 125.5 kW) and the electricity produced by the CHP system.

The optimization problem was formulated as a nonlinear objective function with nonlinear inequality constraints. The definition of these constraints aims to restrict some problem variables according to their physical significance in the system operation.

For instance, the high-pressure air is pre-heated before entering in the CC and so it is required that the temperatures T_2 and T_3 are lower than the temperature of exhaust gases (T_5) at the turbine exit in order to allow an effective heat transfer in the IPH. This physical limitation can be satisfied by the following constraint

$$T_2 \leq T_3 \leq T_5.$$

These temperature values are calculated through the thermodynamic model of the all system. See [11] for more details.

4 Optimization Methods

Several approaches and methods can be used to solve constrained optimization problems. In the literature, there are gradient-based as well as derivative-free procedures that converge to local solutions of problems.

Recent developments show that derivative-free methods are highly demanded by researchers for solving optimization problems in various practical contexts. Derivative free optimization was developed for solving, in general, small dimensional problems (less than 100 variables) in which the computation of the derivatives are relatively expensive or even not available. Problems of this nature, usually arise in engineering applications. On the other hand, the gradient-based methods can only be applied to problems where the objective and the constraint functions are continuous and the derivatives exist.

In this study, two local methods, the Box method and the sequential quadratic programming method, are used in order to solve the present constrained optimization problem.

The Box method, firstly formulated by Box in 1965 [4], is a direct search method to solve constrained optimization problems. Its formulation allows to handle all kind of constrains with the exception of nonlinear equality constrains. The Box method, also known as Complex method when applied to constrained problems, is a direct search method without the need of analytic derivatives to solve constrained optimization problems. This method allows handling all kind of constrains with the exception of nonlinear equality constrains. The solution domain in the n -dimensional space is defined by a polyhedral figure with, at least, $(n + 1)$ vertices, where n represents the number of decision variables. The iterative procedure starts with only one feasible point, but needs a set of

points randomly generated to constitute the complex points. This means that the generated trial points satisfy the simple boundary constraints of the decision variables. Nevertheless, for each generated point, it is required to verify if it satisfies all the other constraints. Thus, the requirement of a feasible initial point, able to satisfy all the constraints of the problem is one disadvantage of the Box method.

Sequential Quadratic Programming (SQP) [14] is one of the most successful methods for the numerical solution of constrained nonlinear optimization problems. It relies on theoretical foundation and provides powerful algorithmic tools for the solution of optimization problems.

The idea of SQP is to model the constrained nonlinear problem at the current point x_k (k is the iteration number) by a quadratic subproblem (QP) and to use the solution of this subproblem to find the new point x_{k+1} . SQP is in a way the application of Newton's method to the Karush-Kuhn-Tucker (KKT) optimality conditions. The QP subproblems which have to be solved in each iteration step should reflect the local properties of the NLP with respect to the current iterate x_k . This is done in such a way that the sequence (x_k) converges to a local minimum x^* of the nonlinear problem as $k \rightarrow \infty$. In this sense, the nonlinear problem resembles the Newton and quasi-Newton methods for the numerical solution of nonlinear algebraic systems of equations [14].

Note that a major advantage of SQP is that the initial point (or any future iterations points) need not be feasible points (solutions that solve the constraints of the problem). However, the difficulty to choose correctly the initial point is an advantage, because the convergence is only guaranteed when the algorithm starts close to the solution point.

5 Numerical Results and Discussion

In this section, the numerical results of a preliminary study are presented and discussed. The optimization problem was solved by two nonlinear optimization methods, in MatLab® environment: the Box method (by implementation) and the Sequential Quadratic Programming (SQP) by using FMINCON (an available command in the optimization toolbox).

When implementing the optimization model, a script file was created to define all the equations that describe the thermodynamic behavior of the physical system and the cost equations which include the most important physical parameters for each of the CHP components, respectively. For both methods, it was also required to define the nonlinear constraints as well as the simple bounds of the decision variables (upper and lower bounds). The main routine, where the algorithm parameters were defined, integrates all the scripts in order to solve the thermoeconomic problem.

In the specific case of the two methods under study, the Box method and the SQP method need an initial point as input for the optimization algorithm. In this study the initial approximation for the six decision variables were given by

$$r_c = 4; \eta_c = \eta_T = 0.85; T_3 = 850\text{K}; T_4 = 1200\text{K} \text{ and } \dot{W} = 90\text{kW}.$$

In order to better compare the accuracy of both methods in finding the optimal solution, the convergence criteria of the SQP method was assumed taking into account the same stopping criteria of the Box method. Thus, the convergence was assumed setting the value of 1.0E-06 for the termination tolerance on the function value and on the constraint violation.

After executing the main routine, the obtained optimal values for the six decision variables with the Box and SQP methods are given in Table II.

Regarding the Box method results, and despite its simplicity, parameters such as the convergence criteria, the explicit constraint violation or complex reflection parameters had to be adjusted, according to the specifications of the optimization problem in study. In this study, it was considered the recommended value for the reflection parameter ($\alpha = 1.3$) and the Box method reached convergence within 640 iterations.

Regarding the SQP results, the objective function converged to the optimal solution within 22 iterations and performed 180 function evaluations. The optimization terminated because the objective function was non-decreasing in feasible directions and the solution was within the value of the function tolerance.

Table 1. Optimum values of the decision variables

Decision Variables	Box Method	SQP Method
r_c	5.786	5.785
η_c	0.8228	0.8228
η_T	0.8614	0.8614
T_3 (K)	953.64	953.63
T_4 (K)	1365.59	1365.6
\dot{W} (kW)	92.86	92.86

It can be concluded that both numerical optimization techniques yield the same feasible local minimum (satisfying the constraints and the upper and lower values of the decision variables). The bounds in the variables guarantee that the optimum solution is within the technical operating capability of the plant.

The data in Table II show that the two methods present the same outcome for the optimal solution. The results yield the conclusion that both methodologies are highly adequate because the results for the decision variable are perfectly acceptable for cogeneration systems in the considered power range under study.

The compressor pressure ratio ($r_c = 5.78$) is relatively higher than the currently available micro gas turbines in the market. However micro gas turbines with a single stage usually have a compression pressure ratio of about 4. The obtained Turbine Inlet Temperature (T_4) of 1365K can be considered a reasonable value, although this result is higher than the expected value (of approximately of 1200 K). The compressor and turbine efficiencies ($\eta_c = 0.8228$ and $\eta_T = 0.8614$) seem to be within the expected values for this kind of systems. According to these results for the optimal solution, the resulting CHP system is able to produce about 93 kW of electrical power.

The developed numerical optimization model was also able to find the optimal values of other important physical variables for the best economical outcome, e.g. exhaust gases temperature at turbine exit (T_5), gas exit temperature (T_7), air mass flow rate (\dot{m}_a) and fuel mass flow (\dot{m}_{fuel}). Some CHP system performance criteria were also assessed: the water heater and internal pre-heater effectiveness, electrical efficiency and the primary energy saving. These results are presented in the Table 2.

Table 2. Optimal values of relevant physical variables

Physical variables		Optimal Value
Air Mass Flow Rate	\dot{m}_a (kg/s)	0.4192
Fuel Mass Flow Rate	\dot{m}_{fuel} (kg/s)	0.0058
Gases Temperature at Turbine Exit	T_5 (K)	985.88
Exit Gas Temperature	T_7 (K)	363
Internal Pre-Heater Effectiveness	ε_{IPH} (%)	93.14
Water Heater Effectiveness	ε_{WHE} (%)	77.09
Electrical Efficiency	η_{el} (%)	33.1
Primary Energy Saving	(%)	14.17

The results shown in the Table 2 disclosed that the thermal energy transferred to the water is maximized, since the temperature of the exhaust gases reaches the minimum value allowed as its lower bound in the optimization model (a temperature of 363 K). The IPH effectiveness is one of the most relevant parameters in optimizing regenerative micro gas turbines. Considering the result, slightly above 90%, the model will lead to a CHP system with an excellent overall efficiency. The obtained electrical efficiency of 33.1% is higher than the actual values observed for real micro turbines (25 to 31%). The CHP system could also provide a PES of 14.2%, respecting the minimum recommended (at least 10%) to be considered a high efficiency cogeneration system for small-scale applications.

The optimal economic output, optimal costs and revenues, are presented in Table 3. Obviously, both methods reached the same value for the objective

Table 3. Results for the optimal costs and revenues of the small CHP system

Economic Output		(€/year)
Capital investment Cost	C_{inv}	-13926
Fuel Costs	C_{fuel}	-37565
Operating Costs	C_{op}	-2089
Income from selling electricity power to the grid	R_{sell}	41453
Avoided cost of conventional heat generation	C_{avoided}	20080
Annual worth of the small scale-CHP system	AW	7953

function ($AW = 7953$). According to the results of the optimization, it is possible to get profits with the cogeneration system operation obtaining a maximum annual worth of 7953€.

The authors remark, that several simulations considering distinct initial points were ran for both methods but the same output as the optimal solution was obtained. It was also carried out two additional tests in which the stopping criteria tolerances were changed, but it was verified that the optimization problem converged to the same optimal solution.

6 Conclusions

In this paper, a model of a small cogeneration system based on a micro gas turbine was developed. The small cogeneration system was able to deliver 125.5 kW of thermal power to fulfill the base heating load of a medium-size building.

In order to select the best fitness of the decision variables, two different optimization approaches, the Box method and the Sequential Quadratic Programming method were implemented, and some simulations were carried out to obtain the maximum profit of the small cogeneration system.

The obtained results for both methods were similar, demonstrating the validity of the implemented approach although they have different characteristics: Box method is a direct search method that only uses the information of the objective function; SQP is a method that uses the information from the derivatives of the function.

Also, both methods are computationally very simple and none of them requires large computer storage. A more detailed sensitivity analysis for the different parameters of the optimization methods should be carried out in order to better confirm their accuracy.

Since Box and SQP are local optimization methods, and we do not assume the convexity of the problem. Hence, there is no guarantee that the optimal solution found is, indeed, the global optimum of the problem. Thus, it would be appropriate, as future work, the use of a global optimization method in order to verify if the obtained solution is, actually, the global maximum of the optimization problem.

Acknowledgments. The first author would like to express her acknowledgments for the support given by the Portuguese Foundation for Science and Technology (FCT) through the PhD grant SFRH/BD/62287/2009. This work was financed by National Funds-Portuguese Foundation for Science and Technology, under Strategic Project and PEst-OE/EME/UI0252/2011.

Second author is gratefully acknowledged for the financial support from FEDER COMPETE (Operational Programme Thematic Factors of Competitiveness) and FCT Project FCOMP-01-0124-FEDER-022674.

References

1. Alanne, K., Saari, A.: Sustainable small-scale CHP technologies for buildings: the basis for multi-perspective decision-making. *Renew Sustain Energy Review* 8(5), 401–431 (2004)
2. Baños, R., Manzano-Agugliaro, F., Montoya, F.G., Gil, C., Alcayde, A., Gómez, J.: Optimization methods applied to renewable and sustainable energy: A review. *Renewable and Sustainable Energy Reviews* 15, 1753–1766 (2011)
3. Bejan, A., Tsatsaronis, G., Moran, M.: *Thermal design and optimization*. John Wiley and Sons Inc. (1996)
4. Box, M.J.: A new method of constrained optimization and a comparison with other method. *Computer Journal* 8(1), 42–52 (1965)
5. Directive 2002/91/EC. Directive of the European Parliament and of the Council. On the energy performance of buildings. *Official Journal of the European Union*, December 16 (2002)
6. Directive 2004/8/EC. Directive of the European Parliament and of the Council, On the promotion of cogeneration based on a useful heat demand. *Official Journal of the European Union*, February 11 (2004)
7. Directive 2010/31/EU. Directive of the European Parliament and of the Council, On the energy performance of buildings (recast). *Official Journal of the European Union*, May 19 (2010)
8. El-Sayed, Y.M.: A Decomposition strategy for the thermoeconomic optimization of a given system configuration. *Journal Energy Resource Technology* 111(3), 41–47 (1989)
9. Gogus, Y.A.: Thermoeconomic Optimization. *International Journal of Energy Research* 29, 559–580 (2005)
10. Lahdelma, R., Hakonen, H.: An efficient linear programming algorithm for combined heat and power production. *European Journal of Operational Research* 148, 141–151 (2003)
11. Leão, C.P., Teixeira, S.C.F.T., Silva, A.M., Nunes, M.L., Martins, L.A.S.B.: Thermoeconomical optimization in the design of small scale and residential cogeneration systems. In: *Mechanical Engineering Congress & Exposition, IMECE 2009, Florida, USA, 13089*, 5 pages (2009)
12. Marques, F.M.: *Análise termo-económica no desenvolvimento de sistemas de cogeração de pequena escala para edifícios*, Master Thesis in Mechanical Engineering, University of Minho, Portugal (2011) (in portuguese)
13. Martins, L.B., Ferreira A.C.M., Nunes, M.L., Leão, C.P., Teixeira, S.F.C.F., Marques, F., Teixeira, J.C.F.: Optimal Design of Micro-Turbine Cogeneration Systems for the Portuguese Buildings Sector. In: *Proceedings of the ASME 2011 International Mechanical Engineering Congress & Exposition, IMECE 2011–64470, (DVD), Denver, Colorado, USA, 8 pages* (2011)
14. Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer Science and Business Media LLC (2006)
15. Onovwiona, H., Ugursal, V.: Residential cogeneration systems: a review of the current technology. *Renewable and Sustainable Energy Reviews* 10, 389–431 (2006)
16. Pehnt, M.: Environmental impacts of distributed energy systems-The case of micro cogeneration. *Environmental Science & Policy* 11(1), 25–37 (2008)

17. Rao, S.S.: Engineering Optimization Theory and Practice, ch. I, 4th edn. John Wiley & Sons, Inc. (2009)
18. Rodriguez-Toral, M.A., Morton, W., Mitchell, D.R.: Using new packages for modelling, equation oriented simulation and optimization of a cogeneration plant. *Chemical Engineering* 24, 2667–2685 (2000)
19. Valdés, M., Duran, M.D., Rovira, A.: Thermoeconomic optimization of combined cycle gas turbine power plants using genetic algorithms. *Applied Thermal Engineering* 23, 2169–2182 (2003)

Global Optimization Simplex Bisection Revisited Based on Considerations by Reiner Horst*

Eligius M.T. Hendrix¹, Leocadio G. Casado², and Paula Amaral³

¹ Arquitectura de Computadores, Universidad de Málaga and Logistics
and Operations Research, Wageningen University

eligius@uma.es

² Dpt. de Arquitectura de Computadores y Electronica, Universidad de Almería
leo@ual.es

³ Department of Mathematics and CMA, Universidade Nova de Lisboa
paca@campus.fct.unl.pt

Abstract. In this paper, the use of non-optimality spheres in a simplicial branch and bound (B&B) algorithm is investigated. In this context, some considerations regarding the use of bisection on the longest edge in relation with ideas of Reiner Horst are reminded. Three arguments highlight the merits of bisection of simplicial subsets in B&B schemes.

Keywords: Global Optimization, simplicial partition, branch and bound, bisection.

1 Introduction

This work is dedicated to Reiner Horst, who encouraged and inspired the study of Global Optimization branch and bound (B&B) methods. In his last (2010) contribution titled “Bisection by global optimization revisited” [8], some considerations were elaborated regarding the use of simplices in branch and bound. Reiner ideas on simplicial partitioning, developed in discussion with his co-workers Micheal Nast and Nguyen Van Thoai, are summarized in the book [11] and elaborated and experimented in the thesis of Ulrich Raber [17]. The main issue in [8] is that “bisection is not optimal”. It is clear that optimality depends on the objective under consideration, and we would like to stress that Reiner had a wider view on the use of simplices than B&B only, namely the typical lower dimensional tessalation in physics and the use of triangulations for finding roots of mappings. This paper focuses on several aspects for which bisecting the longest edge in simplicial branch and bound in Global Optimization may be convenient.

* This paper has been supported by The Spanish Ministry of Science and Innovation (project TIN2008-01117) and Junta de Andalucía (P11-TIC-7176), in part financed by the European Regional Development Fund (ERDF) and by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through PEst-OE/MAT/UI0297/2011 (CMA). Eligius M.T. Hendrix is a fellow of the Spanish “Ramon y Cajal” contract program, co-financed by the European Social Fund.

Previous experience regarding the use of B&B on the unit simplex in applications in mixture design for multinational Unilever, [4], showed that the study of Reiner [7], on splitting the unit simplex by bisection, leads typically to edge lengths of 1 , $\sqrt{3}/2$, $\sqrt{2}/2$ and $1/2$. Moreover, it was shown that, implicitly, this technique leads to samples points over an equidistant grid when using an ε accuracy in the decision space [4]. In the above mentioned application, the practical importance of this design property has connections with robustness considerations, in the sense that finding an acceptable design means that all points in its environment are feasible. Bisection of the longest edge gives relatively ‘round’ partition sets. For running a B&B tree to the bottom, where simplices have at most a size of ε , this feature is convenient. Using radial splitting over the centroid, as suggested in [7], leads to needle shaped subsimplices. Deviating from the midpoint requires keeping track of ε robustness. Concluding *Bisecting over the middle of the longest edge can be convenient for ε robustness considerations.*

A second aspect has to do with implementation issues in B&B. Instead of questioning how many small subsets the B&B search may lead to at the bottom of the tree, Reiner showed his always optimistic perspective of following a subset to be split iteratively from top to bottom of the tree to see how fast it converges to a singleton. In [8] he repeats the proposition that can also be found in [7] and with an extensive proof in [11] that after splitting an n -simplex n times, the longest edge will be shorter than $\sqrt{3}/2$. He also reminded that Beaker Kearfott already published a similar result in 1978. In [4] it is shown that after splitting all edges i.e. going $n(n+1)/2$ deeper in the tree, then the size is at most $1/2$. Up to about $n = 9$ this is even a sharper bound. However, seen from the worst case (pessimistic) B&B perspective this is not encouraging; the number of simplices to be evaluated is astronomically high.

After obtaining in practice millions of subsets to be stored in RAM, there are two optional directions. From a theoretical point of view, this means looking for sharper (and more elegant) bounds. From a practical perspective, it implies designing a convenient way to store and manage the search tree, allowing sorting, easy look up, and workload distribution over several processors. This is a second reason why using the midpoint of the longest edge can be convenient; the same (evaluated) point appears several times as vertex of sub-simplices, allowing repeated re-use of its information without the need to evaluate it. Concluding, *bisecting over the middle of the longest edge can be convenient for storing a B&B tree structure with subsets linked to evaluated vertices.*

The use of simplices in B&B to solve GO problems is common [12,16]. The idea can be found in the work of Ulrich Raber [16], also in the work of Julius and Antanas Žilinskas [20] and Remigijus Paulavičius, J. Žilinskas and Andreas Grothey [14]. Specifically, in [19] the idea of using more elegant partitioning than bisection is discussed. However, these proposals are applicable only to low dimensional cases. Theoretically, it is known that convergence is guaranteed. As discussed, bisection may be a good basis from the computer science perspective, despite it is not efficient from a bounding perspective. In the sequel, a third aspect related to the practical use in B&B is outlined. To address this issue, a

question is posed: how can simplicial partition sets and bisection be used to have early pruning of nodes? This means to develop methods that detect subspaces which cannot contain a global optimizer in an early phase.

Regarding node pruning, we focus on covering methods, based on bounds on first (Lipschitz constant) and second derivative. For higher dimensions, greater than 1, Reiner Horst mentioned in [11] that Lipschitz optimization “does not look very practical”. His focus was rather on B&B; in [10] a B&B view on covering methods is presented. Our study deals with linking the two concepts using so-called non-optimality spheres.

We describe the idea of covering algorithms that can basically also be found in the books of Reiner Horst [9,11] with the aid of simple examples and figures in Sect. 2. In Sect. 3, the concept of non-optimality spheres is presented and a B&B algorithm is given in Sect. 4 based on simplicial partition sets. We discuss how to infer simplicial partition set covering by non-optimality spheres in Sect. 5. We numerically illustrate the concept of using bisection in such a procedure in Sect. 6. This is followed by conclusions in Sect. 7.

2 Covering Algorithms

The generic box-constrained GO problem consists in finding the global minimum f^* of a real valued n -dimensional function $f : S \rightarrow \mathbb{R}$, $S \subset \mathbb{R}^n$, and the corresponding set S^* of global minimum points, where S is a box, i.e.:

$$f^* = f(x^*) = \min_{x \in S} f(x), \quad x^* \in S^* . \quad (1)$$

Covering methods approach this problem by defining iteratively a covering function $\varphi_k(x) \leq f(x)$, where a minimum point of $\varphi_k(x) \leq f(x)$ over S is then used as the next iterate x_{k+1} . A basic method with this property is due to Piyavski and Shubert [5,15,18], who published in parallel about an algorithm where the so-called saw-tooth cover is based on information about the Lipschitz constant. The knowledge of a scalar L is assumed such that

$$|f(x_1) - f(x_2)| \leq L \|x_1 - x_2\| \quad \forall x_1, x_2 \in S . \quad (2)$$

For evaluated points $x_1, \dots, x_i, \dots, x_k$, with function values $f_1, \dots, f_i, \dots, f_k$ the covering function is defined by

$$\varphi_k(x) = \max_{1 \leq i \leq k} (f_i - L \|x - x_i\|) . \quad (3)$$

By keeping track of the best function value as upper bound of the global minimum $U = \min_{1 \leq i \leq k} f_i$, one can show that the algorithm evaluating iteratively the minimum of φ_k leads to a guaranteed approximation of the optimum with accuracy δ , when using $U - \min_x \varphi(x) < \delta$ as stopping criterion.

The real challenge is the application of this concept for dimensions higher than 1, given the scepticism of some authors, Reiner Horst included in [11] that a direct application looks impractical. A continuing report on achievements on

covering methods in Russian is due to the work of Yuri Evtushenko (e.g. [6]) during 40 years. A description of the corresponding algorithm and minimization of φ_k is also described in [13].

A second base for covering algorithms is due to the work of Breiman and Cutler, [2] when using a bound K on the second derivative, such that $K \geq -f''(x), x \in S$ or more general (in higher dimensions) on an overestimate of the negative of the minimum eigenvalue of the Hessian, such that

$$f(x) \geq f(x_1) + \nabla f^T(x)(x - x_1) - \frac{1}{2}K\|x - x_1\|^2 \quad \forall x, x_1 \in S . \quad (4)$$

Analogously to (3), the corresponding covering function is given by

$$\varphi_k(x) = \max_{1 \leq i \leq k} (f_i + \nabla f_i^T(x - x_i) - \frac{1}{2}K\|x - x_i\|^2) . \quad (5)$$

The algorithm of Breiman-Cutler takes iteratively a minimum point of φ_k as next iterate. The original article [2] describes also the approach for the multivariate case where it is necessary to find the minimum of intersecting parabolics leading to polytope shaped regions that are similar to Voronoi diagrams. The method is very elegant, but also very elaborative, as it requires storing information on all evaluated points, intersecting planes and resulting vertices of the polytopes.

Baritomba in [1] showed how (2) and (4) can be relaxed by focussing on the behavior around global optimum x^*, f^* . Let M and K be values such that $f(x) \leq f^* + M\|x - x^*\|, \forall x \in S$ and $f(x) \leq f^* + \frac{1}{2}K\|x - x^*\|^2, \forall x \in S$. So, it is

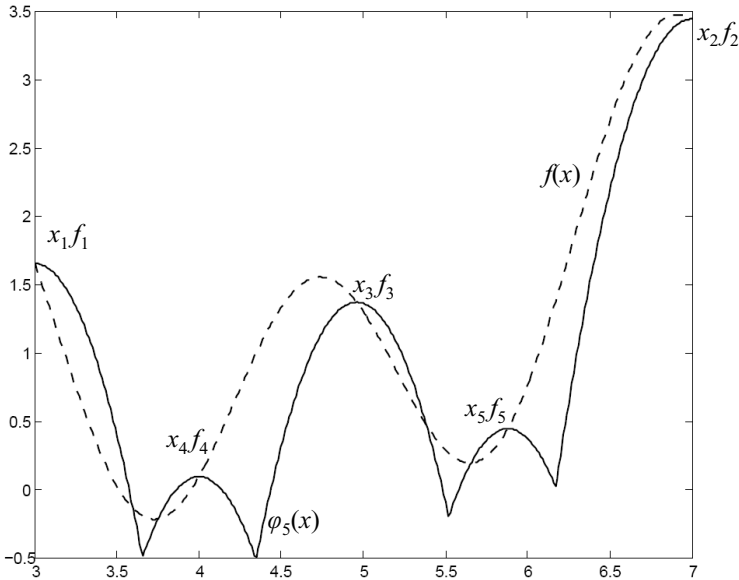


Fig. 1. Iterate is a minimum of (7) for $f(x) = \sin(x) + \sin(3x) + \ln(x)$

not necessary to have a global overestimate of neither Lipschitz constant, nor of the second derivative (or the negative of the minimum eigenvalue of the Hessian in higher dimensions), K . Then one can take as cover

$$\varphi_k(x) = \max_{1 \leq i \leq k} \{f_i - M\|x - x_i\|\} \tag{6}$$

or alternatively

$$\varphi_k(x) = \max_{1 \leq i \leq k} \left\{ f_i - \frac{1}{2}K\|x - x_i\|^2 \right\} . \tag{7}$$

An interesting aspect is that φ is not necessarily an underestimating function of f , but it neither cuts away a global minimum point.

Example 1. Consider function $f(x) = \sin(x) + \sin(3x) + \ln(x)$ on the interval $X = [3, 7]$. We take $K = 10$, as the maximum value of the second derivative is reached close to x^* . Fig. 1 depicts iterates corresponding to a minimum point of (7). Function φ_k is not a lower bounding function, but neither cuts away the global minimum. The minimum point of φ_k is a lower bound for the minimum of f .

The example illustrates the deterministic view of Reiner Horst, where the global minimum can be obtained with a guarantee given certain information. Usually one refers to a bound on derivatives, but the assumptions of Baritompá around the global minimum point do not require the function to be differentiable neither continuous over the whole domain. On the other hand, practically information on M or K is required, which may be as hard to obtain as solving the original problem.

Our interest is the multivariate variants of using (6) and (7) in simplicial branch and bound. Finding iteratively the minimum point of $\varphi_k(x)$ may be a tedious job. However, from a B&B perspective, it is not necessary to know exactly the minimum of φ . Our focus is on the potential of so-called non-optimality spheres, close to the covering concepts of [6] and the concept of infeasibility spheres in [4]. The purpose is to come to simplicial B&B based algorithms applying (6) and (7) in order to illustrate the usefulness of bisection.

3 Non-optimality Spheres

As the name suggests, non-optimality spheres are spheres that are guaranteed not to contain an optimal solution. We start describing how non-optimality spheres can be derived from sample points and global value information. Next, a specific B&B algorithm which uses simplicial partition sets is presented.

Consider sample points $x_1, \dots, x_i, \dots, x_k$, with $f_1, \dots, f_i, \dots, f_k$ as function values, and U the best function value found, $U = \min_{1 \leq i \leq k} f_i$. For a value of M such that

$$f(x) \leq f^* + M\|x - x^*\|, \forall x \in S, \forall x^* \in S^*, \tag{8}$$

a non-optimality sphere BM_i centered at x_i with radius r_i is given by

$$BM_i = \left\{ x \in S \mid \|x - x_i\| < \left(r_i = \frac{f_i - U}{M} \right) \right\} . \tag{9}$$

For a value K with

$$f(x) \leq f^* + \frac{1}{2}K\|x - x^*\|^2, \forall x \in S, \forall x^* \in S^* \tag{10}$$

a non-optimality sphere is given by

$$BK_i = \{x \in S \mid \|x - x_i\|^2 < (r_i^2 = 2\frac{f_i - U}{K})\} . \tag{11}$$

First notice that in (8) and (10) necessarily $M > 0$ and $K > 0$. The definition of the non-optimality sphere radius is obtained by a simple manipulation of (8) and (10), bounding f^* by U and replacing x and $f(x)$ respectively by x_i and f_i . Comparing BK_i as in (11) with the sphere that could be obtained using a similar procedure and the Breiman-Cutler assumption (4), the difference is that the center of the sphere is shifted towards $x_i + \frac{1}{K}\nabla f_i$ and in the radius definition in (11), one should take instead of f_i the value of the top of the parabola $f_i + \frac{1}{2K}\nabla f_i^T \nabla f_i$.

Notice that for a current best point where $f_i = U$, the non-optimality sphere is empty, i.e. it could be an optimum point. Moreover, if the upper bound U goes down during the iterations, the spheres are getting bigger. The fact that the area of the non-optimality spheres can be left out of further consideration is given in the following theorems.

Theorem 1. *Non-optimality sphere BM_i does not contain a global minimum point $x^* \in S^*$.*

Proof. Proof by contradiction. By definition of M , $f(x_i) \leq f^* + M\|x_i - x^*\|$, such that $f^* \geq f_i - M\|x_i - x^*\|$. Let $x^* \in B_i$. Substitution of definition (9) gives

$$f^* \geq f_i - L\|x_i - x^*\| > f_i - f_i + U = U, \tag{12}$$

which contradicts U being an upper bound of f^* . □

For the parabolic non-optimality sphere this is given as follows.

Theorem 2. *Non-optimality sphere BK_i does not contain a global minimum point $x^* \in S^*$.*

Proof. Proof by contradiction. By definition of K , $f(x_i) \leq f^* + \frac{1}{2}K\|x_i - x^*\|^2$, such that $f^* \geq f_i - \frac{1}{2}K\|x_i - x^*\|^2$. Let $x^* \in BK_i$. Substitution of definition (11) gives

$$f^* \geq f_i - \frac{1}{2}K\|x_i - x^*\|^2 > f_i - f_i + U = U, \tag{13}$$

which contradicts with U being an upper bound of f^* . □

In general, a non-optimality sphere may be completely covered by another one depending on the values of M and K . This is interesting from algorithmic perspective, as the covered sample point and its sphere apparently do not add any

information to the search. However, if M is a strict overestimate of the Lipschitz constant

$$M > \frac{|f_2 - f_1|}{\|x_2 - x_1\|} \forall x_1, x_2 \in S, \tag{14}$$

then one sphere cannot be covered by another one.

Theorem 3. *Let M be an overestimate (14), BM_i be defined by (9) and $x_1, x_2 \in S$ with function values f_1, f_2 . Then neither $BM_1 \subset BM_2$ nor $BM_2 \subset BM_1$.*

Proof. A sphere BM_2 of radius r_2 and center x_2 contains a sphere BM_1 with radius r_1 and center x_1 if

$$r_2 \geq r_1 + \|x_2 - x_1\|.$$

W.l.o.g. let $f_2 > f_1$, such that $r_2 > r_1$. Then $BM_2 \subset BM_1$ is not possible, as $r_1 \geq r_2 + \|x_2 - x_1\|$ is not possible. Furthermore, from (9) we have

$$r_2 - r_1 = \frac{f_2 - U}{M} - \frac{f_1 - U}{M} = \frac{f_2 - f_1}{M}.$$

Now using (14) we obtain

$$r_2 - r_1 < \frac{f_2 - f_1}{|f_2 - f_1|} \|x_2 - x_1\| = \|x_2 - x_1\|,$$

so

$$r_2 < r_1 + \|x_2 - x_1\|.$$

So neither $BM_2 \subset BM_1$ nor $BM_1 \subset BM_2$. □

Example 2. Consider the six-hump camel-back function:

$$f(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4 \tag{15}$$

taking as feasible area $S = [-2, 2] \times [-2, 2]$. It has 6 local optimum points two of which describe the set of global optimum solutions. All vertices of S and 16 more generated sample points x_i are evaluated. The maximum eigenvalue of the Hessian goes up to 184. In [2], experiments are done with $K = 9$, as the lower bounding is based on the most negative eigenvalue. In the illustration, a value of $K = 60$ is used.

The resulting Emmentaler set $S \setminus \cap BK_i$, where the optimum still can be located, is drawn in Fig. 2. Similar figures can be made using $M = 38$ as valid upper bound in the determination of BM_i . The spheres close to the vertices of S are relatively big, because the highest function values are attained there. Only 19 spheres are drawn because the one that corresponds to $\min_i f_i$ is empty; a cross marks its center.

As such, the described set is difficult to work with. However, an alternative is to link covering algorithms to B&B as Reiner Horst did in [10]. Specifically, we apply the spheres in a B&B framework where n -simplices are used as suggested by Reiner in [8].

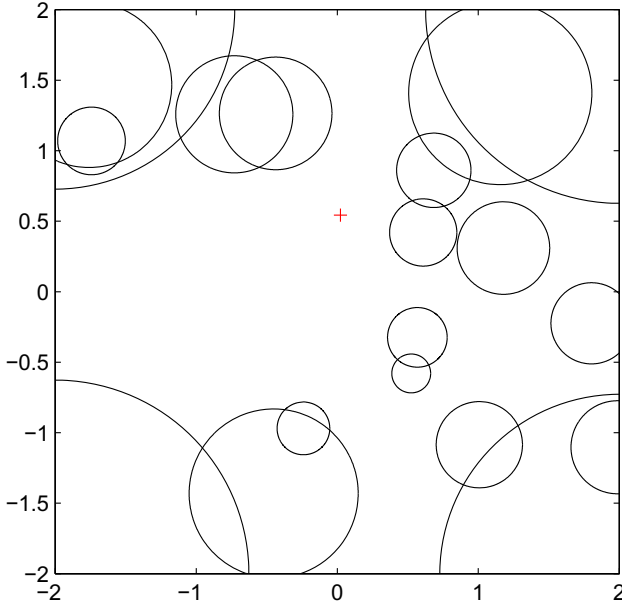


Fig. 2. Emmentaler set for six-hump camel-back evaluated at vertices and 16 additional sample points

4 B&B Simplicial Covering Algorithm

The use of value information on M and K aims at guaranteeing that in the end the best sampled point has a function value that differs less than a predefined accuracy δ from f^* ; $f^* \leq U \leq f^* + \delta$. This target can be reached by having a dense sampling, e.g. a grid. If a value for M is given and there is for every sampling point, another sampling point such that the distance in between them is at most $\varepsilon = 2\frac{\delta}{L}$, then any $x \in S$ is closer than $\frac{1}{2}\varepsilon$ from a sampled point, such that, $f(x) > U - \frac{1}{2}M\varepsilon = U - \delta, \forall x \in S$. In this case U is a δ -accuracy optimal solution.

Similarly, if a value for K is given, one can sample up to an accuracy of points being $\varepsilon = \sqrt{\frac{8\delta}{K}}$ apart to guarantee $f(x) > U - \frac{1}{2}K(\frac{1}{2}\varepsilon)^2 = U - \delta$. The essential of branch and bound is not to sample everywhere dense, but to remove areas where it has been proven that the optimum cannot be located. In the B&B method, the set is subsequently partitioned into more and more refined parts (branching) over which bounds of an objective function value, and in this case, non-optimality spheres can be determined. Parts completely covered by the spheres are deleted (pruning), since these parts of the domain cannot contain optimum solutions.

Algorithm 1. B&B algorithm.

Inputs: - S : box constrained feasible area
 - f : objective function
 - δ : accuracy
 - K : parabolic parameter

Output: - best proven solution x^U

Funcnt B & B Algorithm

1. $\varepsilon := \sqrt{\frac{8\delta}{K}}$, $\Lambda := \{C_1, \dots, C_p\}$ as first partition of S
 2. **for** sample points $x_i \in C_j, C_j \in \Lambda$ EvaluateVertex(x_i)
 3. **for** simplices $C_j \in \Lambda$ EvaluateSimplex(C_j)
 4. **while** $\Lambda \neq \emptyset$
 5. Take one subset C from list Λ according to a selection rule.
 Subdivide C into two new subsets C_{new_1} and C_{new_2} by splitting
 over the longest edge, generating new point x_k .
 6. EvaluateVertex(x_k),
 7. EvaluateSimplex(C_{new_1}), EvaluateSimplex(C_{new_2})
 8. return x^U
-

A possible algorithm based on bisection is outlined (see Algorithm [1](#)). The method starts with a partitioning of set S into simplices C_1, \dots, C_p to be stored as first elements of a list Λ of subsets (partition sets) and stops when the list Λ is empty. We also store the generated sample points x_i and their function value f_i on which the radius of the non-optimality spheres is based. Finally we keep track of points, that are proven to have a function value which differs less than δ from the global minimum f^* .

A generated subset C_k is not stored in Λ , if it can be proven that it is covered. In Sect. [5](#), results on proving coverage are discussed. Moreover, partition sets smaller in size than ε are discarded. The branching concerns the further refinement of the partition. This means that one of the subsets is selected to be split into new subsets. A selection rule determines the subset to be split next.

As discussed before based on the considerations in [7](#), an advantage of bisection splitting along the longest edge is due to the shape of the partition sets. The length of the longest edge is at most twice the size of the shortest edge. Therefore the sets can never get a needle shape.

Algorithm 2. Evaluate subset, decide to put on list based on cover

Funcnt EvaluateSimplex (C); global A, U, ε

1. **if** $size(C) > \varepsilon$
 2. Cover check of C by $\cup BK_i$
 3. **if** C not proven to be covered
 4. store C in A
-

Algorithm 3. Evaluate a point and update global information.

Funct EvaluateVertex (x); global A, U, x^U

1. Determine $f(x)$ either from stored points or evaluate
 2. **if** $f(x) < U$
 3. $U := f(x)$ and $x^U := x$ *Update global information*
 4. Update all BK_i and remove all $C_k \in A$ that are covered *Pruning*
-

5 Check on Covering a Simplex by Spheres

The question if simplex C is covered by spheres $B_i = \{x \mid \|x - v_i\| \leq r_i\}$ centered at its vertices v_i has been dealt with extensively in [3]. Notice that the vertices v_i are a subgroup of the evaluated points; $\{v_i \in C\} \subset \{x_1, \dots, x_k\}$. Even the question of covering the simplex by spheres at the vertices is not for each instance easy to verify. The following three rules are useful:

1. check first if one of the spheres alone covers C , i.e. $\max_j \|v_j - v_i\| < r_i$.
2. if an interior point $x \in C$ is covered by the intersection of spheres $x \in \cap_{v_i \in C} B_i$, all the simplex is covered, i.e. $C \subset \cup_{v_i \in C} B_i$. One can try a weighted average of the vertices.
3. the best point to check is the so called θ -point where $\|\theta - v_i\|^2 - r_i^2 = \|\theta - v_j\|^2 - r_j^2, \forall v_i, v_j \in C$. Even if θ is not interior, but covered, the whole simplex C is covered.

In the algorithmic context, the first rule is the easiest to check and should be tried first. The determination of the θ -point requires solving a set of n linear equalities. Consider the vertices v_1, \dots, v_{n+1} of C . Equating

$$(\theta - v_1)^T(\theta - v_1) - r_1^2 = (\theta - v_i)^T(\theta - v_i) - r_i^2, \quad i = 2, \dots, n + 1 \quad (16)$$

and bringing the terms with θ to the left hand side gives

$$2(v_i - v_1)^T \theta = r_1^2 - r_i^2 + v_i^T v_i - v_1^T v_1, \quad i = 2, \dots, n + 1 . \quad (17)$$

Example 3. Consider the following three spheres in 2-dimensional space:

$$v_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, v_3 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, r_1^2 = 4, r_2^2 = 3, r_3^2 = 1.$$

Point $\theta = \begin{pmatrix} 2.6 \\ 2.7 \end{pmatrix}$ can be determined equating the two planes (17) between v_1 and v_2 and between v_1 and v_3 , see Fig. 3. The corresponding solution has equal values $\|\theta - v_j\|^2 - r_j^2 = 10.05$ for the three vertices, v_1, v_2 and v_3 .

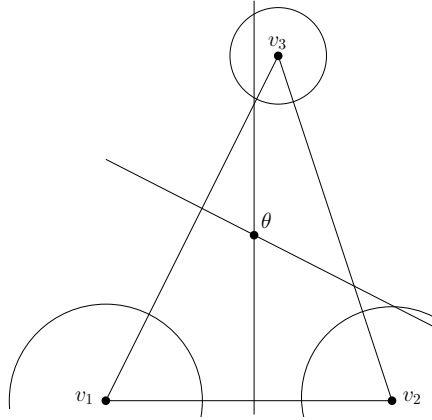


Fig. 3. Determination of the θ -point

It is interesting that the θ -point is closely related to the vertices that Breiman-Cutler keep track of in their algorithm. In fact, in [3] it is shown that if θ is interior with respect to C , then it is a global minimum point of ψ_C , defined similar as φ_k in (7), where one only considers the vertices of C . To be more precise:

$$\psi(x) = \max_{v_i \in C} \{ \|x - v_i\|^2 - r_i^2 \} . \tag{18}$$

Using (11), where $r_i^2 = 2\frac{f_i - U}{K}$, we can redefine

$$\psi_C(x) = \max_{v_i \in C} \{ f_i - \frac{1}{2}K\|x - v_i\|^2 \} . \tag{19}$$

Notice that $\psi_C(x) \leq \varphi_k(x)$, because $\{v_i \in C\} \subset \{x_1, \dots, x_k\}$. As has been shown in [3], $l(C) := \psi(\theta) \leq \min_{x \in C} \psi(x)$ is a lower bound of ψ_C over C . In that sense, $l(C)$ is also a lower bound of φ_k over C . The consequence of this theoretical results is that to check the cover, θ can be computed to find $l(C)$. If $l(C) > U$, then C cannot contain an optimum solution. For the underestimate based on M [8], one can also redefine the function ψ of (18). However, in that case also the ratio between radii (r_i/r_j) depends on the best function value found, U . That means, that also the θ -point depends on U . So, one can construct a similar test, but if an update of the global upper bound U has been found, the ratio changes, such that the θ -point is shifted.

Now getting back to the main question of our research that deals with the use of bisecting the longest edge by the midpoint. In the empirical work of running the algorithms, we found that if the θ -point and therefore minimum point of ψ , has more tendency to be inside the simplex under consideration, than in the case of needle shaped simplices. Concluding, *bisecting over the middle of the longest edge can be convenient for checking the cover of a simplex by non-optimality spheres centered at its vertices.*

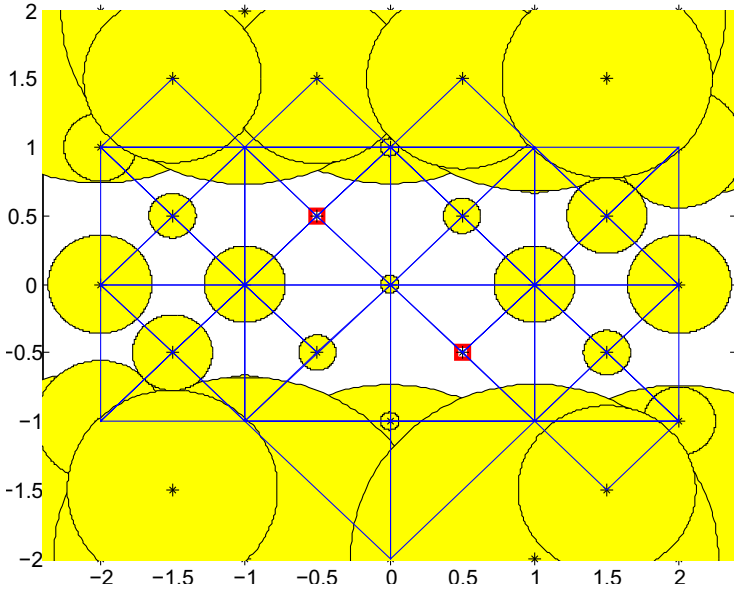


Fig. 4. Progress of a simple covering algorithm on six-hump after 50 iterations. The left over simplices, the evaluated points and their corresponding non-optimality spheres are depicted. The best points found are given by a small square.

If an inspected simplex C is not covered by the spheres at its vertices, it may still be covered completely by spheres centered at other points in $\{x_1, \dots, x_k\} \setminus \{v_i \in C\}$. To check this individually, one can run over the list of evaluated points $x_1, \dots, x_j, \dots, x_k$ and check whether

$$\max_{v_i \in C} \|x_j - v_i\|^2 < r_j^2 . \tag{20}$$

Intuitively, a sphere has more tendency to cover a ‘round’ simplex than a needle shaped one. In the illustration, (20) is used on the bisected partition sets.

6 Numerical Illustration

How does the development of using non-optimality spheres in simplicial B&B look like? The presented algorithm is rather generic as many details can be filled in. A simple illustration is given without any pretention to outperform other covering based algorithms. The algorithm was applied to the six-hump camel-back function, where an accuracy of $\delta = 0.0001$ and $K = 60$ were used. The only used cover check is the validation of (20) for all evaluated points and a breadth-first-search selection was applied. A list of ns simplices is maintained and the number $ndel$ of deleted simplices and number nf of function evaluations is measured during the iterations it in Table 1.

The algorithm converges after 622 iterations returning the global minimum points. Notice that at each iteration two simplices are evaluated and that about half of them are not put on the list in the first place. Figure 4 sketches the progress after 50 iterations. Proceeding, U is updated and consequently spheres increase. Figure 5 shows the state after 150 iterations. These figures also show well that many simplices are covered by a set of non-optimality spheres, but not by a single one, so test (20) is quite rough. The main interpretation of the illustration is that in fact more ‘round’ simplices have an earlier covering by individual spheres than needle shaped simplices, advocating the use of bisection as division rule.

Table 1. Progress of the B&B algorithm on six-hump, $\delta = 0.0001$ and $K = 60$

it	50	200	400	622
ns	36	96	19	0
$ndel$	16	106	383	624
nf	39	129	285	454
U	-1.13	-0.98	-1.02	-1.03

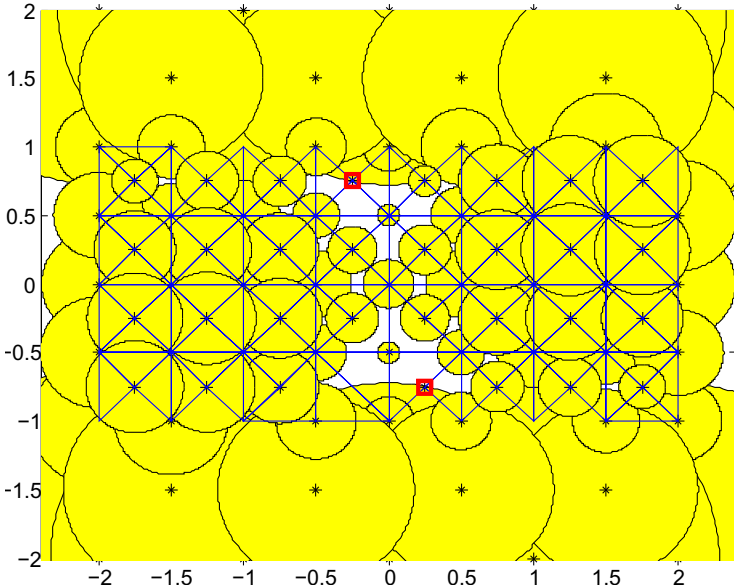


Fig. 5. Progress of a simple covering algorithm on six-hump after 150 iterations. The left over simplices, the evaluated points and their corresponding non-optimality spheres are depicted. The best points found are given by a small square.

7 Conclusions and Future Work

In a recent publication Reiner Horst stated that bisection is “not optimal” referring to volume considerations and convergence rates within a branch and bound tree.

In this paper we discuss several aspects for which the use of bisection in simplicial B&B may be convenient due to the feature of leading to relatively ‘round’ partition sets, and implicitly sampling over an equidistant grid. Bisection of the longest edge over the midpoint appears to be convenient for

- Robustness considerations searching for feasible ε -spheres;
- storage issues in branch and bound trees;
- checking the cover of a simplex by non-optimality spheres.

The first two observations follow from experience solving practical design problems by B&B. To elaborate the latter, a generic B&B algorithm has been outlined and several properties regarding non-optimality spheres in this context have been elaborated. As reported, relatively ‘round’ simplices appear to be convenient. Like Reiner, we also looked into other ways to divide simplices. Although convenient in low dimensional applications, regular equilateral subdivisions cannot be extended to higher dimensional branch and bound methods. Further research can even be focussed on overlapping equilateral subdivisions.

References

1. Baritomba, W.P.: Customizing methods for global optimization, a geometric viewpoint. *Journal of Global Optimization* 3, 193–212 (1993)
2. Breiman, L., Cutler, A.: A deterministic algorithm for global optimization. *Mathematical Programming* 58, 179–199 (1993)
3. Casado, L.G., García, I., Tóth, B.G., Hendrix, E.M.T.: On determining the cover of a simplex by spheres centered at its vertices. *Journal of Global Optimization* 50, 654–655 (2011)
4. Casado, L.G., Hendrix, E.M.T., García, I.: Infeasibility spheres for finding robust solutions of blending problems with quadratic constraints. *Journal of Global Optimization* 39, 557–593 (2007)
5. Danilin, Y., Piyavski, S.A.: An algorithm for finding the absolute minimum. *Theory of Optimal Decisions* 2, 25–37 (1967) (in Russian)
6. Evtushenko, Y., Posypkin, M.: Coverings for global optimization of partial-integer nonlinear problems. *Doklady Mathematics* 83, 1–4 (2011)
7. Horst, R.: On generalized bisection of n -simplices. *Mathematics of Computation* 66(218), 691–698 (1997)
8. Horst, R.: Bisection by global optimization revisited. *Journal of Optimization Theory and Applications* 144, 501–510 (2010)
9. Horst, R., Pardalos, P.M., Thoai, N.V.: *Introduction to Global Optimization, Non-convex Optimization and its Applications*, vol. 3. Kluwer Academic Publishers, Dordrecht (1995)
10. Horst, R., Tuy, H.: On the convergence of global methods in multiextremal optimization. *Journal of Optimization Theory and Applications* 54, 253–271 (1987)

11. Horst, R., Tuy, H.: Global Optimization (Deterministic Approaches). Springer, Berlin (1990)
12. Locatelli, M., Raber, U.: On convergence of the simplicial branch-and-bound algorithm based on ω -subdivisions. *J. Optim. Theory Appl.* 107, 69–79 (2000)
13. Mladineo, R.H.: An algorithm for finding the global maximum of a multimodal multivariate function. *Mathematical Programming* 34, 188–200 (1986)
14. Paulavičius, R., Žilinskas, J., Grothey, A.: Investigation of selection strategies in branch and bound algorithm with simplicial partitions and combination of lipschitz bounds. *Optimization Letters* 4, 173–183 (2010)
15. Piyavski, S.A.: An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics* 12, 57–67 (1972) (in Russian)
16. Raber, U.: A simplicial branch-and-bound method for solving nonconvex all-quadratic programs. *Journal of Global Optimization* 13, 417–432 (1998)
17. Raber, U.: Nonconvex All-Quadratic Global Optimization Problems: Solution Methods, Application and Related Topics. Ph.D. thesis, Trier University (1999)
18. Shubert, B.O.: A sequential method seeking the global maximum of a function. *SIAM Journal of Numerical Analysis* 9, 379–388 (1972)
19. Zilinskas, A., Clausen, J.: Subdivision, sampling, and initialization strategies for simplicial branch and bound in global optimization. *International Journal of Computers and Mathematics with Applications* 44, 943–955 (2002)
20. Zilinskas, A., Zilinskas, J.: Global optimization based on a statistical model and simplicial partitioning. *International Journal of Computers and Mathematics with Applications* 44, 957–967 (2002)

Application of Variance Analysis to the Combustion of Residual Oils

Manuel Ferreira and José Carlos Teixeira

University of Minho, Mechanical Engineering Department, Guimarães, Portugal
{ef,jt}@dem.uminho.pt

Abstract. Although the disposal of residual oils represents a major industrial problem it also opens an opportunity due to the energy recovery potential. If the combustion process is properly controlled the economic revenue could be added to the environmental benefits. For this purpose a test facility was developed which is based upon the application of an effervescent atomizer on a 200 kW furnace. The operating conditions were optimized through the application of a Taguchi technique for experimental planning. This enabled the analysis of 4 independent variables and 3 interactions. The CO concentration was the chosen as the control variable. The analysis on the variance data shows that the swirl and the Air Liquid Ratio are the most relevant variables for the optimization of the combustion.

Keywords: Effervescent atomization, Taguchi method, Combustion.

1 Introduction

The disposal of used oils is currently a problem of major concern. One of the main reasons refers to the volume of used oil, which in Portugal amounts to 35,000 tones per year of collected oil. Amongst the technical solutions available, the energy recovery through combustion or re-refining back to a virgin base oil are those preferable. Although there is considerable debate on the merits and drawbacks of either solution, the combustion solution is still one with great potential. The reasons are two fold: a) it represents the use of an energy resource with a high heating value, thus reducing the demand on the conventional fuels; b) the economic benefits relatively to the re-refining solution.

One of the main vectors of directive 75/439/EC on used oils, amended in 1987, is that, among the different options for recovery, priority is given to the regeneration over their incineration. However, several studies clearly demonstrate that, most of the member states of the EU do not favor regeneration of used oil but, on the contrary, are widely using used oil as fuel in industrial applications [1].

Out of the 1,730 kt of used oil accounted per year in the EU, roughly 50% is used as an energy source in the EU [1], mostly in cement kilns (35% of the burnt oil). Other options are based upon the use of conventional liquid fuel boilers. However, this solution raises various issues of concern, which result from the fuel

characteristics: high viscosity, carbon deposits, and particulate emissions. These characteristics yield a problematic atomization (requiring pre-heating) and subsequent deficient combustion resulting in soot and gaseous emissions.

It is well understood that pollutant formation can be mitigated through correct combustion, which in turn depends upon the atomization of the liquid. Of particular interest is the NO_x formation, which is closely related with the droplet size. One of the main mechanisms of NO_x formation depends on the temperature and residence time of the combustion mixture, which should be the lowest (thermal NO_x). This requirement can be met by producing a nearly homogeneous air/fuel mixture and burning far from stoichiometric conditions (lean or rich).

Previous studies have provided a very comprehensive characterization of an effervescent atomizer [2,3]. Because of the internal dynamics of the nozzle, the results have shown that a very fine spray can be produced even at very low operating pressures and with no previous heating required. The outer orifice is fairly large and the nozzle obstruction is likely to be reduced. These characteristics make the effervescent atomizer very promising for used oil combustion.

Effervescent atomization is a method of twin-fluid atomization that involves bubbling a small amount of gas into the liquid fuel stream upstream the discharge orifice of the atomizer. This technique was first developed by Lefebvre and his co-workers in the late 1980s [4,5,6,7]. The term "effervescent" was later introduced by Buckner and Sojka [8]. Chawala [9] attributed the better atomization performance of twin-fluid techniques over single-fluid techniques to the substantial difference in the speed of sound between single and two-phase media. The sonic velocity in a liquid/gas mixture is substantially lower than that in either the gas or liquid phases.

Over the past decade, various experimental studies have been carried out to determine the performance and spray characteristics of effervescent atomizers over a wide range of operating conditions. Sovani et al [10] presented a very comprehensive review of the effervescent atomization. Of particular interest is the application in combustion systems, especially with low value and poorly refined fuels containing high levels of impurities and widely varying physical properties. Features like lower injection pressures, smaller drop sizes, smaller gas flow rates and larger nozzle orifices make this type of atomizer very promising for applications with such fuels. Another advantage of the effervescent atomizer-produced sprays is the presence of air (atomizing gas) in the spray core, which increases the air/fuel mixing process, yielding a reduction in the pollutant emissions.

Practical applications of the effervescent atomizer in combustion systems include gas turbine combustors, furnaces and boilers, IC engines and incinerators [10]. Sankar et al [11] developed a swirl effervescent atomizer for application in industrial and residential boilers. The combustion studies only included qualitative observations being that the flame produced by kerosene combustion was completely blue as the most relevant. This suggested an absence of soot, indicating a complete combustion of the fuel that resulted from the fine atomization of the liquid fuel.

Loebker and Empie [12] designed an effervescent atomizer for spraying a pulping industry by-product, called black liquor (a viscous liquid of widely varying composition with up to 80% solid suspension), into a heat recovery boiler. The

performance of this atomizer was compared with a conventional spray system Vee-Jet™ nozzle, used with the black liquor. They found that while the near nozzle structure of the Vee-Jet nozzle showed a mesh of interwoven, unbroken strands of liquid, effervescent atomizer showed much smaller liquid fragments and drops.

The main objective of this study was to apply a variance analysis to the optimization of the operating parameters of an effervescent atomizer applied to the combustion of used recycled oils.

2 Test Facility

The effervescent atomizer used in this study was based on the design presented in [3]. In order to enable its assembly in the burner setup some modifications have been introduced. Figure 1 shows the detailed design of the plain-orifice effervescent atomizer. The aerator consisted of a brass tube with 6.4 mm inside diameter and 82.9 mm long, perforated with 96 holes with 0.75 mm diameter, arranged in a 8x12 staggered layout. The oil and air flows were injected through the top of the atomizer, as shown in Figure 1, to enable the assembling into the burner. The atomizer's body had two straight holes, 10 mm in diameter, to accommodate the two ignition electrodes.

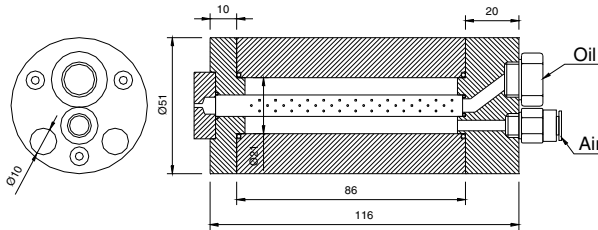


Fig. 1. Detailed design of the plain-orifice effervescent atomizer

Figure 2 shows schematically the main components of the test rig: furnace (1), oil burner (2), auxiliary gas burner (3) and propane gas supply (4), oil supply system (5), atomizing (or primary) air supply (6), secondary air supply (7), cooling system (8) and flue gases exhaust (9).

The furnace consists of a cylindrical combustion chamber, with an inside diameter of 0.5 m and 2.7 m long. The furnace is of modular design with five water-cooled steel segments: one, closer to the nozzle, 1 m long; a second with a length of 0.8 m and three with 0.4 m in length. The segment closer to the burner (1 m long) was lined with a 0.115 m thick layer of refractory. All of them had a cooling water jacket 0.18 m in thickness. The outer surfaces were insulated with a ceramic fiber 0.05 m thick. Along the furnace the segments had seven pairs of diametrically opposed windows with 0.110 m diameter and 0.4 m apart from each other, in order to provide optical access to the flame. Various steel pipes with ½" nominal diameter were also

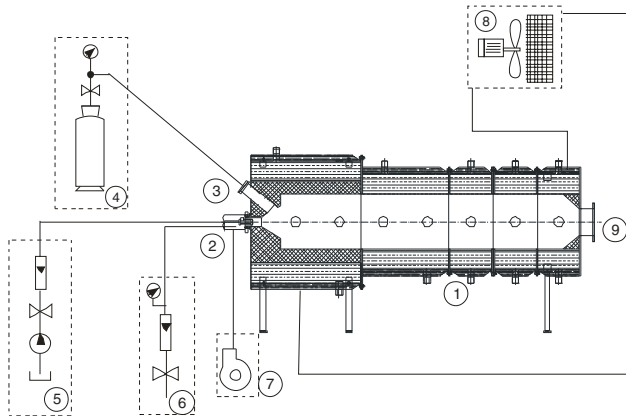


Fig. 2. Layout of the test rig

introduced through the furnace wall to enable other measurements such as furnace static pressure and flame temperature. A set of 21 thermocouples were located at various points inside the furnace walls to monitor wall temperatures. Flue gases exit the furnace through a stack with an internal diameter of 0.25 m and 10 m high, made of stainless steel, insulated with mineral fiber.

The oil burner included the effervescent atomizer and a rotary vane tangential swirl generator. The swirl intensity was regulated by adjusting the angle of the rotary vanes (Figure 3). The secondary air was injected tangentially between the burner gun (Figure 1) and the burner tube with a diameter of 0.085 m. This burner was controlled by an oil control Danfoss BHO 64 connected to ignition electrodes, flame photo-detector, secondary-air pressure sensor and an oil electro-valve.

The furnace also incorporated a single-port atmospheric-type gas burner for heating the refractory prior to oil firing. This burner was controlled by a specific gas burner control Pactrol CSA6, performing ignition, flame detection and gas shut off. After being turned off, this burner was carefully sealed to avoid unwanted air entering the furnace.

The oil supply system, already described elsewhere [3], comprised a spur gear pump connected to an oil container (1 m³), a helical screw flow meter and a pressure gauge. The atomizing air (dry and filtered) was supplied by the compressed air mains. A rotameter with a pressure gauge and a needle valve enabled the control and measurement of pressure and air flow-rate. Both flow meters were individually calibrated.

Secondary air was supplied by a centrifugal fan into the swirl generator (Figure 3). The flow-rate was measured at the entrance of the fan with a bell shaped inlet section. The flow-rate was determined by measuring the static pressure at the throat of the bell shaped entrance, which was calibrated by calculating the discharge coefficient, CD . This was determined by measuring the velocity profile of a transparent pipe downstream of the ventilator exit, using a 2D measuring LDA system. An average CD of 0.96 was obtained, demonstrating the high efficiency of the bell shaped entrance.

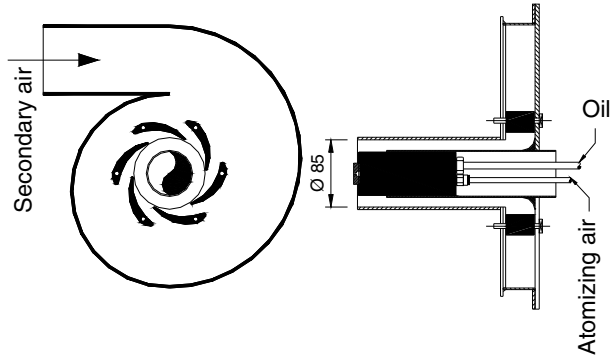


Fig. 3. Schematic of the swirl generator

Used recycled oil supplied by AUTO VILA, reference OQ1 was used throughout the experiments. The most important properties were determined: $\rho=898 \text{ kg/m}^3$; $\nu=46 \text{ mm}^2/\text{s}$; higher heating value= 44.6 MJ/kg ; lower heating value= 41.8 MJ/kg .

3 Methodology

Taking the furnace and nozzle configuration as a given constant, the efficiency of the combustion process is dependent upon a wide range of parameters. Some are operating variables; others are design features. The choice of the most appropriate configuration for the combustion operation requires the application of an optimization procedure to the effervescent atomizer, described in the previous section. For this purpose and to evaluate the most relevant parameters in the system design, an experiment plan was organized according to the Taguchi method and subsequently processed according to a variance analysis, ANOVA [13].

This method is essentially made up of 3 phases: 1) selection of parameters; 2) experiments planning and 3) data analysis and interpretation.

In order to capture changes in gradient over the range, it was decided to evaluate the influence of each variable in 3 levels. In this way a set of 27 experiments defines a orthogonal matrix, L27, with 13 columns. They can be used to evaluate the influence of 7 parameters and 3 interactions (each one requires 2 columns, [13]).

The choice of appropriate variables and the corresponding levels is crucial to the method success. Based upon the experience obtained in the design (which included a number of preliminary runs) and the understanding the atomization/combustion processes, the number of variables was reduced to 4, as listed in Table 1. All the experiments were carried out at a constant flow rate (9.9 kg/h) and for one geometric configuration of the atomizer (3 orifices, 1 mm in diameter, at a diverging angle of 15°). The table also includes the values used for the three levels.

Table 1. Variables and corresponding levels used in the Taguchi plan of experiments

FACTOR		levels		
		1	2	3
A	Excess of air	0.20	0.30	0.40
B	Oil feeding pressure (bar)	4.5	5.0	5.5
C	<i>swirl</i>	0	0.25	0.5
D	ALR	0.25	0.35	0.45

Table 2. Orthogonal L27 matrix for the Taguchi plan

	Exc. Air	Pressure			Swirl				ALR				
	A	B	AxB		C	AxC		BxC	D	e	BxC	e	e
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2	2	2	2	2	2
3	1	1	1	1	3	3	3	3	3	3	3	3	3
4	1	2	2	2	1	1	1	2	2	2	3	3	3
5	1	2	2	2	2	2	2	3	3	3	1	1	1
6	1	2	2	2	3	3	3	1	1	1	2	2	2
7	1	3	3	3	1	1	1	3	3	3	2	2	2
8	1	3	3	3	2	2	2	1	1	1	3	3	3
9	1	3	3	3	3	3	3	2	2	2	1	1	1
10	2	1	2	3	1	2	3	1	2	3	1	2	3
11	2	1	2	3	2	3	1	2	3	1	2	3	1
12	2	1	2	3	3	1	2	3	1	2	3	1	2
13	2	2	3	1	1	2	3	2	3	1	3	1	2
14	2	2	3	1	2	3	1	3	1	2	1	2	3
15	2	2	3	1	3	1	2	1	2	3	2	3	1
16	2	3	1	2	1	2	3	3	1	2	2	3	1
17	2	3	1	2	2	3	1	1	2	3	3	1	2
18	2	3	1	2	3	1	2	2	3	1	1	2	3
19	3	1	3	2	1	3	2	1	3	2	1	3	2
20	3	1	3	2	2	1	3	2	1	3	2	1	3
21	3	1	3	2	3	2	1	3	2	1	3	2	1
22	3	2	1	3	1	3	2	2	1	3	3	2	1
23	3	2	1	3	2	1	3	3	2	1	1	3	2
24	3	2	1	3	3	2	1	1	3	2	2	1	3
25	3	3	2	1	1	3	2	3	2	1	2	1	3
26	3	3	2	1	2	1	3	1	3	2	3	2	1
27	3	3	2	1	3	2	1	2	1	3	1	3	2

Another important decision is the choice of control variables. These should represent the characteristic that is the most relevant in terms of operation. In combustion systems, the efficiency is paramount and CO or HC concentration in the flue gases could either be used as control variables. In the present case, the CO concentration was selected. Throughout the experimental program it was registered the concentration of CO, CO₂ and O₂ along with the temperature in the furnace and the shape and dimensions of the flame.

For each factor the number of degrees of freedom equals the number of levels minus 1, being in the present case, 2. Interactions between two factors have 4 (2x2) degrees of freedom. As each column in the L27 matrix has 2 degrees of freedom, each interaction between independent variables requires the use of two columns. It was decided that the study should focus in: excess air/pressure; excess air/swirl and pressure/swirl. Table 2 summarizes the experimental plan according to the L27 matrix.

This matrix details the various levels defined for the 4 variables and the resulting interactions for the 3 combinations selected. Columns 10, 12 and 13 are not used.

4 Results and Discussion

The implementation of this experimental plan is detailed in Table 3, for the 4 variables, which results from the combination of Tables 1 and 2.

The table also details the experimental results. Although the control variable is the CO concentration (corrected for 8% O₂), the data for the CO₂ and O₂ levels and Flame Temperature are also included as they may bring further insight into the discussion. The objective is the minimization of CO concentration.

In any optimization process the deviation from the optimal value depends on both the mean value and data dispersion. In this way the variance analysis (ANOVA) was implemented. This requires the mean and variance data for each variable which can be used to test the significance of the data and the contribution of noise and uncertainties.

The variance analysis is based upon the determination of the signal to noise ratio (S/R), according to:

$$S/R = -10 \log(\bar{X}^2 + \sigma_x^2) \quad (1)$$

where \bar{X} is the mean and σ_x^2 the variance.

For each experimental run the CO concentration was recorded over a period of time, such as represented in Figure 4. From this time series both the average, \bar{X} , and the variance σ_x^2 were calculated.

Table 3. Experimental plan and results

Test #	A	B	C	E	Variable				Flame Temp °C
	Excess of Air	Pressure	Swirl	ALR	CO ₂	O ₂	CO	CO (8% O ₂)	
1	0.20	4.5	0	0.25	12.19	3.92	123.7	94.0	854
2	0.20	4.5	0.25	0.35	11.39	4.97	38.0	30.8	871
3	0.20	4.5	0.5	0.45	11.48	4.69	28.9	23.0	855
4	0.20	5	0	0.35	11.15	5.33	18.4	15.3	851
5	0.20	5	0.25	0.45	11.04	5.28	22.3	18.5	863
6	0.20	5	0.5	0.25	11.57	4.59	36.2	28.7	857
7	0.20	5.5	0	0.45	11.66	4.51	37.6	29.6	831
8	0.20	5.5	0.25	0.25	11.43	4.72	31.2	24.8	871
9	0.20	5.5	0.5	0.35	11.52	4.67	19.7	15.6	862
10	0.30	4.5	0	0.35	10.84	5.56	25.8	21.7	818
11	0.30	4.5	0.25	0.45	10.50	6.04	29.6	25.7	826
12	0.30	4.5	0.5	0.25	11.00	5.40	18.4	15.3	858
13	0.30	5	0	0.45	10.85	5.59	27.3	23.0	820
14	0.30	5	0.25	0.25	10.50	6.08	21.6	18.8	865
15	0.30	5	0.5	0.35	10.96	5.48	19.8	16.6	869
16	0.30	5.5	0	0.25	10.56	5.97	165.4	141.9	830
17	0.30	5.5	0.25	0.35	10.67	5.84	19.0	16.2	844
18	0.30	5.5	0.5	0.45	10.72	5.73	12.6	10.7	846
19	0.40	4.5	0	0.45	10.11	6.34	57.2	50.7	808
20	0.40	4.5	0.25	0.25	9.94	6.74	23.9	21.8	843
21	0.40	4.5	0.5	0.35	10.02	6.64	26.2	23.7	823
22	0.40	5	0	0.25	9.26	7.36	335.1	355.5	790
23	0.40	5	0.25	0.35	9.83	6.88	30.0	27.6	856
24	0.40	5	0.5	0.45	9.95	6.79	23.7	21.7	820
25	0.40	5.5	0	0.35	10.10	6.61	40.2	36.2	810
26	0.40	5.5	0.25	0.45	9.69	7.05	37.8	35.2	839
27	0.40	5.5	0.5	0.25	9.69	7.04	22.6	21.1	842

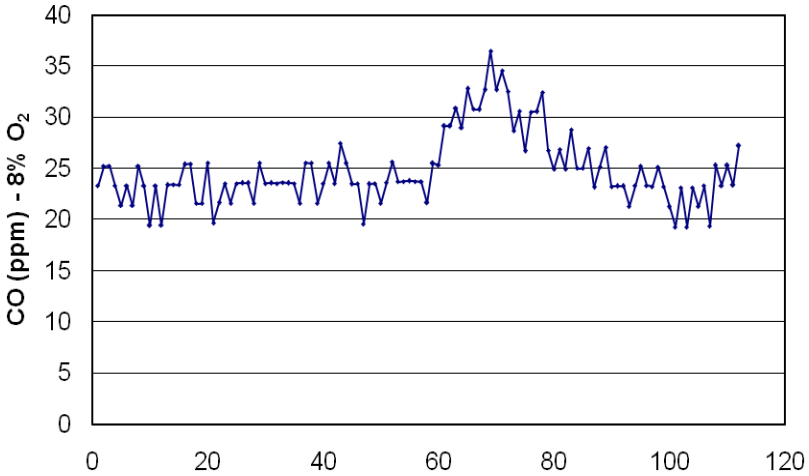


Fig. 4. Time series (in s) of CO concentration

The *S/R* ratio, calculated for each response, is presented in Table 4 for each one of the three levels assigned to each variable. The bottom line shows the maximum difference for each parameter which can be used to assess the relative relevance of each parameter.

Table 4. Variance analysis for *S/R* ratio

<i>S/R</i>	A	B	AxB	C	AxC	BxC	D	e	BxC	e	e		
1	-28.54	-29.40	-32.63	-29.23	-34.29	-27.63	-28.00	-29.38	-32.84	-28.99	-27.98	-27.71	-32.93
2	-27.16	-29.32	-27.33	-28.36	-27.57	-29.05	-30.06	-28.90	-26.77	-29.13	-29.95	-30.32	-28.17
3	-31.74	-28.72	-27.48	-29.85	-25.57	-30.75	-29.38	-29.15	-27.83	-29.31	-29.51	-29.40	-26.34
dif.	4.58	0.68	5.29	8.72	3.11	1.98	6.07	0.32	1.98	2.61	6.59		

The data shows that the swirl (C) is the most significant variable affecting the CO reduction, followed by the ALR (D) and the excess of air (A). Injection pressure has a negligible influence. In addition it is observed that unaccounted factors may contribute to the reduction of CO (right hand column). The strong influence of the swirl is supported by early observations that a correct adequacy of the flow rotation index to the combustion chamber and the spray pattern is paramount to an efficient combustion. Similarly the ALR is relevant for its direct contribution to the spray atomization and a correct stoichiometry in the core of the spray.

Table 5 presents the response of the system based upon the CO concentration, which shows a pattern that closely follows that just discussed for the *S/R* ratio. This is expected as the objective of the experimental program is to minimize the response of the system.

Table 5. Variance analysis for CO concentration

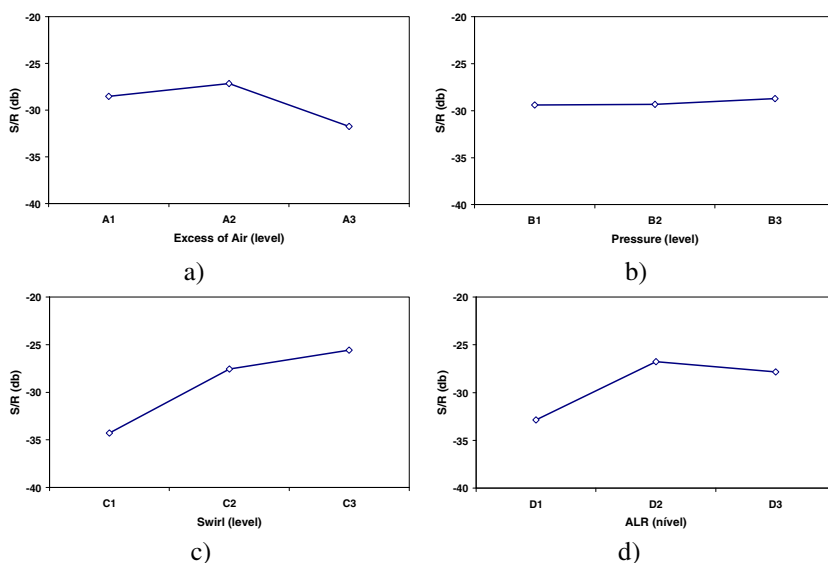
Average	A	B	AxB		C	AxC		BxC	D	e	BxC	e	e
1	31.14	34.07	80.15	33.18	85.34	29.56	29.56	34.40	80.21	32.71	30.96	29.15	80.74
2	32.22	58.40	24.18	36.37	24.36	36.35	62.12	57.72	22.64	38.36	39.22	61.63	26.99
3	65.93	36.82	24.96	59.73	19.59	63.38	37.61	37.18	26.44	58.22	59.12	38.51	21.55
dif.	34.79	24.33	55.97		65.75	33.82		28.16	57.57	25.51		32.48	59.19

Figure 5 shows the response to the different parameters for the three levels assumed. This analysis enables to assess the direction that each parameter has on CO reduction.

As mentioned the injection pressure has a negligible contribution to the CO reduction (Figure 5-b). The influence of the swirl is monotonic within the range tested. In other words, by increasing the swirl results in a more efficient combustion (lower CO concentration). It should be stressed that this analysis is based on the wider spray cone angle (3 orifices at a diverging angle of 15 °) which proved to be highly dependent on the combustion air rotation. For a narrower spray cone angle, this influence may be opposite.

The ALR contribution (Figure 5.d) shows that an increase in ALR is beneficial for low CO (level 1 to 2) but may contribute to a reduction in efficiency at high ALR (level 2 to 3). In fact, a detailed analysis of the influence of the ALR upon the spray dynamics shows that the atomization efficiency levels off at high ALR (Ferreira, 1999).

For the excess air a similar pattern is observed as the combustion efficiency decreases for high excess air (level 2 to 3). By increasing the excess air, the air fuel mixture becomes leaner, the flame temperature reduces (with the combustion kinetics) which leads to an increase in CO concentration.

**Fig. 5.** Influence of individual factors in the system response

From the data analysis it should be highlighted the strong interaction between the excess air and the injection pressure. However, taking into account that the injection pressure has a negligible influence (when considered alone) and because there is strong evidence that unaccounted factors are important, one may conclude that the interaction between the excess air and the injection pressure is the direct result of random errors. Figure 6-a) shows no definitive trend. The other parameters show no cross interaction as the trend lines are basically parallel (Figure 6-b,c).

In the variance analysis, F is the ratio between the variance of a parameter and the variance of the error. This ratio is shown in Table 6. For a certain confidence level, a high value of F means that the variance associated to a certain factor is greater than the variance associated to error. In other words, the influence of that parameter is significant. Defining the confidence level and knowing the number of degrees of freedom (df), a critical F can be determined. For a 99% confidence level and 2 degrees of freedom this is 5.72.

The data shows that the F ratio associated to the swirl is higher than the critical ratio. Therefore one may conclude that with a confidence level of 99% the swirl intensity has a contribution of up to 27% in the reduction of CO concentration. On the other hand the data for ALR shows that a confidence level of 99% is too high to conclude on its significance in the reduction of CO concentration. By reducing the confidence level the influence of this parameter may be accepted. In fact considering a confidence level of 95% (critical F of 3.44) it can be concluded that the ALR has an influence of up to 11% in the reduction of CO concentration.

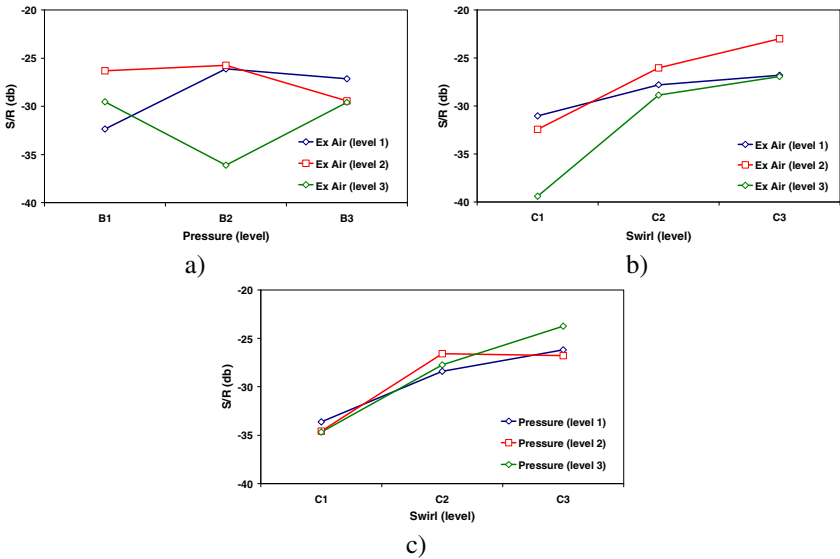


Fig. 6. Interaction between factors

Also from the variance analysis it may be concluded that the experimental error has a higher contribution (60.86%) than either the swirl intensity or the ALR. Various factors may contribute to this: unaccounted variables, inappropriate levels for the variables, inaccurate control of the variables and instabilities during operation.

Table 6. ANOVA table. Determination of the statistically meaningful data

ANOVA	Exc. air	Pressure			Swirl		ALR				Error	
	1	2	3 e 4	5	6 e 7	8 e 11	9	10	12	13	Experi.	total
	A	B	AxB	C	AxC	BxC	D	e	e	e		
df	2	2	4	2	4	4	2	2	2	2	22	26
sq	99.34	2.51	173.74	375.60	63.57	20.36	189.46	0.47	31.61	208.44	600.04	1165.10
var	49.67	1.25	43.43	187.80	15.89	5.09	94.73	0.23	15.81	104.22	27.27	44.81
pool	y	y	y	n	y	y	n	y	y	y	n	
F				6.89			3.47					
sq'				321.05			134.91				709.14	1165.10
%				27.56			11.58				60.86	100.00

In fact a detailed analysis of the CO concentration (see Figure 4) shows the occurrence of occasional instabilities in the operating conditions, most likely due to uncontrolled conditions. During the furnace operation acoustic instabilities were observed in the combustion process which can result from fluctuations in the air/fuel pressure/flow rate (which may lead to conditions outside the stable operation) and the on-off operation of the cooling loop.

Another point worth mentioning concerns the application of the Taguchi method. Typically the selection of variables and their levels requires some historic background and experience on the process which is not the case at present; no similar analysis has been reported in the literature. In a certain way the preliminary tests (not reported here) were used to overcome this limitation, but it is certainly a source of concern. Nonetheless, the technique proved its potential.

5 Conclusions

The present work reports the application of a variance analysis coupled with a Taguchi method for experiment planning to the optimization of a combustion process. This is based upon the application of an effervescent nozzle to the atomization of residual oils. The experimental program was based on the measurement of the CO concentration as an objective function associated with the combustion efficiency. The following conclusions can be drawn:

1. The Taguchi method was used to plan the set of experiments based on the analysis of the contribution of 4 variables at 3 levels. In this way, 27 experiments enable the assessment of 4 variables and 3 interactions.

2. The swirl intensity and the ALR are the most significant individual variables contributing to the reduction of CO concentration. Injection pressure has a marginal influence.
3. The analysis of the interaction between parameters shows that no reliable conclusions can be obtained.
4. The variance analysis enable the conclusion that with a confidence level of 99% the swirl intensity contributes up to 27% to the reduction of CO concentration. As far as the ALR is concerned, its contribution to the same objective function is down to 11% though with a lower confidence interval (95%).
5. The error associated with variance analysis suggests that unaccounted factors may blur the contributions of individual variables. Various causes may be identified, though instabilities in the operation of the furnace may be a prime factor.

References

1. Monier, V., Labouze, E.: Critical Review of Existing Studies and Life Cycle Analysis on the Regeneration and Incineration of Waste Oil. European Commission, DG Environment (2001)
2. Bates C.J., Bowen, P., Teixeira, J.C.F.: Influence of exit orifice characteristics on transition between effervescent atomization flow regimes. In: Proc. of the ICLASS 2000 (2000)
3. Ferreira, M., Teixeira, J.C.F., Bates, C.J., Bowen, P.J.: Detailed investigation of the influence of fluid viscosity on the performance characteristics of plain-orifice effervescent atomizer. *Atomization and Sprays* 11, 107–124 (2001)
4. Lefebvre, A.H., Wang, X.F., Martin, C.A.: Spray characteristics of Aerated-liquid pressure atomizers. *AIAA J. Prop. Power* 4(4), 293–298 (1988)
5. Roesler, T.C., Lefebvre, A.H.: Studies on aerated-liquid atomization. In: ASME Winter Annual Meeting, Boston, Massachusetts, Paper 87-WA/HT-17 (1987)
6. Roesler, T.C., Lefebvre, A.H.: Studies on aerated-liquid atomization. *Int J. Turbo. Jet Engines* 6, 221–230 (1989)
7. Wang, X.F., Chin, J.S., Lefebvre, A.-H.: Influence of gas injector geometry on atomization performance of aerated-liquid nozzles. *Int. J. Turbo Jet Engines* 6, 271–280 (1989)
8. Buckner, H.E., Sojka, P.E.: Effervescent atomization of high viscosity fluids. Part 1: Newtonian liquids. *Atomization and Sprays* 1, 239–252 (1991)
9. Chawala, J.B.: Atomization of liquids employing the low sonic velocity in liquid/gas mixture. In: Proc. of the Third International Conf. on Liquid Atomization and Spray Systems, pp. LP/1A/5/1–LP/1A/5/7 (1985)
10. Sovani, S.D., Sojka, P.E., Lefebvre, A.H.: Effervescent Atomization. *Progress in Energy and Combustion Science* 27, 483–521 (2001)
11. Sankar, S.V., Robart, D.M., Bachalo, W.D.: Swirl Effervescent Atomizer for Spray Combustion. *ASME HTD* 317-2, 175–182 (1995)
12. Loebker, D., Empie, H.J.: High Mass Flow-rate Effervescent Spraying of High Viscosity Newtonian Liquid. In: Proceedings of the 10th Annual Conference on Liquid Atomization and Spray Systems, Ottawa, ON, Canada, pp. 253–257 (1997)
13. Rhoss, P.J.: *Aplicações das Técnicas de Taguchi na Engenharia da Qualidade*. McGraw-Hill (1991)

Warehouse Design and Planning: A Mathematical Programming Approach

Carla A.S. Geraldés¹, Maria Sameiro Carvalho², and Guilherme A.B. Pereira²

¹ Centro ALGORITMI and Polytechnic Institute of Bragança, Portugal
carlag@ipb.pt

² Centro ALGORITMI and University of Minho, Portugal
sameiro@dps.uminho.pt, guilherme.pereira@algoritmi.uminho.pt

Abstract. The dynamic nature of today's competitive markets compels organizations to an incessant reassessment in an effort to respond to continuous challenges. Therefore, warehouses as an important link in most supply chains, must be continually re-evaluated to ensure that they are consistent with both market's demands and management's strategies. A number of warehouse decision support models have been proposed in the literature but considerable difficulties in applying these models still remain, due to the large amount of information to be processed and to the large number of possible alternatives. In this paper we discuss a mathematical programming model aiming to support some warehouse management and inventory decisions. In particular a large mixed-integer nonlinear programming model (MINLP) is presented to capture the trade-offs among the different inventory and warehouse costs in order to achieve global optimal design satisfying throughput requirements.

Keywords: Supply chain management, Warehouse models, Inventory management, Mathematical modelling.

1 Introduction

Market competition requires continuous improvement in the design and operation of supply chains. A supply chain can be considered as a network of entities whose efficiency and effectiveness is highly determined by the performance of the overall network (see Fig. 1).

In a supply chain network, products need to be physically moved from one location to another. During this process, they may be buffered or stored at certain facilities (warehouses) for a certain period of time for strategic or tactical reasons. Within this context, warehouses play an important role in supply chains and are a key aspect in a very demanding, competitive and uncertain market.

On the other hand, modern supply chain principles compel companies to reduce or eliminate inventory levels. Additionally warehouses require capital, labour, and information technologies, which are expensive resources. So, why do we still need warehouses?

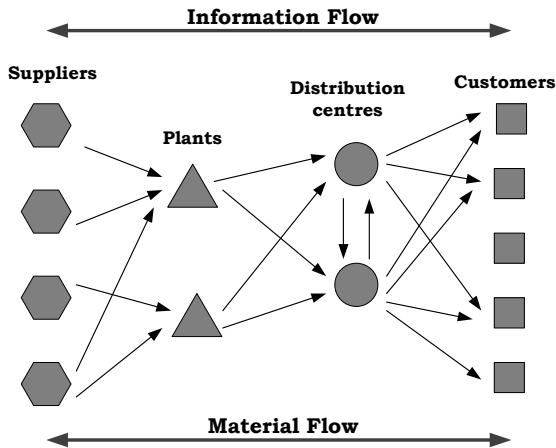


Fig. 1. A typical supply chain network

According to Bartholdi and Hackman [1] there are four main reasons why warehouses are useful:

1. To consolidate products in order to reduce transportation costs and provide customer service;
2. To take advantage of economies of scale;
3. To provide value-added processing services, and
4. To reduce response time.

Thus, warehouses will continue to be an important node at the logistic network by the fact that if a warehouse cannot process the orders quickly, effectively, and accurately, then all the supply chain optimization efforts will suffer (see Tompkins [12]).

In distribution logistic where market competition requires higher performances from the warehouses, companies are compelled to continuously improve the design and planning of warehouse operations. Furthermore, the ever increasing variety of products, the constant changes in customer demands and the adoption of agile management philosophies also bring new challenges to reach flexible structures that provide quality, efficiency and effectiveness to the logistics operations.

Some major decisions involved in the warehouse design and operation problems are illustrated in Fig. 2 (see Gu et al. [4]). Warehouse design and planning decisions typically run from a functional description, through a technical specification, to equipment selection and determination of the layout. The overall structure decision determines the material flow patterns within the warehouse, the specification of the functional areas and the flows between the areas. Sizing and dimensioning decisions determine the total size of the warehouse as well as the space allocation among functional areas. Layout definition is the detailed configuration within a functional area and equipment decisions define

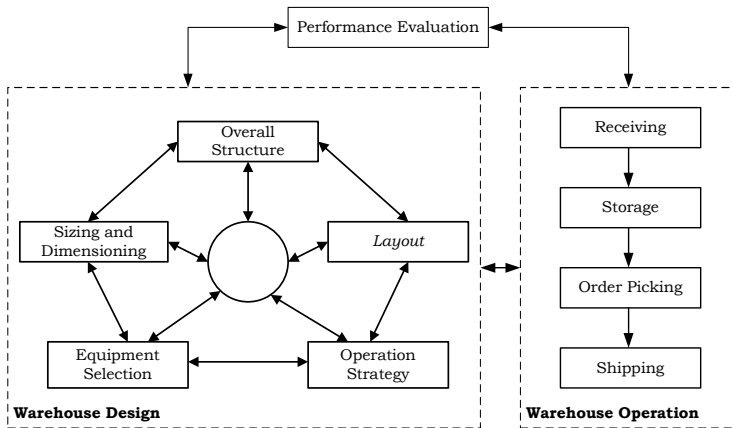


Fig. 2. A framework for design and operation problems (adapted from Gu et al. [4])

an automation level for the warehouse and identify equipment types. Finally operating policies refer to storage, picking and routing decisions.

Hassan [7] presented a framework for the design of a warehouse. The proposed framework accounts for several factors and operations of warehousing such as:

1. Specification of warehouse type and purpose;
2. Analysis and forecasting of the demand;
3. Definition of operating policies;
4. Establishment of inventory levels;
5. Class formation;
6. Definition of functional areas and general layout;
7. Storage partition;
8. Selection of equipment for handling and storage;
9. Design of aisles;
10. Determination of space requirements;
11. Location and number of Input/Output points;
12. Location and number of docks;
13. Arrangement of storage;
14. Zone formation.

Once these warehouse decisions are strongly interrelated, warehouse design is a highly complex task where conflicting objectives impose specific trade-offs.

Despite the various decision models available in scientific literature, the majority addresses isolated or simplified problems in order to provide the best solution. However, most of the real problems are unfortunately not well-defined and often cannot be reduced to multiple isolated sub-problems. Therefore, warehouse design often requires a mixture of analytical skills and creativity. Anyhow, research aiming an integration of various decisions models and methods is badly

needed in order to develop a methodology for systematic warehouse design (see Rouwenhorst et al. [10]).

Furthermore, as the literature review in next section shows, most research efforts have been dedicated to warehouse operations decisions instead of design decisions. This is not surprising since design decisions models are more difficult to develop and treat analytically once they require the integration of several and complex issues.

In this paper we present a warehouse and inventory mathematical model that jointly integrates issues concerning:

- The size of the warehouse;
- The external storage additional capacity, if needed;
- The replenishment quantities and reorder points of products to be stored.

Our aim is to test an integrated approach that takes into account inventory and some warehouse design decisions.

In the Section 2 of this paper we will present a brief literature review on warehouse design and planning issues. The purpose of this section is not limited to the specific studied problem but also covers other important topics in warehouse literature. The Section 3 will present the proposed warehouse sizing and inventory model formulation and a description of the methodology used to solve the model. Computational results will be presented and summarized in Section 4, and finally some conclusions and future work directions are reported in Section 5.

2 Literature Review

Warehousing is concerned with all the material handling activities that take place within a warehouse. It includes the receiving of products, storage, order-picking, accumulation, sorting and shipping operations. Basically, one can distinguish two types of warehouses: *distribution warehouses* and *production warehouses*. According to Van den Berg and Zijm [14], a distribution warehouse is a warehouse in which products from different suppliers are collected (and sometimes assembled) for delivery to a number of customers. On the other hand, a production warehouse is used for the storage of raw materials, semi-finished products and finished products in a production facility.

There are many activities that occur at a warehouse. Typically, distribution warehouses receive products - Stock Keeping Units (SKUs) - from suppliers, unload products from the transport carrier; store products, receive orders from customers, assemble orders, repackage SKUs and ship them to their final destination. Frequently, products arrive packaged on large scale units and are packaged and shipped on small units. For example, SKUs may arrive in full pallets but must be shipped in cases. Fig. 3 shows the typical functional areas and flows within warehouses.

At the receiving area products (or items) are unloaded and inspected to verify any quantity and quality inconsistency. Afterwards, items are transferred

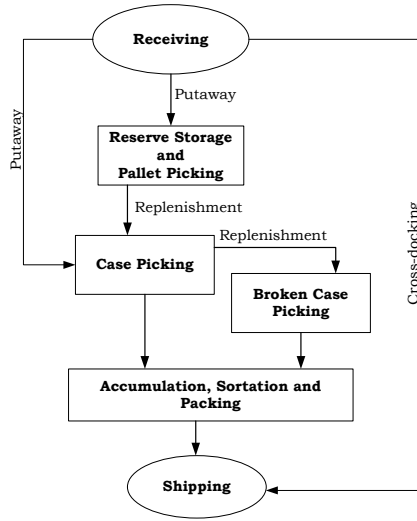


Fig. 3. Typical warehouse functions and flows

to a storage zone or are directly placed to the shipping area (this is called a cross-docking operation). We can distinguish two types of storage areas: reserve storage area and forward or picking area. The reserve area is the place where the products stay until they are required by customers' orders. The picking area is a relatively small area, typically used to store fast moving products. Most of the flows between these areas are the result of replenishment processes. Order picking is one of the most important functions in most warehouses. SKUs are retrieved from their storage positions based on customers' orders and moved to the accumulation and sorting area or directly to the shipment area. The picked units are then grouped by customers order, packaged and stacked on the right unit load and transferred to the shipping area.

2.1 Warehouse Design and Planning

Warehouse design can be defined as a structured approach of decision making at distinct decision levels in an attempt to meet a number of well-defined performance criteria. At each level, multiple decisions are interrelated and therefore it is necessary to cluster relevant problems that are to be solved simultaneously. According to Rouwenhorst et al. [10] a warehouse design problem is a "coherent cluster of decisions" and they define decisions to be coherent when a sequential optimization does not guarantee a globally optimal solution.

The design of a warehouse is a highly complex problem. It includes a large number of interrelated decisions involving warehouse processes, warehouse resources and the organization of the warehouse (see Heragu [8]). Rouwenhorst et al. [10]

classify the management decisions concerning warehousing into strategic decisions, tactical decisions and operational decisions. Strategic decisions are long term decisions and always mean high investments. The two main issues are concerned with the design of the process flow and with the selection of the types of the warehousing systems. Tactical management decisions are medium term decisions based on the outcomes of the strategic decisions. The tactical decisions have a lower impact than the strategic decisions, but still require some investments and should therefore not be reconsidered too often. At the operational level, processes have to be carried out within the constraints set by the strategic and tactical decisions made at the higher levels. In this level, the concern includes the operational policies such as storage policies and picking and routing operations.

After determining warehouse location and its size, layout decisions must include areas definition and what size should be allocated to each functional area. The forward-reserve problem (FRP) is the problem of assigning products to the functional areas. In this problem the critical decision concerns the choice of products that will be stored in the forward area. Van den Berg et al. [13] proposed a binary programming model to solve the FRP in the case of unit load replenishment, and presented efficient heuristics that provide tight performances guarantees. These replenishments can occur during busy or idle picking periods. The objective was to minimize the number of urgent or concurrent replenishments of the forward area during the busy periods. Although addressing this problem is a strategic decision problem, it is strongly associated upon some tactical problems such as how the items will be distributed among the functional areas. However, the approach usually adopted is to solve the problems sequentially by generating multiples alternatives for the functional area size problem and then determine how the products can be allocated for each of the alternatives.

Gray et al. [5] developed an integrated approach for the design and operation of a typical order-consolidation warehouse. This approach included warehouse layout, equipment and technology selection, item location, zoning, picker routing, pick generation list and order batching. Due to the complexity of the overall problem, they developed a multi-stage hierarchical decision approach. This hierarchical approach used a sequence of coordinated mathematical models to evaluate the major economic trade-offs and to reduce the decision space to a few number of alternatives. They also used simulation technique for validation and fine tuning of the resulting design and operating policies.

Heragu et al. [8] developed a mathematical model and a heuristic algorithm that jointly determines the size of the functional areas and the allocation of the product in a way that minimizes the total material handling and storage costs. The proposed model uses real data readily available to a warehouse manager and considers realistic constraints.

Gerales et al. [2] adapted the mixed-integer programming model proposed by Heragu et al. [8] to tackle the storage allocation and assignment problems during the redesign process of a Portuguese company warehouse.

More recently Strack and Pochet [11] presented a robust approach that integrates aspects such as: (i) the size of the functional areas; (ii) the assignment and allocation of products to storage locations in the warehouse; (iii) the replenishment decision in the inventory management. This is probably the most integrated decision model found in this area, nevertheless still assumes fixed and known capacity for the warehouse.

2.2 Inventory Decisions

The adoption of new management philosophies compels companies to eliminate or reduce inventory levels. In addition to warehouse management decisions, an appropriate inventory policy may result in a reduction of the total warehousing costs and can also improve the efficiency of the operating policies within the warehouse. The aim of inventory management is to minimize total operating costs satisfying customer service requirements (see Ghiani et al. [3]). To accomplish this, an optimal ordering policy must answer questions, for each SKU, such as when to order and how much to order.

Two different inventory policies arise (see Hadley and Whitin [6]): continuous review policy and the periodic review policy. The first policy implies that the stock level will be monitored continuously. Whenever the inventory on hand decreases to a predetermined level, referred to as the *reorder point*, a new order is placed to replenish the inventory level. The placed order is a fixed quantity that minimizes the total inventory costs and is normally called the *economic order quantity*. In the second policy, the inventory level is checked at specific fixed time periods and an order decision is made to complete the stock to a desired upper limit. In this system the inventory level is not monitored at all during the time interval between orders, so it has the advantage of little or no required record keeping. The disadvantage is less direct control. Such system also requires that a new order quantity must be determined each time a periodic order is made. The operating costs taken into account in both inventory policies are the acquisition cost, the holding cost and the shortage cost.

These basic policies can be adapted to take into account special situations such as single or multi-item models with or without a constraint on the total storage space, deterministic or stochastic demands, lost sales, etc. For more details and examples see Ghiani et al. [3] and Nahmias [9].

3 Mathematical Programming Model

Storage may be considered one of the major warehouse functions. Some fundamental decisions occur during the design and planning of a warehouse: (i) how much inventory should be kept in the warehouse for each SKU; (ii) how frequently and at what time should the inventory for a SKU be replenished, and (iii) what should be the size of the warehouse.

The first two decisions lead to the traditional inventory management problem and the last one is probably one of the most important aspect in designing a

warehouse. Once warehouse size is determined, it will act as a constraint that may last for a long period of time.

Generally there are two basic choices for warehousing. The first is to operate a owned (or private) warehouse and the second is to rent storage space from a public warehouse. Depending on which is least expensive a warehouse manager may use only one type or adopt a mixed strategy - both owned and rented storage space.

In this section we present a mathematical programming model that integrates the warehouse sizing problem and the inventory decisions. The mathematical model determines the optimal overall size for a warehouse in a way that minimizes the inventory and warehouse costs. In developing the model we also assume that external storage space is available to rent to allow the selection of the most adequate storage alternative: storage space ownership, rented warehouse space or a combined solution. Furthermore, an inventory control policy will be based on continuous review policy (reorder point system).

The problem can be stated as follow: a distribution company delivers a set of i products to its customers. Customers demand is well defined and unit acquisition, processing, storage and inventory carrying costs are known. Given warehousing costs (both own storage space and rented space) the objective is to determine the size and type (own or rented) of storage space required as well as the quantity of each product to order and respective reorder point, for a level of service and that minimizes total system costs.

3.1 A MINLP Model for the Warehouse Design Problem

The following notation, adapted from Strack and Pochet [11], is used:

Parameters:

i	: Product number ($i = 1, \dots, I$)
$CostCar$: Inventory carrying cost
$CostAcqu_i$: Acquisition cost of product i
$CostShort$: Shortage cost
$CostRecp$: Reception cost
$CostCapaS$: External capacity cost
$CostCapaW$: Capacity cost of the private warehouse
$CostCapaFW$: Private warehouse capacity fixed cost
$E(U_i)$: Expected value of the demand of product i
L	: Supply lead time
d_i^L	: Demand of product i during L
μ_i^L	: Average demand of product i during L
σ_i^L	: Standard deviation of demand of product i during L
M	: Large positive number

Decision variables:

$$w = \begin{cases} 1 & \text{if we have a private warehouse} \\ 0 & \text{otherwise} \end{cases}$$

$CapaW$: Capacity of the private warehouse

$CapaS$: External additional storage capacity

Q_i : Replenishment quantity of product i

r_i : Reorder point of product i

The general formulation of the model can be stated as:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^I Costcar \times \left(\frac{Q_i}{2} + r_i - \mu_i^L \right) + \sum_{i=1}^I CAcqui \times E(U_i) \\ & + \sum_{i=1}^I CostShort \times \frac{E(U_i)}{Q_i} \times \int_{r_i}^{\infty} (d_i^L - r_i) f(d_i^L) dd_i^L \\ & + \sum_{i=1}^I CostRecp \times \frac{E(U_i)}{Q_i} + CostCapaS \times CapaS \\ & + CapaW \times w \times (CostCapaW + CostCapaFW), \end{aligned} \tag{1}$$

subject to:

$$\sum_{i=1}^I (Q_i + r_i - \mu_i^L) \leq CapaW + CapaS, \tag{2}$$

$$CapaW \leq Mw, \tag{3}$$

$$CapaW \geq w, \tag{4}$$

$$Q_i, r_i \geq 0, \tag{5}$$

$$CapaW, CapaS \geq 0, \tag{6}$$

$$w \in \{0, 1\}. \tag{7}$$

The objective function (1) is the expected warehouse and inventory costs per period. Concerning the inventory costs we have taken into account: carrying cost, acquisition cost and shortage cost. The warehouse costs are composed by the reception cost, the additional external storage capacity cost and the costs of the private warehouse. The integrity of the model is ensured by the capacity constraint (2) that guaranties that the required storage capacity is met. Constraints (3)-(4) serve to include the costs of the private warehouse. Finally, a set of variables must be nonnegative (5)-(6) and another is considered binary (7). The above model considers inventory and warehouse size decisions since it integrates both issues supporting decision makers defining the best warehousing alternative, taking into account space requirements determined by customers' demand.

3.2 Methodology

The above model jointly integrates inventory decisions and size decisions (both ownership and rented storage space). It is a mixed-integer nonlinear programming model with a large number of variables when real cases are considered.

To evaluate the computational performance involved in solving the proposed integrated model, experimental tests were performed using LINGO 12.0 solver which uses branch-and-bound algorithm for integer nonlinear problems. All tests were performed on an Intel Core 2Duo 1.4 GHz CPU and 3GB RAM.

4 Computational Results

Instances for different scenarios were randomly generated to assess the behaviour of the model when the number of products increases (see Table 1). Table 2 shows parameter values used to generate the testing problems.

Table 1. Analysed scenarios

Scenario	I	II	III	IV	V
SKU [units]	10	100	500	1000	5000

Table 2. Parameter values for the numerical examples

Parameter	Value
$CostCar$	3
$CostShort$	50
$CostRecp$	5
$CostCapaS$	20
$CostCapaW$	3
$CostCapaFW$	10
$E(U_i)$	Uniform [1, 50]
d_i^L	$N(\mu_i^L, \sigma_i^L)$

The computational results for the different testing cases are shown in Table 3. As it can be seen it was possible to analytically solve to optimality all the test scenarios in a very satisfactory computational time. In general the algorithm is very efficient and converge to the optimal solution in a very short time. Nevertheless, the computational time of LINGO rises as the problem size increases.

Table 3. Model computational results

	Scenario				
	I	II	III	IV	V
Total variables	13	103	503	1003	5003
Nonlinear variables ^a	12	102	502	1002	5002
Iterations	203	647	1762	4020	16021
CPU time [mm : ss]	00 : 03	00 : 11	00 : 41	01 : 20	14 : 01
State	Global Opt.	Global Opt.	Global Opt.	Global Opt.	Global Opt.

^a Variables involved in the nonlinear relationships of the model.

In order to point out some more features of the proposed model, more experiments were conducted for a warehouse problem with 500 SKUs (scenario III). Again parameter values from Table 2 were used and service levels between 95% and 99% were considered.

Table 4 shows that a private warehouse can be less expensive than renting all the storage capacity, even considering higher amounts of stock. When the privately operated warehouse has an upper storage limit, the total optimal storage capacity equals the optimal capacity for the strategy of using only a public warehouse. This happens whenever the storage upper limit of the private warehouse is less than the optimal storage solution of the public warehouse.

Table 4. Warehouse dimension and cost

w	M	$CapaS$	$CapaW$	Total Cost
0	—	694	—	51060
	<i>unlimited</i>	0	842	45724
	300	394	300	48960
1	500	194	500	47560
	600	94	600	46860
	700	0	700	46149
	900	0	842	45724

Additional results also show that the inventory costs have a significant impact in warehouse management (see Table 5). As expected the integration of an appropriate inventory policy may result in a reduction of the total warehousing costs.

The gathered results also indicate that a mixed strategy (both owned and rented) would be better than renting all the storage space. Warehouse management should then achieve high levels of utilization for owned storage space and

Table 5. Warehouse total costs

w	M	Inventory Costs	Warehouse Costs	Total Costs
0	—	23908	27152	51060
	<i>unlimited</i>	23878	21846	45724
	300	23908	25052	48960
1	500	23908	23652	47560
	600	23908	22952	46860
	700	23903	22246	46149
	900	23878	21846	45724

use additional rented space on a short-term basis to meet peak space requirements.

As it can be seen in Table 6, different cost elements have different impacts in warehouse total cost.

Table 6. Warehouse inventory and operational costs

w	M	$CostCar$	$CostAqu$	$CostShort$	$CostRecp$
0	—	1056	21436	1416	13272
	<i>unlimited</i>	1279	21436	1163	10900
	300	1056	21436	1416	13272
1	500	1056	21436	1416	13272
	600	1056	21436	1416	13272
	700	1065	21436	1402	13146
	900	1279	21436	1163	10900

Since numerical tests were performed without having the real value of the different parameters involved in the model, and in order to highlight some relevant issues in the performance of the model, it is important to investigate how much influence some fluctuations in different cost elements may have in the optimal solution. For this purpose some of these costs will be selected and will vary (-10% and 25%) one at a time.

Denoting x and x^* as the initial parameter value and the new value, and $w(x)$ and $w(x^*)$ as the size of the storage capacity we will measure the impact on warehouse size of some variations using the following equation:

$$\frac{w(x) - w(x^*)}{w(x^*)} \times 100\%. \tag{8}$$

Table 7 summarizes the impact of cost fluctuations in the warehouse optimal size. A detailed look at the different cost changes reveals that the reception cost

(*CostRecp*) has a larger impact in warehouse size. As expected, an increase in this cost forces the model to increase the optimal order quantities and consequently the warehouse storage size increases. In a similar way decreasing the reception cost causes a decrease of the storage size area. On the other hand, the carrying cost (*CostCar*) and the shortage cost (*CostShort*) have lower impacts in the optimal solution with changes in warehouse size less than 1.2%.

Table 7. Impact in warehouse size (%)

Parameter	Variation	<i>CapaS</i>	<i>CapaW</i>
<i>CostCar</i>	-10	-0.71	-0.59
	25	1.01	1.20
<i>CostShort</i>	-10	0.58	0.48
	25	-0.99	-1.17
<i>CostRecp</i>	-10	4.83	4.73
	25	-9.51	-9.55
<i>CostCapaS</i>	-10	-4.67	—
	25	10.86	—
<i>CostCapaFW</i>	-10	—	-3.55
	25	—	8.08

The results also show that despite the inventory costs have a significant impact in the total warehouse costs, also variations in the warehouse costs mean significant impact in the storage size area. For example, the model is very sensitive to the external capacity cost (*CostCapaS*) where an increase of 25% causes a reduction of 10.86% of the rented area. In a similar way a 25% increase in the cost of the owned warehouse (*CostCapaFW*) causes an reduction of 8.08% of the size storage area.

The fluctuations that were carried out in some of the parameters of the model revealed that the size of the warehouse is more sensitive to changes of some parameters than to others. Nonetheless, the warehouse size solution is quite robust once the proposed model integrates the different costs in a way that balances the trade-offs among them, minimizing the effect of the change.

Finally it should be noted that in our numerical results the parameter variations were done one at a time. Future research might also be performed to investigate joint effects of different parameters on the quality of the solution of the model.

5 Conclusions and Future Work

Most of the times, inventory decisions and warehouse design decisions are independently considered. In fact, a single decision model that integrates several

decisions concerning warehouse design and planning is very complex due to the tremendous amount of information to be processed, to the large number of existing alternatives, to the existence of various and often conflicting objectives and to the uncertainty inherent in the material flow into, through and out of the warehouse.

Throughout this work our aim was to show the value of integrating inventory decisions and warehouse design decisions, in particular the strategic decision that concerns the size of a warehouse. For that purpose a mathematical programming model was proposed and discussed.

The proposed mathematical programming model jointly integrates: (i) the size of a private storage warehouse; (ii) the external additional storage capacity, if needed; and (iii) the replenishment quantities and reorder points of the products.

Although the model is a mixed-integer nonlinear programming model with a large number of variables, usually difficult to solve by general optimization packages, it was possible to solve to optimality some test scenarios in a very satisfactory computational time. Nevertheless for large instances the computational time increases considerably.

Computational results were obtained for five different scenarios randomly generated. The gathered results also suggest that, for this particular situation, a mixed strategy, both owned and rented storage capacity, would have lower total costs than renting all the storage space. High levels of utilization should be achieved for owned storage space and additional rented space should be used to meet peak space requirements. It has also been shown that inventory costs have a significant impact in warehouse system costs.

A sensitivity analysis was performed to observe the impact of cost fluctuations in the warehouse optimal size. For this purpose some costs were selected and varied one at a time. The results show that the size of the warehouse is more sensitive to changes of some parameters than to others.

Even though the presented model integrates two important decisions concerning the design of a warehouse, many other decisions were not included. For example the size of the functional areas inside the warehouse; the assignment and allocation problem of the products; the picking and routing strategies, etc.

In summary, despite some advances in integrated approaches, further research focusing integrated models where different processes in the warehouse are jointly considered (and its corresponding dynamic nature), is still required. Given the prevalence of warehouses in the supply chain networks we believe that such research achievements can have a significant impact in the supply chain performance.

References

1. Bartholdi, J., Hackman, S.: Warehouse & Distribution Science (2006). Release 0.76, <http://www.warehouse-science.com>
2. Gerales, C., Carvalho, S., Pereira, G.: A warehouse design decision model. In: Proceedings of the International Engineering Management Conference 2008, IEMC-Europe 2008, Estoril, Portugal (2008) ISBN: 978-1-4244-2289-0, IEEE Catalog Number: CFP08EMS

3. Ghiani, G., Laporte, G., Musmanno, R.: *Introduction to Logistics Systems Planning and Control*. John Wiley & Sons Ltd, England (2004)
4. Gu, J., Goetschalckx, M., McGinnis, L.F.: Research on warehouse operation: A comprehensive review. *European Journal of Operational Research* 177, 1–21 (2007)
5. Gray, A., Karmarkar, U., Seidman, A.: Design and operation of an order-consolidation warehouse: models and application. *European Journal of Operational Research* 58, 14–36 (1992)
6. Hadley, G., Whitin, T.: *Analysis of Inventory Systems*. Prentice-Hall, Englewood Cliffs (1963)
7. Hassan, M.: A framework for the design of warehouse layout. *Facilities* 20(13/14), 432–440 (2002)
8. Heragu, S., Du, L., Mantel, R.J., Schuur, P.C.: Mathematical model for warehouse design and product allocation. *International Journal of Production Research* 43(2), 432–440 (2005)
9. Nahmias, S.: *Production and Operation Analysis*, 3rd edn. McGraw-Hill International Editions (1997)
10. Rouwenhorst, B., Reuter, B., Stokrahm, V., van Houtum, G., Mantel, R., Zijm, W.: Warehouse design and control: Framework and literature review. *European Journal of Operational Research* 122, 515–533 (2000)
11. Strack, G., Pochet, Y.: An integrated model for warehouse design and planning. *European Journal of Operational Research* 204, 35–50 (2010)
12. Tompkins, J.A., White, J.A., Bozer, Y.A., Tanchoco, J.M.A.: *Facilities Planning*. John Wiley & Sons, New York (2003)
13. Van den Berg, J., Sharp, G., Gademann, A., Pochet, Y.: Forward-reserve allocation in a warehouse with unit-load replenishments. *European Journal of Operational Research* 111, 98–113 (1998)
14. Van den Berg, J., Zijm, W.: Models for warehouse management: Classification and examples. *International Journal of Production Economics* 59, 519–528 (1999)
15. Van den Berg, J.: A literature survey on planning and control of warehousing systems. *IEEE Transactions* 31, 751–762 (1999)

Application of CFD Tools to Optimize Natural Building Ventilation Design

José Carlos Teixeira¹, Ricardo Lomba¹, Senhorinha F.C.F. Teixeira²,
and Pedro Lobarinhas¹

¹Department of Mechanical Engineering, University of Minho, Guimarães, Portugal
{jt, rsl, pl}@dem.uminho.pt

²Department of Production Systems, University of Minho, Guimarães, Portugal
st@dps.uminho.pt

Abstract. The use of natural ventilation systems may contribute considerably to the reduction of the energy consumption, while providing adequate comfort levels and hygiene standards for the occupants. Computational Fluid Dynamics (CFD) techniques are becoming increasingly attractive in the design of ventilation systems. In this work, tests on a validated CFD model, which simulates the air flow inside a standard building, were carried out in order to obtain a suitable tool to predict ventilation performance and therefore optimize the building ventilation design. The model solves the mass, momentum and energy for the air flow, coupled with the k - ϵ turbulence model. The equations are solved by a FV discretization technique in a structured grid. Appropriated boundary conditions and the dimension of the domain were studied for more accuracy in numeric simulation. The influence of the free stream velocity profile and wind direction upon the efficiency of a natural ventilation system under isothermal conditions has been tested. The results obtained so far confirm the validity of the implemented model and its possible use for the optimal design of natural ventilation systems.

Keywords: Natural ventilation, Sustainable systems, Computational Fluid Dynamics (CFD).

1 Introduction

Increasing demands for thermal comfort and concerns regarding health and hygiene levels in confined spaces are driving the necessity for better ventilation systems. Ventilation is certainly a major tool in providing a desired level of user satisfaction. In order to regulate the indoor air parameters, it is essential to have suitable tools to predict ventilation performance in buildings [1].

The natural ventilation is an important and cost effective technique that, when properly implemented, contributes to the quality of indoor air as it decreases the pollutant concentration. In addition it reduces the death rate caused by respiration issues, frequently due to bad quality of indoor air. Finally, it also improves the thermal comfort of space and reduces the quantity of energy consumed by air conditioning systems.

However, natural ventilation is often neglected for other technologies such as mechanical ventilation, because it presents some difficulties, mostly due to the low velocity of the wind and inappropriate orientation of the ventilation openings to the wind direction. When natural ventilation is applicable to a specific architecture design, it usually requires a complex study of flow simulation to assure the air flow is appropriately distributed and regions of low interior air quality are absent. In turn, this increases the cost of the project. Traditional techniques for the design of ventilation systems are of difficult application in such conditions. Factors such as thermal gradients, wind velocity, aperture layout are paramount to their correct operation.

CFD models have become more and more popular in predicting ventilation performance [1]. The CFD technique numerically solves a set of partial differential equations for the conservation of mass, momentum, energy and other scalar quantities such as, turbulence intensities and species concentration. The solution provides detailed information on the spatial distribution of air velocity, pressure and temperature. Despite simplifications in the simulation studies, the results given by the CFD analysis are normally a realistic approximation of a real-life system. Its flexibility and cost effectiveness makes the CFD software a very powerful tool in the engineering research field [2].

Due to the increasing interest in using natural ventilation to reduce energy demand in buildings and to improve indoor air quality, the applications of CFD models for natural ventilation design are becoming popular. CFD three-dimensional modeling has been extensively explored in studying climatic and geometrical parameters in different fields [3-6]. El-Agouz [7] associated CFD two-dimensional air flow simulation with the energy equation, to study the effect of an internal heat source in a room.

Some authors used experimental data [8] and theoretical models [9] on the subject contributing to a deeper knowledge regarding natural ventilation, validating and giving credibility to CFD tool in the study of natural ventilation.

CFD models have been used to improve other building simulation tools so that ventilation performance can be accurately predicted. Wang and Chen [10] coupled CFD with a multizone airflow program and results from a building simulation method have been used as boundary conditions for a CFD 3D study [11]. These authors conclude that coupled simulation can better predict the indoor flow simulations. The coupling of multizone air models and thermal building simulation seems useful in the optimization of natural ventilation with buildings at the stage of the design and behavior of the occupants [12].

The creation of a robust optimization scheme aiming to assist office building designers is still a challenge where CFD techniques combine with optimization algorithms, such as, Genetic Algorithms [13] and Artificial Neural Network models [14].

The main objective of the present study is the test of a numerical CFD simulation of the air flow inside a standard building, implemented in the software Fluent [15]. The knowledge regarding the influence of the incidence angle of the wind and the free stream velocity profile of the wind upon the natural ventilation allows for the introduction of architectural details in the building at an early phase of the project.

This problem has not been properly addressed in the past. Therefore, this study is focused on the influence of the velocity profiles of the wind coupled with the angle of incidence of the wind upon the efficiency of natural ventilation systems for an isothermal situation. The ability of the CFD tool to handle these important aspects of the natural ventilation gives insight in obtaining an optimized design of natural building ventilation systems.

The paper is organized as follows. In section 2, the main equations of the CFD model are described as well as its numerical solution. Section 3 presents the geometry, mesh and the different boundary conditions test cases used for the simulations. The main results are presented and discussed in Section 4 and Section 5 presents the main conclusions and ideas for future work.

2 CFD Model

The air flow inside the building and its surroundings is solved as steady, incompressible, Newtonian and viscous turbulent. The governing equations of conservation of mass and momentum are solved with appropriate modeling procedures to describe the effects of turbulence fluctuations.

The model involves the simultaneous solution of the equations for mass and momentum conservation on a 3D framework. The main equations are now presented as well as, the numerical solution.

2.1 Mathematical Modeling

For any fluid property, its instantaneous value is obtained by adding a mean value with a fluctuating component. Taking velocity, this results

$$u_i = \bar{u}_i + u_i' \quad (1)$$

The conservation equations for turbulent flows are obtained from those for laminar flows using a time averaging procedure usually known as Reynolds averaging.

Mass conservation implies that the mass entering a control volume equals the mass flowing out, creating a balance between input and the output flows for a certain volume. This concept is mathematically expressed by Eq. 2, assuming constant the fluid properties[16-17]. Dropping the overbar on the mean velocity, \bar{u} , the ensemble-averaged mass conservation equation yields

$$\rho \frac{\partial u_i}{\partial x_i} = 0 \quad (2)$$

where ρ stands for density, x_i ($i=1, 2, 3$) or (x, y, z) are the three-dimensional Cartesian coordinates and u_i or (u_x, u_y, u_z) are the Cartesian components of the velocity vector u .

The momentum equations for the steady state turbulent flows are described by Eq. 3:

$$\frac{\partial}{\partial x_j}(\rho u_i u_j) = \frac{\partial}{\partial x_j} \left(\mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \left(\frac{2}{3} \mu \frac{\partial u_i}{\partial x_i} \right) \right) - \frac{\partial p}{\partial x_i} + \rho g_i + F_i + \frac{\partial}{\partial x_j}(\overline{\rho u_i u_j}) \quad (3)$$

where μ represents the viscosity, p represents static pressure, g represents gravitational acceleration and F the source term for the momentum equation. The effect of turbulence is incorporated through the "Reynolds stresses" terms, $\overline{\rho u_i u_j}$. These are related to mean flow quantities via a turbulence model. The most commonly used model for turbulence calculations is the k - ϵ successfully used in several fields of engineering. It is classified as a Reynolds-averaged Navier–Stokes (RANS) based turbulence model, considering the eddy viscosity as linear, it is a two equation model. This model accounts for the generation of turbulent kinetic energy (k) and for the turbulent dissipation of energy (ϵ). The model formulation described is typically called the Standard k - ϵ , and it is calculated by Eq. 4 and 5 [17].

$$\frac{\partial}{\partial x_i}(\rho k u_i) = \frac{\partial}{\partial x_j} \left[\left(\mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] + G_k - \rho \epsilon \quad (4)$$

$$\frac{\partial}{\partial x_i}(\rho \epsilon u_i) = \frac{\partial}{\partial x_j} \left[\left(\mu + \frac{\mu_t}{\sigma_\epsilon} \right) \frac{\partial \epsilon}{\partial x_j} \right] + C_{1\epsilon} \frac{\epsilon}{k} G_k - C_{2\epsilon} \rho \frac{\epsilon^2}{k} \quad (5)$$

In these equations, G_k represents the generation of turbulent kinetic energy due to the mean velocity gradients, σ_k and σ_ϵ are the turbulent Prandtl numbers for the k and ϵ , respectively. $C_{1\epsilon}$ and $C_{2\epsilon}$ are model constants. Turbulent viscosity, μ_t , is modelled according to the Eq. 6, by combining the values of k and ϵ .

$$\mu_t = \rho C_\mu \frac{k^2}{\epsilon} \quad (6)$$

This model uses, by default, the following values for the empirical constants: $C_{1\epsilon} = 1.44$, $C_{2\epsilon} = 1.92$, $C_\mu = 0.09$, $\sigma_k = 1.0$, $\sigma_\epsilon = 1.3$

Appropriate boundary conditions have to be assumed, in order to correctly define the computational domain.

2.2 Numerical Solution

FLUENT uses a control volume based technique to solve the conservation equations for mass, momentum, and turbulence quantities [16-17]. The domain is divided into discrete control volumes where the governing equations are integrated to obtain the algebraic equations for the unknowns (velocities, pressure and scalars). The

integration of the differential equations in each control volume results in a finite-difference equation that conserves each quantity on a control-volume basis.

Because FLUENT defines the discrete control volumes using a non-staggered storage scheme (all variables are stored at the control volume cell center), interpolation schemes are needed to determine the face values of the unknowns from the stored values at the cell center. The standard discretisation scheme was used for the pressure and the second order upwind scheme for the momentum, turbulent kinetic energy and turbulent dissipation rate equations.

The discretized equations are solved sequentially and the SIMPLE algorithm has always been used in the present application. This type of algorithm is based on using a relationship between velocity and pressure corrections in order to recast the continuity equation in terms of a pressure correction calculation. In this way, the calculated velocity and pressure fields satisfy the linearized momentum and continuity equations at any point.

FLUENT does not solve each equation at all points simultaneously and so an iterative solution procedure is used with iterations continuing until the convergence criterion specified has been achieved.

The algebraic equation for each variable is solved using a Line Gauss-Seidel procedure (LGS) and the user can specify the direction in which the lines are solved (the direction of the flow or alternate directions) and the number of times the lines are solved in order to update a given variable within each global iteration cycle. To speed up the convergence achieved by the LGS procedure, Fluent uses a Multigrid acceleration technique by default to solve the pressure and enthalpy equations.

The new calculated values of a given variable obtained in each iteration by the approximate solution of the finite difference equations are then updated with the previous values of the variable using a under relaxation technique. The user can choose the best relaxation factors for each variable in order to achieve a better convergence.

3 Simulation Test Cases

The objective of the present simulation is to highlight the performance of the CFD model [15] implemented, using commercially available CFD Fluent software. The building geometry and the surroundings are described as well as the mesh used in the numerical calculations.

In this research work, different configurations for the boundary conditions have been tested according to the intensity and angle of incidence of the wind.

3.1 Geometry and Grid

The basic geometry considers a building with parallelepiped shape, 6 m x 6 m x 3 m. In it, an opening (150 mm x 300 mm) in the front wall is located 150 mm above the floor level and placed horizontally in the middle of the facade. Another opening, in

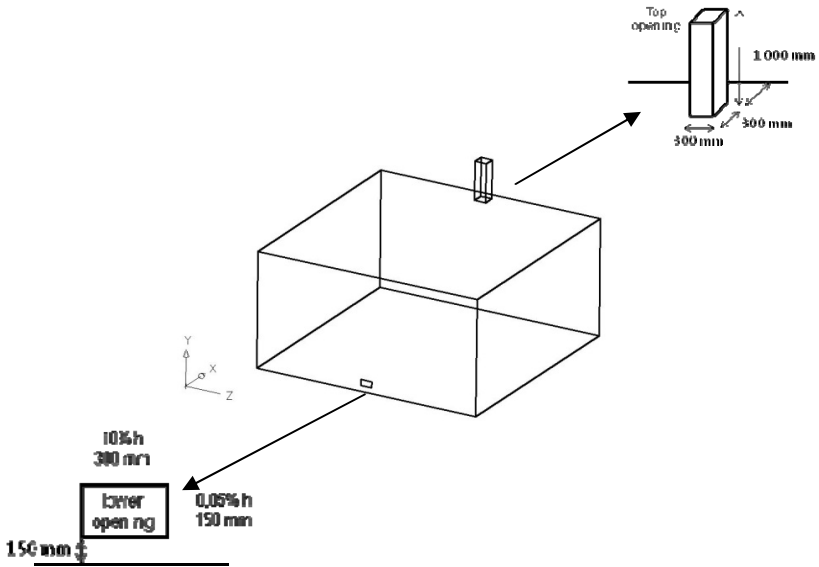


Fig. 1. Geometry of the test case [15]

the shape of a chimney, is located on the ceiling. It is 225 mm from the back wall and has a cross section of 300 mm x 300 mm. The chimney is 1 m high. Figure 1 illustrates the geometry described above.

As already discussed [15], because the inflow and outflow of the building result from the fluid flow interaction with the geometry of the building and surrounding area, a kind of “box-the-air” called domain, is created surrounding the building. This procedure is crucial to accurately simulate conditions of natural ventilation and in this way, the characteristics of the fluid velocity in the building openings are not locally imposed but calculated.

The appropriate dimension of the domain relatively to the building for which the airflow near the limits of the building is not influenced by its size has been carried out [15]. The task was to determine an appropriate size for the computational domain in order that the velocity profiles in the neighborhood of the building are not influenced by its dimension. It was concluded that a domain with overall dimensions of 30 m x 30 m x 9 m is appropriate for an accurate assessment of the influence of the building in the flow field in its vicinity.

Due to the huge dimension of the domain, it is not suitable to use a refined and regular mesh for whole domain. Therefore, in the more important regions the mesh should be more refined at the expense of other zones. The building is the most important zone, particularly around the openings and areas surrounding of building. The cell type was the same in all domain (Hexaedron–submap), with a total amount of approximately 2 million cells. The transitions between meshes with different sizes

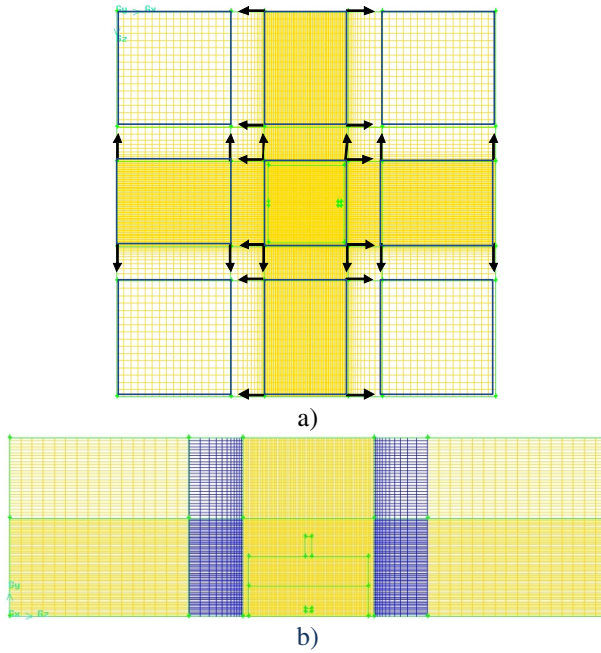


Fig. 2. Configuration of the mesh used: a) in top view b) lateral view [15]

were made by a boundary-layer type in which the size is expanded at a geometric ratio 1.219. So, cells have an average size of 75 mm in central area and are of 550 mm in remote areas. Figure 2 is illustrative of the mesh used.

3.2 Boundary Conditions

In the literature, different alternatives have been chosen for configuring the boundary conditions applicable in such cases. A comprehensive study has been made [15] in order to define the most appropriate boundary conditions (Figure 3).

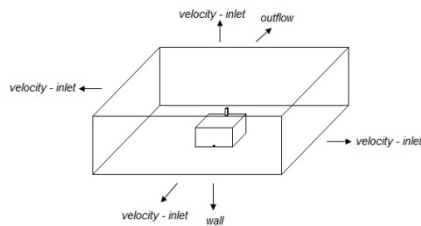


Fig. 3. Configuration of the boundary conditions for 0 angle of incidence of the wind

The frontal face (upstream) assumes a velocity inlet condition and the bottom a solid wall. For the 'free' surfaces of the control volume, a velocity-inlet condition appears to be the most appropriated and an outflow condition was defined on the back (downstream) surface.

The velocity profile representing the atmospheric boundary layer was defined. The velocity is normal to the frontal surface and tangential to the side walls. This profile is a function of the height (y , from the ground) and terrain roughness and is given by the equation [18]:

$$U(y) = U(y_1) \left(\frac{y}{y_1} \right)^\alpha \quad (7)$$

where $U(y)$ is the wind speed at height y (m/s); $U(y_1)$ is the wind speed at reference height (m/s); y_1 is the reference height (m) [typically 10 m]; y is the height (m) and α is the coefficient of the terrain roughness (here assumed as $\alpha=0.25$, for an housing area). On the top wall of the domain, the boundary condition was set as a tangential velocity in direction of the wind velocity, with a magnitude equal to that given by Eq. (7) for the corresponding elevation.

Figure 4 depicts the external domain and the corresponding boundary conditions, as previously described. In the center of Figure 4, the building and the ventilation openings are shown.

The volume between the building and the domain was defined as fluid, with the physical properties of air. The system was assumed with no thermal gradients being the temperature of the building walls and the air set at 20 °C.

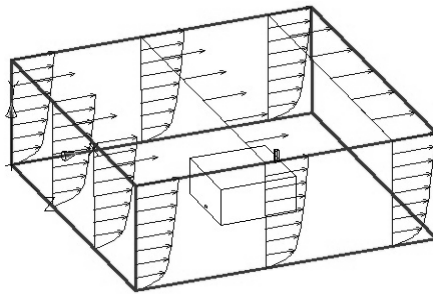


Fig. 4. Scheme of the velocity profiles applied [15]

The simulations were carried out for 3 different velocity profiles, defined by the free stream velocity of 1, 2 and 3 m/s referred at the reference height of 10 m. As far as the wind direction is concerned, 5 different angles of incidence of the wind were tested: 0°, 22.5°, 45°, 67.5° and 90°, as shown in Figure 5.

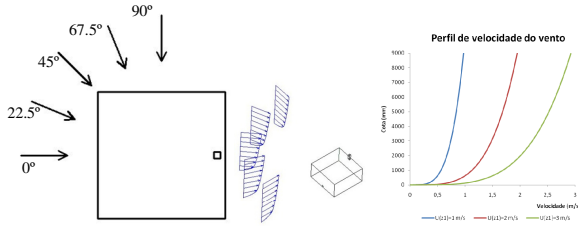


Fig. 5. Angles of incidence of the wind and profiles [15]

The boundary conditions applied in the limits of the domain were similar to those described above. However, its details were adjusted as a function of the angle of incidence of the wind. Figure 6 shows the configurations adopted, taking the reference case for an angle of incidence of 0° (Figure 3).

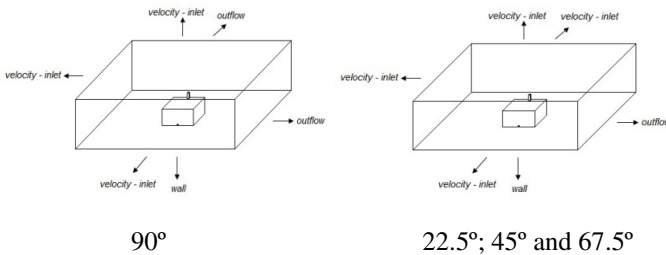


Fig. 6. Configuration of the boundary conditions for different angles of incidence of the wind

4 Numerical Results and Discussion

For all the test cases, an analysis of the flow pattern was made around the building on a horizontal plane at the level of the lower opening. Figure 7 shows the velocity field in the vicinity of (and in) the building for an incidence angle of 0°. The figure on the left hand side (at the opening high) shows that the opening provides an influx of fresh air into the building with a localized velocity that dissipates downstream through the inner space; in fact the velocity for most of the volume is close to zero. The maximum velocity inside the building at a height of 1.5 m is 0.038231m/s (for a free stream velocity of 3 m/s). This is well below the maximum acceptable by standards for building comfort defined by the construction codes of 0.2 m/s. In the back wall of the building two symmetrical recirculation vortices are formed (“3”; Figure 7-a). Similar patterns are observed along the leading edge of the side walls, in which the reattachment is delayed with increasing the air free stream velocity. At the leading edge of the top wall a recirculation vortex is also present (Figure 7-b).

Figure 8 shows the flow pattern for an incidence angle of 90° , which identifies a wind velocity parallel to that ventilation opening. In this situation, the flow is diverted around the building, creating a stagnation point in the center of the upstream wall (“1” in Figure 8). Flow separation occurs at the edge of the building though flow reattachment occurs well upstream of the ventilation opening as can be observed in Figure 8-c). However for a higher free stream velocity ($U(10)=3$ m/s) a strong separation occurs at the building edge (zone “2” in Figure 8-a) and flow reattachment is prevented into well downstream the side wall of the building. The extent of the recirculation zone can be observed in the detail represented in Figure 8-b) where a reverse flow is still evident downstream the side wall in the vicinity of the ventilation opening. This causes flow patterns that penetrates into the building itself as observed by comparing the right hand sides of Figures 8-b) and c); the flow direction inside the building is reversed.

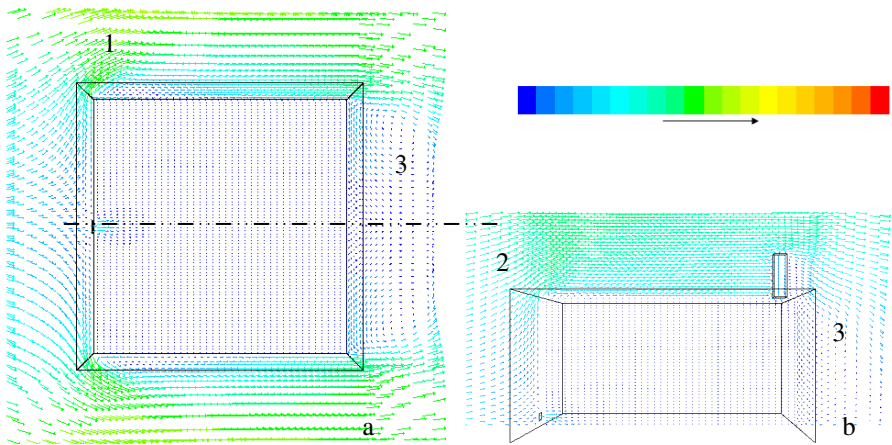


Fig. 7. - a) Distribution of velocity field in the horizontal plan passing through the lower opening for the profile $U(10) = 2$ m/s, b) in the central vertical plane. Incidence angle of 0° .

However, for the low intensity velocity profiles, the air entry into the lower opening is similar to all cases previously observed, i.e., in the direction of the main flow, as shown in Figure 8-c). It is also observed that even at 90° incidence some degree of ventilation is provided. Nonetheless the maximum velocity is approximately an order of magnitude below that observed for a 0° incidence angle (0.004428 m/s).

An alternative analysis consists in assessing the Pressure Coefficient - C_p (Figure 9). The data refers to an incidence angle of 0° . It is observed that the lower opening always acts as the inlet section for the building ventilation as it is located in the upstream wall, facing directly the wind. The local pressure coefficient for the 3 regions represented in Figure 9 follows the relationship $C_{p3} > C_{p2} > C_{p1}$. This behavior was observed for all angles of incidence of the wind and for all the velocity profiles tested. Also the location and extent of the recirculation areas (Figure 7) matches the location of negative pressure coefficients.

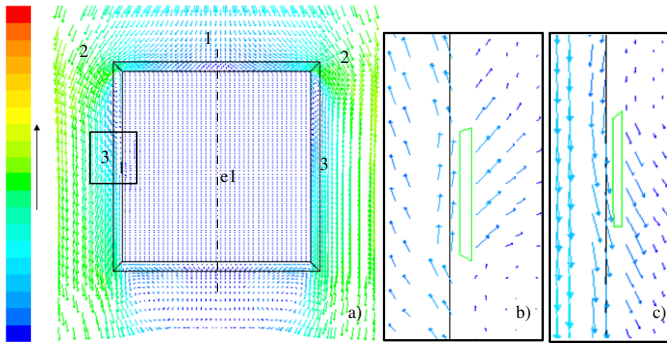


Fig. 8. - a) Distribution of velocity field in the horizontal plan passing through the lower opening for the profile $U(10)=2$ m/s, b) detail near the lower opening for the profile $U(10)=3$ m/s, c) detail near the lower opening for the profile $U(10)=2$ m/s

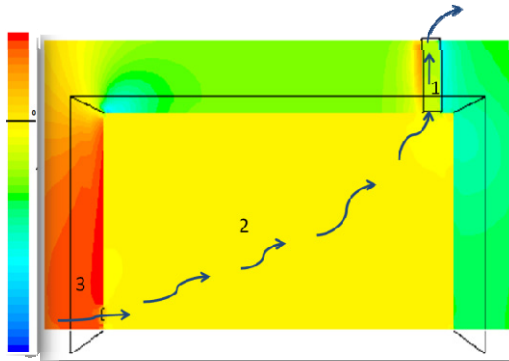


Fig. 9. Distribution of pressure field in the mid plane containing the openings (incidence angle of 0°)

Using for comparison the criterion "area-weighted average velocity magnitude", that for the free stream velocity of $U(10)=3$ m/s, the highest average velocity inside the building is observed (Figure 9). The data refers to the velocity field calculated at an elevation of 1.5 m. Nonetheless the actual average velocity is very low, below 0.0071 m/s. The profile for a free stream velocity of $U(10)=1$ m/s shows the lowest average velocity and a lower dependence upon the angle of incidence. In this case, the influence of incidence angle upon the lower velocity profile is virtually negligible. From the data represented in Figure 9, it is observed that by increasing the free stream velocity the maximum average velocity tends to occur for an incidence angle between 22.5° and 45° and not at 0° of incidence.

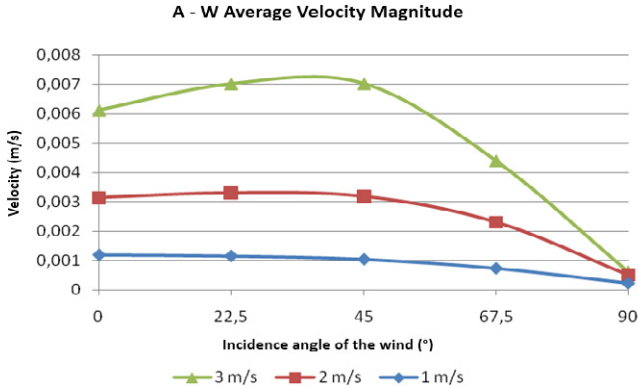


Fig. 10. Average velocity inside the building (1.5 m)

Figure 11 shows the variation of the pressure coefficient at the inlet opening with the incidence angle and the wind velocity. The data is in agreement with the average velocity inside the building. It shows that this parameter is less sensitive to the wind orientation with the decrease in the free stream velocity. The occurrence of negative coefficients for large incidence angles (particularly at high velocity) suggests that the flow is not fully reattached at that point.

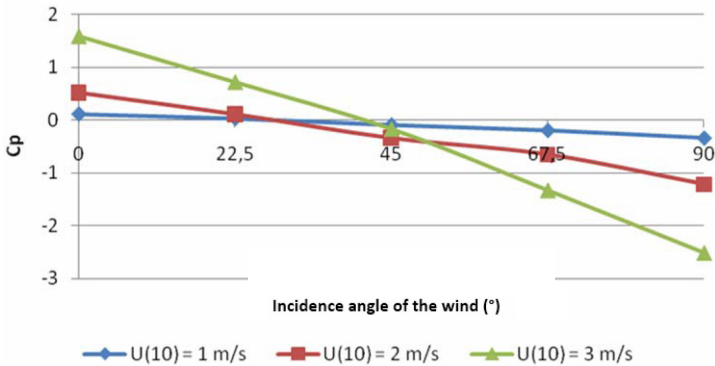


Fig. 11. Pressure coefficient at the inlet opening

One of the most important factors in evaluating the efficiency of a ventilation system (either natural or mechanical) is the rate of air renewal. By referring this as the number of volumes displaced per unit of time (hr^{-1}), the value of the air renewal per hour decreases with the increase of the incidence angle of the wind, as shown in Figure 12.

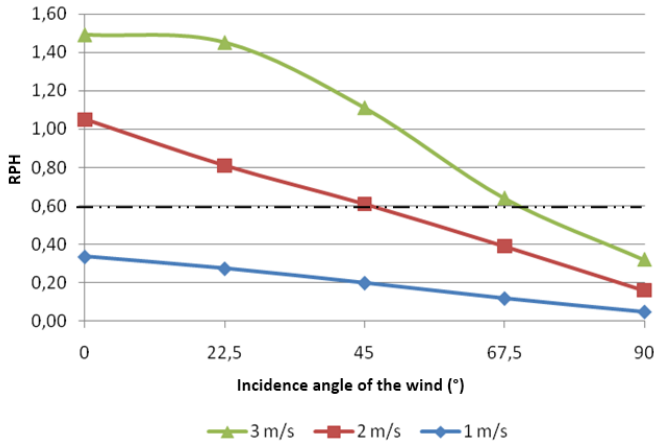


Fig. 12. Value of RPH according to the angle of incidence

The increase in the wind velocity is associated with an improvement in the air renewal, hence ventilation efficiency. The largest numbers of air renewal occur for an angle of incidence 0°, as opposed to the angle 90°. Bearing in mind the value of 0.6 hr⁻¹ [19] as a minimum for maintaining an indoor air quality, the case for U(10)=1 m/s does not promote ventilation at acceptable levels. It should also be referred that the reduction in the ventilation efficiency is small for angles below 45°. Changes in the openings layout and size would be required to keep the ventilation at acceptable levels. Another alternative would be to split the ventilation openings into 2 on adjacent walls in order to guarantee that at least one would always be oriented at a favorable angle.

Figure 13 shows the rate of air renewal as a function of the variation in the pressure coefficient between the entrance opening and the chimney. It is observed that all the data follows a single trend, independent of the air flow orientation.

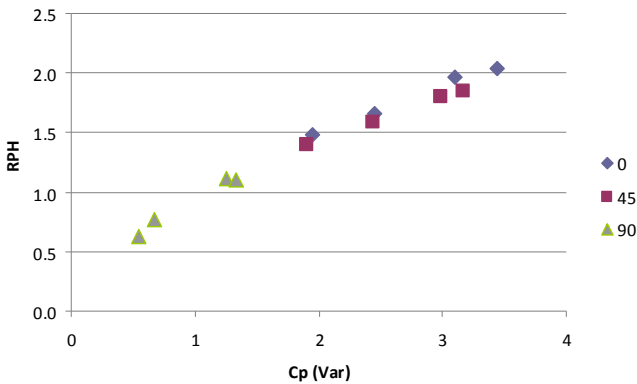


Fig. 13. Value of RPH according to the angle of incidence

5 Conclusions

This work reports the application of a CFD model to describe the natural ventilation inside a building under isothermal conditions. The computational domain was extended well outside the building in order to guarantee that the flow draft through the building is driven by the interaction of the atmospheric boundary layer with the building and its apertures.

From this work it can be concluded that CFD is a valuable tool for the flow analysis inside a building and therefore to assess the efficiency of natural ventilation systems. It was observed that the flow orientation relatively to the aperture is a major factor in providing an effective ventilation. Nonetheless, even for an orientation of 90° , some degree of ventilation is observed. The velocity inside the building is always below the comfort threshold level for indoor air quality. The rate of air renewal comply with the minimum of air renewals per hour for approximately 47% of all the cases tested. Particularly for a profile $U(10)=3$ m/s which shows a 80% efficiency.

The flow patterns in the vicinity of the outer walls of the building show the occurrence of recirculation vortexes over the back wall and on the leading edges on the roof and side walls. The location and size of such vortexes is well correlated with the distribution of the pressure coefficient along the surface of the building. For the highest wind velocity at an orientation of 90° , the recirculation vortex along the side extends into the ventilation aperture yielding an inlet velocity into the building with an orientation contrary to the main flow.

Given that the influence of the angle of incidence is reduced for values below 45° it is suggested that splitting the openings into adjacent walls may be effective to ensure an adequate ventilation, regardless of the wind orientation. The air ventilation rate was well correlated with the difference in pressure coefficients between the inlet and the outlet (chimney) apertures.

References

1. Chen, Q.: Ventilation performance prediction for buildings: a method overview and recent applications. *Building and Environment* 44, 848–858 (2009)
2. Ferziger, J.H., Peric, M.: Computational methods for fluid dynamics, 3rd edn. Springer (2003), <http://linkinghub.elsevier.com/retrieve/pii/S0898122103900460> (cited November 24, 2011)
3. Asfour, O.S., Gadi, M.B.: Using CFD to investigate ventilation characteristics of vaults as wind-inducing devices in buildings. *Applied Energy* 85(12), 1126–1140 (2008)
4. Lin, Z., Chow, T.T., Tsang, C.F., Fong, K.F., Chan, L.S., Shum, W.S., Tsai, L.: Effect of internal partitions on the performance of under floor air supply ventilation in a typical office environment. *Building and Environment* 44(3), 534–545 (2009)
5. Norton, T., Grant, J., Fallon, R., Sun, D.-W.: Optimising the ventilation configuration of naturally ventilated livestock buildings for improved indoor environmental homogeneity. *Building and Environment* 45(4), 983–995 (2010)

6. Bangalee, M.Z.I., Lin, S.Y., Miao, J.J.: Wind driven natural ventilation through multiple windows of a building: A computational approach. *Energy and Buildings* 45, 317–325 (2012)
7. El-Agouz, S.A.: The effect of internal heat source and opening locations on environment natural ventilation. *Energy and Buildings* 40(4), 409–418 (2008)
8. Su, X., Zhang, X., Gao, J.: Evaluation method of natural ventilation system based on thermal comfort in China. *Energy and Buildings* 41(1), 67–70 (2008)
9. Kaye, N.B., Ji, Y., Cook, M.J.: Numerical simulation of transient flow development in a naturally ventilated room. *Building and Environment* 44(5), 889–897 (2008)
10. Wang, L., Chen, Q.: Theoretical and numerical studies of coupling multizone and CFD models for building air distribution simulations. *Indoor Air* 17(5), 348–361 (2007)
11. Wang, L., Wong, N.H.: Coupled simulations for naturally ventilated rooms between building simulation (BS) and computational fluid dynamics (CFD) for better prediction of indoor thermal environment. *Building and Environment* 44(1), 95–112 (2008)
12. Oropeza-Perez, I., Østergaard, P.A., Remmen, A.: Ventilated buildings optimization by using a coupled thermal-airflow simulation program. In: *Proceedings of Building Simulation 2011, 12th Conference of International Building Performance Simulation Association*, Sydney, November 14–16, pp. 2666–2671 (2011)
13. Zhou, L., Haghghat, F.: Optimization of ventilation system designs and operation in office environment, Part I: Methodology. *Building and Environment* 44(4), 651–656 (2009)
14. Stavrakakisa, G.M., Zervasa, P.L., Sarimveisb, H., Markatosa, N.C.: Optimization of window-openings design for thermal comfort in naturally ventilated buildings. *Applied Mathematical Modeling* 36(1), 193–211 (2012)
15. Teixeira, J.C., Lomba, R.S., Lobarinhas, P.M., Seabra, E., Silva, L.F.: The Influence of Boundary Conditions on the Natural Ventilation in Buildings. In: *International Conference on Engineering Education ICEE 2010, Gliwice, Poland, July 18–22* (2010)
16. Versteeg, H.K., Malalasekera, W.: *An introduction to computational fluid dynamics: the finite volume method*. Longman, Harlow (1995)
17. ANSYS. ANSYS FLUENT Theory Guide (Internet). ANSYS Inc., Canonsburg (2009), <http://www.ansys.com>
18. Tantasavasdi, C., Srebric, J., Chen, Q.: Natural ventilation design for houses in Thailand. *Energy and Buildings* 33(8), 815–824 (2001)
19. RCCTE, Regulamento das Características de Comportamento Térmico dos Edifícios, DL 80/2006 de 4 de Abril, Diário da República (67), I Série – A, pp. 2468–2513 (2006)

Middleware Integration for Ubiquitous Sensor Networks in Agriculture*

Junghoon Lee¹, Gyung-Leen Park^{1,**}, Min-Jae Kang²,
Ho-Young Kwak³, Sang Joon Lee³, and Jikwang Han⁴

¹ Dept. of Computer Science and Statistics

² Dept. of Electronic Engineering

³ Dept. of Computer Engineering

⁴ Jeju National University, 690-756, Jeju-Do, Republic of Korea

Jinwoo Soft Innovation, Jeju-Do, Republic of Korea

{jhlee,glpark,minjk,kwak,sjlee}@jejunu.ac.kr, hmurdoc@jinwoosi.co.kr

Abstract. This paper first presents our framework architecture of ubiquitous sensor networks for agricultural and livestock farms, and then designs integrative middleware capable of efficiently managing the interaction between sensors and sensor applications. For the sensor network platform now under integration test, a block-level error recovery scheme is proposed to cope with the burst error pattern in wireless communication channels. The server-side middleware sends per-block acknowledgment packets, while sensors summarize the unacknowledged block and then transmit the summary when the connection gets reachable again. The block size and other summarization parameters are decided on the stream establishment. For operator control supports, two communication paths are established, one between remote PCs and embedded control boxes via TCP/IP connections, and the other between mobile terminals and smart motes via cellular network connections. Our middleware design can contribute to improving the reliability and correctness of sensor data processing by systematically coordinating interactions between all components involved in ubiquitous sensor networks.

Keywords: Ubiquitous sensor network, integrative middleware, interaction coordination, block-level error control.

1 Introduction

Each sensor network employs different sensors and differently deals with collected data according to the given system goal, even though its common operation is to collect and analyze sensor data, finally deciding an appropriate control action [1]. The control action may include storing and sophisticated mining of large volume of sensor data stream [2]. It is natural for each sensor network to be specialized according to its application area, reliability level, data scalability and

* Corresponding author.

** This research was supported by the MKE (The Ministry of Knowledge Economy), through the project of Region technical renovation, Republic of Korea.

the like, but sensor networks belonging to the same category working in the similar environment can have a common framework. Here, the small difference can be efficiently overcome by integrative middleware which coordinates interactions between sensor nodes and data handlers. This common framework can be deployed without redesigning the whole components such as data exchange protocols, message formats, and sensor data analyzers. It can even invite a new application to the underlying sensor networks [3].

Our project team has been researching and developing an integrative USN (Ubiquitous Sensor Network) for agricultural and livestock farms as well as fisheries, which are major industries in our province, namely, Jeju island, Republic of Korea [4]. Our USN, currently under development and testing, continuously monitors the environmental change in temperature, humidity, lightness, wind speed, CO₂, and NH₂ levels, while biosensors which captures the disease of a livestock. In this design, a sensor network can be built by selectively integrating those components required by an application type which can be characterized by involved sensors, coverage area such as a village or farms, and control reactions. This system pursues the following 5 development goals and now is in the middle of the final integration test.

1. composite sensor device and relevant remote control module
2. USN middleware for farms, fisheries, and livestock farms
3. intelligent control technology for farming environment
4. smart farming management information system
5. testbed construction and field test

In the phase of the integration test, the availability of sound sensor data is much more important for the correct operation of sensor applications than expected. For the integrative design of USNs, the middleware plays a key role of managing sensors and activating software agents. In this regard, this paper designs USN middleware including a block-level data recovery mechanism which attempts to minimize the effect of message loss by summarizing each message block. In addition, control paths and corresponding user interfaces will be developed for remote sensor network access. This middleware-supported framework can efficiently host a new sensor network application.

This paper is organized as follows: After outlining the problem in Section 1, Section 2 shows the sensor network architecture we have built. Section 3 designs sensor network middleware focusing on how to process the stream data. Section 4 implements the communication path for remote control. Finally, Section 5 concludes this paper with a brief introduction of future work.

2 System Architecture

Under the research and technical project named *Development of convergence techniques for agriculture, fisheries, and livestock industries based on the ubiquitous sensor networks*, a common sensor network framework has been built taking advantage of intelligent information technologies [5]. This framework provides an

efficient and seamless runtime environment for a variety of monitor-and-control applications on ubiquitous sensor networks. As shown in Figure 1, the sensor network field consists of diverse sensors exchanging messages by the IEEE 802.15.4 Zigbee protocol [6]. Through the USN gateway, sensor data messages are forwarded to a stream manager, which implements a variety of rules for detecting events of interest. The stream manager also determines the lifetime of each stream according to the requirement given by an application which can be from monitor-and-control agents to sophisticated cultivation planners [7]. The stream data will be dispatched to the appropriate application which is registered in the manager in advance. This architecture can integrate new sensors, rules, and applications.

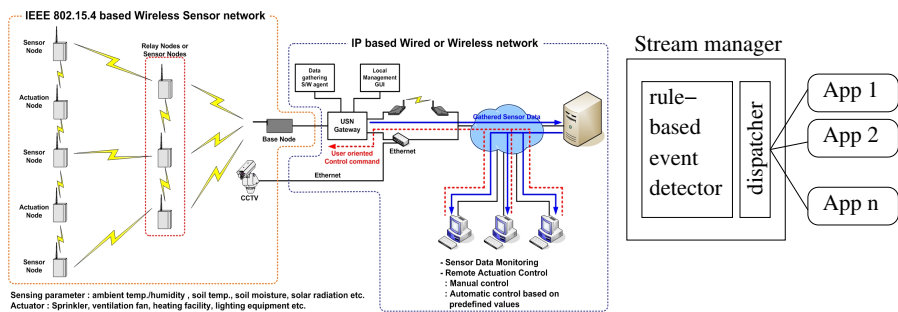


Fig. 1. Ubiquitous sensor network framework design

For a sensor network, our system implements a sensor node capable of reliably collecting sensor data mainly and delivering control actions to actuators in the agricultural sensor network [8]. Basically, the sensor nodes are installed at fixed location in agricultural sensor networks, so a sensor board can provide a stable power to them. A composite sensor board is consist of a main control unit, a sensor interface base board, an actuator interface & control board, a global network interface module, and a sensor signal acquisition board. It extends the data transmission range by regulating power consumption according to the device operation type. Here, the lightweight 6LowPAN and IPv6 protocol is extended, while a group of sensor devices can share pin connections. The sensor board also implements a program logic which converts raw sensor values to human-friendly ones. For example, the humidity sensor just captures the voltage level and the sensor board maps it to the 0 ~ 100 % humidity range. In the same way, anometer sensor values are converted to wind speed and direction.

3 Middleware Design

3.1 Sensor Stream Data Path

The main targets of our sensor network framework are listed in Table 1. Greenhouses, water tanks, and cattle sheds are the places where our USN will be

prospectively installed [9]. Each USN can work autonomously, but multiple USNs can be merged into a single management domain. Common operations and components for their sensor networks can be defined as each category has many common features in terms of sensors and cultivation management. For example, temperature monitoring and control is required in all of three places. Even though the water tank and the cattle shed bring up different species, they have the same cultivation and environmental management targets.

Table 1. Target network classification

	Crops	Cultivation	Requirement
greenhouse	flowering plant,	fruit tree type	temperature, illumination,
	fruit-vegetable,	species feature	sunshine, gas,
	fruit tree	cultivation management	water, soil
water tank	halibut,	species feature	temperature, oxygen,
	porgy,	cultivation management	water pollution, salinity,
	tuna	environment management	water level
cattle shed	pig,	species feature	temperature, illumination,
	chicken	cultivation management	gas, wind speed,
		environment management	humidity

In addition, Figure 2 shows the inter-module communication diagram for sensor data interfaces. Sensor nodes are connected to the gateway through the Zigbee network path. Embedded control boxes connect to the gateway using RS-232C serial communication on 115,000 *bps*. Moreover, u-MultiSensorMote, which will be explained later, establishes a connection to mobile phones using CDMA 2000 1X on necessity basis. For better event delivery on this communication path, our middleware classifies the message priority into 5 levels. Each message container includes this priority value along with sensor network id, sensor node id, and sensing type specification.

3.2 Middleware Support

Systematic middleware coordinates the interaction between the application layer and low-level sensors. For the sake of analyzing a great volume of sensor data, it mainly filters and synthesizes sensor streams [10], creating value-added context information. Then, an agent-based architecture distributes real-time data, forwarding a specific event to the appropriate application, which is registered in the directory service via the open interface. The middleware cooperatively works as service, server-side, and in-network agents. The service middleware agent includes a service discovery module, application agents, and a middleware query handler. The server-side middleware mainly manages the sensor stream, activating and relinquishing the sensor data streams according to the order given by applications through the service middleware [11].

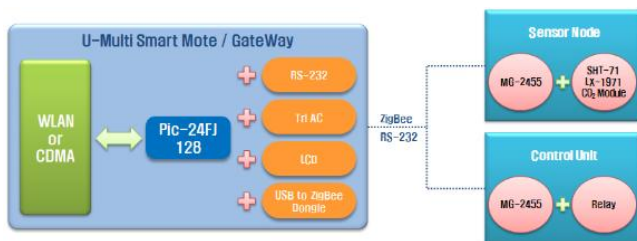


Fig. 2. Sensor stream data path

The in-network middleware, running in sensor nodes, provides a common abstract interface to the sensor hardware [12]. By means of this interface and meta-data specification, the server-side middleware can manipulate the sensor node operation. The stream manager collects the data from sensor nodes during the lifetime of a sensor stream. It continuously tracks sensor readings and detects a change in them. The queries can be classified by server-side, snapshot, continuous, event, and event snapshot queries. The sensor stream is an isochronous (or synchronous) traffic, consisting of message streams that are generated by their sources on a continuing basis and sent to a stream handler also on a continuing basis [13]. Sensor values are likely to contain error due to many reasons such as temporary board power instability, sensor node fault, and network disconnection. Hence, it is necessary to first filter false sensor readings, namely, check whether the sensor value is in the reasonable range. Additionally, if the current value is too much different from the previous one, it is considered to be invalid. Only the valid readings are stored in the database for further processing.

The wireless network is not so stable compared with wired counterpart [14]. Sometimes, sensor nodes can be isolated from the network and their data cannot be sent to the server-side middleware. Moreover, if a critical event such as fire occurs, sensors around the spot detect the status change and attempt to report simultaneously, resulting in a temporal traffic jam in the gateway node. Some messages are lost and the middleware can't correctly grab what is really going on. Such network instability is unavoidable in wireless sensor networks. In our field test for wireless communication, the sensor board operation is significantly affected by the weather condition and likely to work poorly when it is windy and rainy. In this case, the middleware can request data retransmission to the sensor node. However, the legacy per-message error recovery is impractical between sensors and the middleware, as it incurs so many control packets such as acknowledgment, retransmission, and the like. In the wireless environment, the error pattern is bursty. If a channel hopping mechanism is not exploited in the data link layer, a block of messages is almost completely lost [15].

Accordingly, our system designs a block-level error recovery scheme. The middleware sends acknowledgement messages back to the sensor node when it receives a block of messages. The block size, namely, the number of messages in a block, depends on the sensor type and the sensor environment as previously shown in Table 1. The middleware can know whether the block is well received

by the sequence number and its timer based on the report period of the sensor stream. It can send an acknowledgement message even if just a few messages are lost and how many message losses it can tolerate is also a tunable parameter. The sensor node keeps the report messages until the block is acknowledged, so the sensor buffer will overflow for a long breakage, which is not uncommon in wireless networks. If not, it summarizes the block of messages, before discarding all of them, as illustrated in Figure 3. Each message is embraced by STX (Start Transmission) and ETX (End Transmission) fields, while its payload length is less than 20 bytes.

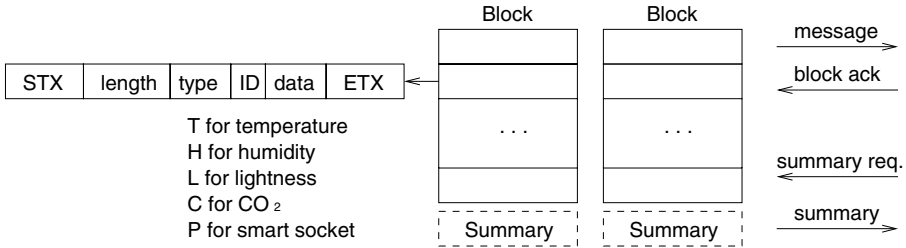


Fig. 3. Ubiquitous sensor network framework design

We define several options a sensor node can take for the summary of each block, while the middleware selects a specific option when the connection is established. The options include average, max, min, last, and counter. The first four are self-explanatory, while the counter option counts the number of sensor readings which meet the specific condition in the message block. For example, it can count the number of sensor records higher than 5 mph in a wind speed sensor. Moreover, another options agreed between the sensor board and the middleware can be integrated easily. For example, a pattern can be specified and captured in the sensor board processor. In this way, each unacknowledged block is reduced to a couple of bytes and sent to the middleware when the network becomes reachable again on the retransmission request from the middleware. Even if the middleware agent cannot recover the whole lost messages, it can keep monitoring and conducting stream-level analysis.

3.3 Data Analysis

The series of event patterns and interpreted knowledge are embedded in the analysis module to recognize conditions of interest instantly. Moreover, our system opens an interface to define the relationship between the environmental aspect and facility control equipments. Through this interface, we can set both the trigger condition of control actions and new event detection logic [12]. The location of each sensor is registered in the spatial database. The location record is specified by a WGS coordinate which consists of latitude and longitude. Along with the digital map representing the geographic information, we can conduct

a spatial analysis on the relationship between two or more sensor values and on terrestrial effects to sensor readings. A sensor stream is established whenever an application wants to monitor the sensor data. For such continuous query processing, the period of sensor reports can be adjusted according to the application requirement in our design. After all, even if the stream message contains neither a time stamp nor a location tag, the sensor application can know the spatio-temporal information.

Based on the long-term history data, it is possible to find an optimal interaction model between agricultural facilities in a physical space and ubiquitous computing environments in the cyber space. We can identify the critical environmental change to keep the best condition for the involving facilities. For efficient data analysis and comparison, our framework digitizes the sensor values and clusters into a fixed number of groups in this stage. Then diverse pattern matching utilities are being developed to support even complex queries.

4 Remote Control

For the remote control of the controller devices, our middleware implements two communication paths and corresponding user interfaces. The first is the communication between a remote PC and smart sockets. Here, an embedded control box intermediates their interactions. An operator can access the sensor network information either on the remote PC or by direct control of an embedded control box. The operator issues a series of commands, while the smart socket responds to them following the request-and-response transaction semantic. In addition, smart sockets can notify a specific event, for example, the bound condition violation for a sensing value, automatic alert, and power disconnection or breakage. Such information will be sent to the operator without a specific status request command. This control interface exchanges messages via TCP/IP protocol, while messages are converted to the Zigbee format in the sensor network area. The message format is already shown in Figure 3.

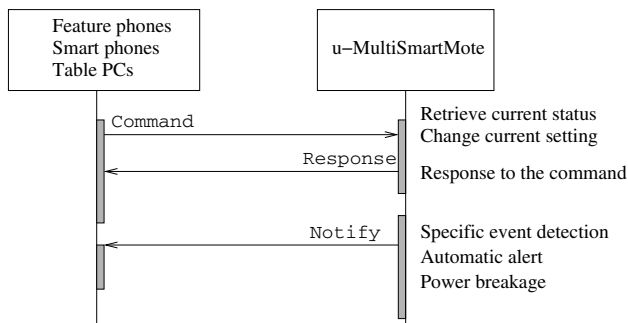
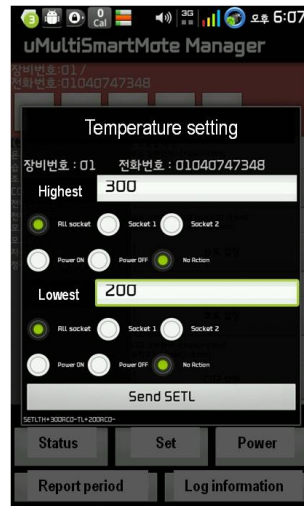


Fig. 4. U-MultiSmartMote interaction

The next communication path can be established between a mobile terminal and smart notes. We call it u-MultiSmartMote. The mobile terminal includes feature phones, smart phones, and tablet PCs. The interaction is functionally the same as the first control interface except that its communication path is built upon SMS (Short Message Service) on wireless telephony network instead of TCP/IP as shown in Figure 4. Figure 5 shows its interface implementation for smart phones. To begin with, the bottom menu includes the current status retrieval, operation parameter retrieval, power mode selection, report period setting, and log information display. In the middle of the window, there are 4 menu buttons for temperature, humidity, lightness, and CO₂ level, respectively. Each of them provides the submenu to specify the permissible range of sensor readings. We can set the upper and lower bounds for the individual sensors and if a sensor value gets out of this range, the sensor controller will notify to the smart mote manger program. For this interface, the remote controller must register its phone number.



(a) Main menu



(b) Temperature setting

Fig. 5. Smart phone user interface

5 Concluding Remarks

In this paper, we have presented a design of USN middleware which cooperatively works in server-side, service, and in-network agents to coordinate interactions between sensor and sensor applications. The requirement in agricultural sensor network has been defined according to crops and cultivation types for greenhouses, water tanks, and cattle sheds. For better accuracy and reliability of sensor applications, our middleware designs a block-level error recovery scheme to cope with the burst pattern in wireless communication channels. It is

based on the block summary specification, making the stream operation less affected by the lost data. For the efficient management of sensor networks, two communication paths are established along with corresponding interfaces. The first one is between remote PCs and embedded control boxes, while the second one between mobile terminals and smart motes. Each interface follows the request-and-response semantic while notification can take place when the specific condition is met for temperature, humidity, lightness, and CO₂ level.

This framework can be commonly applied to agricultural farms, livestock farms, and fisheries, avoiding the whole system redesign. At this stage, we are now in the course of extensive integration test of our sensor network, which consists of composite sensor board design, middleware development, intelligent stream analyzer, and farming management information system. After all, our sensor network is customized for small or medium scale farms and fisheries. The customization can make and develop our new business model.

References

1. Culler, D., Estrin, D., Srivastava, M.: Overview of Sensor Networks. *IEEE Computer* 37, 41–49 (2004)
2. Esposito, F., Basile, T.M.A., Di Mauro, N., Ferilli, S.: A Relational Approach to Sensor Network Data Mining. In: Soro, A., Vargiu, E., Armano, G., Paddeu, G. (eds.) *Information Retrieval and Mining in Distributed Environments*. SCI, vol. 324, pp. 163–181. Springer, Heidelberg (2010)
3. Sigrimis, N., Antsaklis, P., Groumpos, P.: Advances in Control of Agriculture and the Environment. *IEEE Control Systems*, 8–12 (2011)
4. Lee, J., Kim, H., Park, G., Kwak, H., Kim, C.: Intelligent Ubiquitous Sensor Network for Agricultural and Livestock Farms. In: Xiang, Y., Cuzzocrea, A., Hobbs, M., Zhou, W. (eds.) *ICA3PP 2011, Part II*. LNCS, vol. 7017, pp. 196–204. Springer, Heidelberg (2011)
5. Lee, J., Park, G., Kim, H., Kim, C., Kwak, H., Lee, S., Lee, S.: Intelligent Management Message Routing in Ubiquitous Sensor Networks. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) *ICCCI 2011, Part I*. LNCS, vol. 6922, pp. 537–545. Springer, Heidelberg (2011)
6. Gislason, D.: *ZIGBEE Wireless Networking*, Newnes (2008)
7. Madden, S., Franklin, M., Hellerstein, J., Hong, W.: The Design of an Acquisitional Query Processor for Sensor Networks. In: *ACM SIGMOD* (2003)
8. Lee, J., Park, G., Kim, H., Kwak, H., Lee, S., Lee, J., Kang, B., Kim, Y.: Design of a Composite Sensor Node in Agricultural Ubiquitous Sensor Networks. In: Kim, T.-H., Adeli, H., Fang, W.-C., Vasilakos, T., Stoica, A., Patrikakis, C.Z., Zhao, G., Villalba, J.G., Xiao, Y. (eds.) *FGCN 2011, Part I*. CCIS, vol. 265, pp. 53–58. Springer, Heidelberg (2011)
9. Revenaz, A., Ruggery, M., Martelli, M.: Wireless Communication Protocol for Agricultural Machines Synchronization and Fleet Management. In: *International Symposium on Industrial Electronics*, pp. 3498–3504 (2010)
10. Woo, H., Mok, A.K.: Real-Time Monitoring of Uncertain Data Streams Using Probabilistic Similarity. In: *Proc. of IEEE Real-Time Systems Symposium*, pp. 288–300 (2007)

11. Yingyou, W., Zhi, L., Xuena, P., Hong, Z.: A Middleware Architecture for Sensor Networks Applied to Industry Solutions of Internet of Things. In: International Conference on Intelligent Control and Automation, pp. 50–54 (2011)
12. Lee, J., Park, G., Kwak, H., Kim, C.: Efficient and Extensible Data Processing Framework in Ubiquitous Sensor Networks. In: International Conference on Intelligent Control Systems Engineering, pp. 324–327 (2011)
13. Golab, L., Oszu, M.: Issues in Data Stream Management. *ACM SIGMOD Record* 32, 5–14 (2003)
14. Ruiz, L., Nogueira, J., Loureiro, A.: MANNA: A Management Architecture for Wireless Sensor Networks. *IEEE Communication Magazine* 41, 116–125 (2003)
15. Song, S., Han, S., Mok, A., Chen, D., Nixon, M., Lucas, M., Pratt, W.: WirelessHART: Applying wireless technology in real-time industrial process control. In: IEEE Real-Time and Embedded Technology and Applications Symposium, pp. 377–386 (2008)

Usage Pattern-Based Prefetching: Quick Application Launch on Mobile Devices

Hokwon Song^{1,2}, Changwoo Min^{1,2}, Jeehong Kim², and Young Ik Eom²

¹ Samsung Electronics Co., Ltd., Suwon, Korea

² School of Information and Communication Engineering

Sungkyunkwan University, Suwon, Korea

{hokwon,multics69,jjilong,yieom}@ece.skku.ac.kr

Abstract. The startup time of applications is very important as a user perspective performance. If page faults occur frequently in the startup time, the user experience is subjected to an adverse effect. To reduce page faults, the prefetching scheme is used in the traditional OS. Previous studies proposed various schemes, but the most research was conducted for desktop PCs or special embedded devices. We propose the usage pattern-based prefetching scheme which is suitable to mobile devices. Therefore, this paper focuses on the user's applications usage patterns and the improvement of the startup time of application on mobile devices. To inspect the usage patterns, we collect the dataset of the application usage and then analyze collected data. Additionally, considering mobile devices which have relatively poor hardware resources, the lightweight prediction model is employed in the new scheme. The proposed scheme is implemented on both Android 2.2 and Linux kernel 2.6.29. It is tested on the emulator and evaluated by using the dataset. The startup time is improved about 5%, and the accuracy of the prediction is shown up to 59% for the practical dataset.

Keywords: Prefetching, Usage pattern, Mobile device.

1 Introduction

Recently, the embedded system has extended the coverage to the land of desktop PCs on the rapid development of hardware and software. Gradually, the change made users interested in the performance of mobile devices. The performance can be divided into the system performance, like CPU speed, memory size and display resolution, and the user perspective performance such as the startup time, the impression of a color and the user experience.

We focus on the improvement of the application's startup time. Although the startup time is one of the important user perspective performance factors, and the performance of hardware has evolved considerably compared to the previous, the startup time is still unsatisfying [11]. The main cause of the startup problem is the file IO [9, 11]. Since the file IO is much slower than CPU and main memory [7], and a task should be waiting until a page fault handler completes loading memory [8].

Obviously, the startup time increases when the access to secondary storage to load pages occurs frequently. Therefore, the startup time will be dramatically improved if a page fault does not occur. In the OS field, the prefetching scheme is a traditional solution for reducing page faults [1, 2, 3, 10]. The number of page faults is decreased by loading pages in advance of launching the application.

This paper proposes the Usage Pattern-based Prefetching scheme for mobile devices called UPP. UPP predicts the next application based on the user's application usage patterns and fetches the memory pages of a predicted application in advance. Our study introduces the lightweight prediction model and the special triggering time for mobile devices. Mobile devices have more scanty resources than desktop PCs and a different lifetime of applications.

UPP is implemented on both Android 2.2 and Linux kernel 2.6.29 and evaluated by testing on the emulator.

This paper makes the following contributions:

- Observation of the application usage patterns between each application by using the practical dataset
- Suggestion of the lightweight prediction scheme
- Development of the Usage Pattern-based Prefetching called UPP
- Implementation and evaluation of UPP.

The paper is organized as follows. In Section 2, we review other related work and discuss their efforts from the point of mobile devices. Section 3 analyzes the collected workload to understand the application usage patterns of users. In Section 4, we describe how to implement UPP, and provide the detail experimental setup and evaluate the experimental results. Finally, we conclude the paper and comment our future work in Section 5.

2 Related Work

2.1 Prefetching Scheme

In Linux and Windows, the prediction-based prefetching scheme is adopted for desktop PCs. The representatives are Preload [9] and SuperFetch [3]. They load the file-backed pages of an application which is expected to be executed in the near future. Thus, the prefetcher should monitor and analyze the user's access patterns. In case of [9], the prefetcher runs periodically to gather data and prefetch fault pages. Markov's probability model is employed to predict the next application. Additionally, they consider a multi-user environment.

The effectiveness of the prefetching scheme is influenced by the accuracy of prediction model. Scheduler-Assisted Prefetching [1] finds the next task by using a scheduler's queue. The pages of the next task are loaded into memory when the time quantum of current task is depleted until the base line. However, this scheme is proposed under the limited environment which frequently leads to swap-out due to heavy memory workload and does not consider the cold startup.

In studies of [4] and [5], they proposed a RT-PLRU scheme to find the optimal paging strategy. The scheme is focused on the time constraint of real time systems based on NAND flash. The NAND flash memory has been widely used as a secondary storage in the embedded systems. The RT-PLRU is a page replacement policy of the combination of pinning and LRU. The pinning scheme preloads pages and keeps them into main memory. It shows that the prefetching scheme contributes to satisfaction of real-time requirement even if the NAND-based system has the high read speed.

2.2 Smartphone Usages

Diversity in Smartphone Usage [6] is a comprehensive study of the smartphone use. The study found several characterized application usages of users activities. The application usage session described that users did not use installed applications as the same frequency. Specially, it is the same interest of our study. The authors analyzed the individual propensity of the user activities and their impact on energy consumption and use of network. We also collect the dataset of application usage and try to find patterns between the applications. Finally, the accuracy of the proposed prediction model is evaluated based on the dataset.

3 The Dataset Analysis

In this section, we analyze the traced data which is in order of launched applications on the smartphone. The relation between the applications will be inspected as application usage patterns.

3.1 The Dataset Gathering

Our work is based on the practical workloads which are collected by the monitoring applications on android mobile phones for a week. The application generates a log file in which the information of the launched applications is accumulated such as a start-time, the application's name and the binary's path. The dataset is summarized in Table 1.

Table 1. The overview of the dataset

User	# of applications launching	# of installed applications
User 1	220	82
User 2	389	128
User 3	191	69
User 4	317	207

3.2 Application Popularity

The number of installed applications and the number of launched applications for each user are shown in Fig.1 (a). Although the total installed applications are from 80 to 200, users were launched only partial applications from 15 to 45. Fig.1 (b) illustrates the application popularity ratio which is concentrated in 18~31% applications. The usage frequency of application is different, and only some applications are used concentrically. This usage pattern is similar with the research of [6].

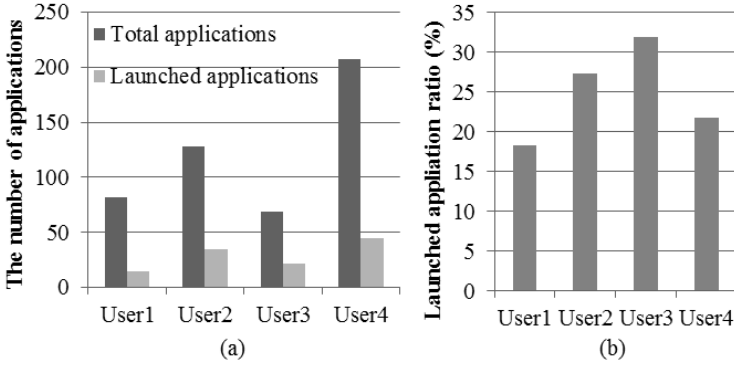


Fig. 1. (a) # of installed applications and # of launched applications for each user (b) The application popularity ratio

3.3 Application Usage Patterns

To find the user’s application access patterns, the data is classified into the ordered list of applications which are consecutively launched after each application finishes. It is required to define the following notation.

- $T(Index\ application)$: the ordered list of traced applications which are consecutively executed after an Index application finishes.

For example, if a user sequentially executes applications like App1, App2, App3, App1, App3, App1, App1 and App4, the reorganized data are represented by $T(App1)=\{App2, App3, App1, App4\}$, $T(App2)=\{App3\}$, $T(App3)=\{App1, App1\}$ and $T(App4)=Null$. Fig.2 depicts how to collect each $T(App)$ regarding the example.

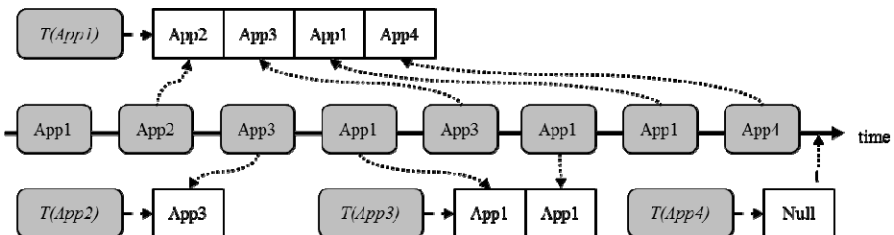


Fig. 2. Illustration of $T(App)$ at time t

In case that the number of elements is smaller than four, it is difficult to find a meaningful pattern. Hence, only the $T(APP)$ which is greater than four is analyzed to study the usage patterns. Fig. 3 and Fig. 4 illustrate the proportion of applications that starts after the application whose title is that of each graph, finishes. In Fig. 3, the user 1 executes Browser consecutively after using the Browser at the rate of 60% and launches MMS after Contacts was closed at the 55% chance. As shown in Fig. 4, the probability of beginning BeyondPod after the BeyondPod successively is 51%. KakaoTalk is expected to be used again after the KakaoTalk at the 53% rate. In the same way, the usage patterns of user 3 are launching Browser after Twitter and Facebook at the rate of 83% and 50% for respectively. The many higher relations between the applications are observed in other sets.

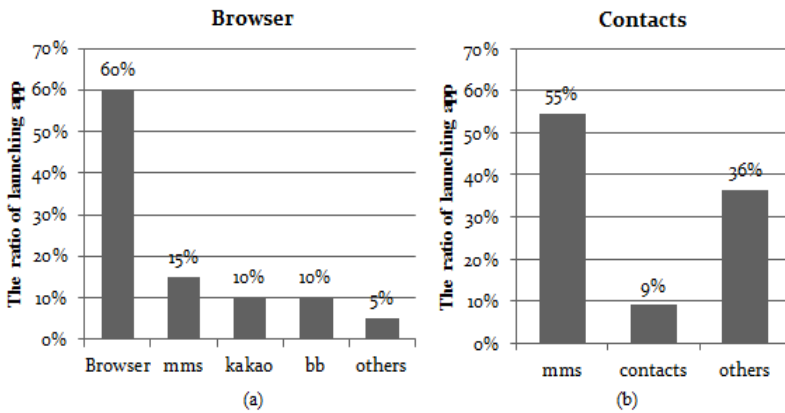


Fig. 3. Usage pattern of user 1

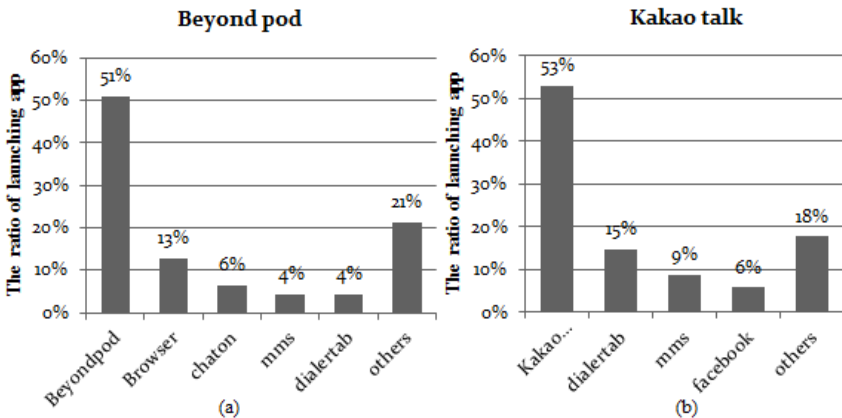


Fig. 4. Usage pattern of user 2

4 Design of UPP

4.1 The Application Prediction Model

The first-order Markov prediction model is applied to Behdad’s Preload [9]. Although the Markov model is a renowned probability model, the more lightweight model is required for mobile devices since mobile devices have poor hardware resources compared to desktop PCs, and they are battery powered.

Considering the usage patterns observed in previous sessions, our prediction model requires some features like the following. Firstly, the information of the application that is consecutively executed is important. Secondly, the recency of the application use as well as frequency information should be reflected, since the usage pattern changes variously according to time. Finally, it should have low run-time overhead to reduce the power consumption and the burden of computation time.

We propose a Window and Weighted Sum-based prediction scheme called WWS that acquires a candidate based on the weighted sum of elements in the sliding window. The similar approach, a Window-based Direct Address Counting (WDAC) is adopted in the study [10].

The following notations are defined to explain the WWS mechanism.

- $W(\text{Element application})$: the sum of element application by WWS.
- $C(\text{Index application})$: the name of an element application which has the maximum $W(\text{Element application})$ among $T(\text{Index application})$.

Fig. 5 depicts the algorithm on how to select the candidate when App1 finishes. App1 is the index application and the element applications are App3, App2, App3, App4, App3 and App5 sequentially. The window size is assumed to six. The results of $W(\text{element})$ are as follows; $W(\text{App2})=0.4$, $W(\text{App3})=1.8$, $W(\text{App4})=0.8$ and $W(\text{App5})=1.2$. $C(\text{App1})$ is App3 which has the largest value. Therefore UPP loads the pages of App3, when App1 finishes.

When updating the list, the new comer is added at the end of window, and the head of window is deleted. To manage WWS, the memory is required as much as $\text{window size} * \text{sizeof}(\text{Application identifier})$ for each application. It is a small amount. The computation time is also negligible. Thus, WWS is a suitable scheme for mobile devices.

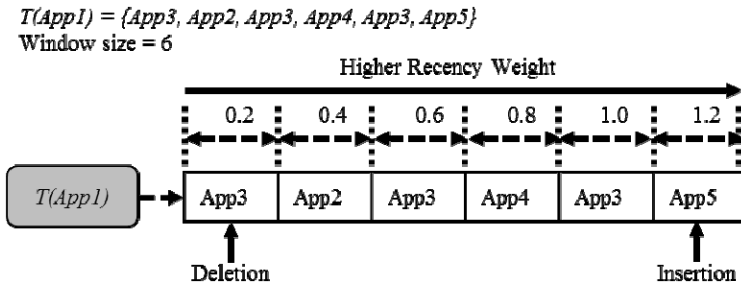


Fig. 5. WWS example

4.2 The Selection of Pages for Prefetching

We focus on accessing pages to induce file IO operations. This kind of access can be divided into two that are a major page fault that occurs when demand paging and explicit file IO call such as read and write. Thus, UPP gathers the information of pages from a major page fault and file IO calls during the application's startup time.

4.3 The Triggering Time of Prefetching

The application is launched through the main screen application (or called the home launcher) which has the entry points of all applications. In other words, the activating time of the main screen is the interval between applications. As shown in Fig. 6, the interval is enough to load fault pages. Thus, the prefetching is triggered when the main screen is activated.

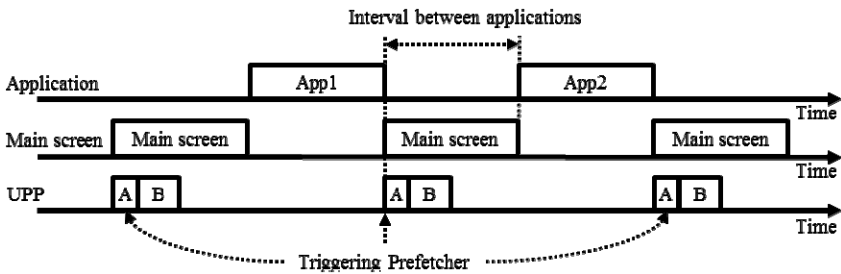


Fig. 6. The triggering time of Prefetching (A: Predicting the next application, B: Loading pages)

5 Evaluation

We explain the overall mechanism of UPP and discuss the results of experiment in this section. The goal of the experiment is to evaluate the effectiveness of WWS on the dataset and to improve startup time by the prefetching.

5.1 Implementation

UPP is composed of three components which are concretizing prefetching lists, tracing application usage and prefetching. Fig. 7 illustrates the mechanism of UPP. The activity of concretizing prefetching lists is represented by Arabic numbers (1~6). The capital letters (A, B) show the flow of tracing application usage which is $T(App)$. The prefetching activity is predicting a candidate application $C(App)$ and loading memory pages that are referenced in the prefetching list. The small letters (a~c) are explained as the prefetching sequence. The activities are implemented on both Android 2.2 and Linux kernel 2.6.29.

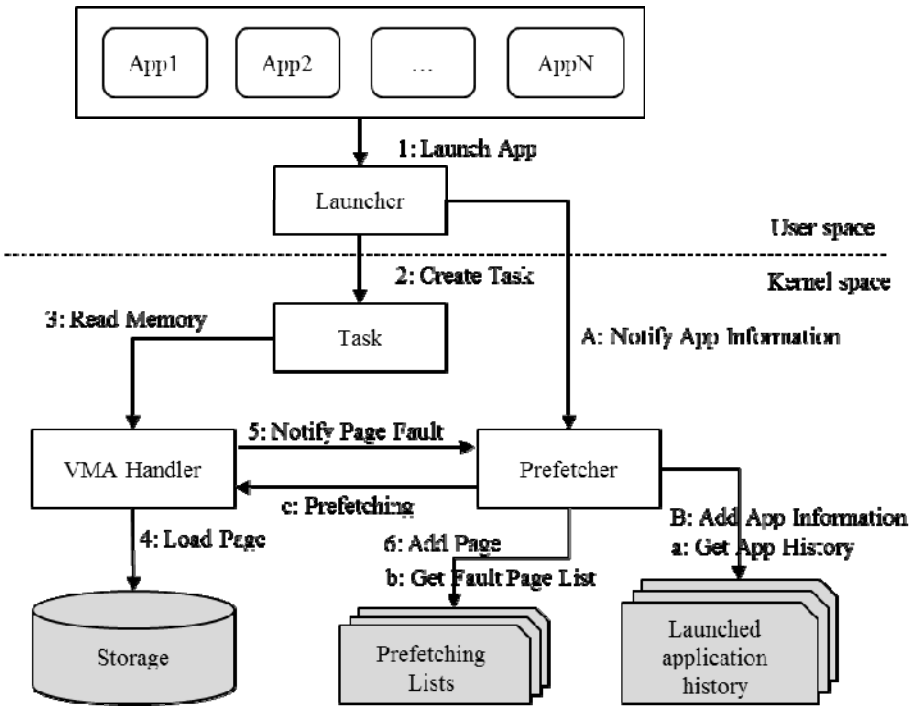


Fig. 7. UPP mechanism

5.2 Experimental Setups

Experiment for WWS

The accuracy of WWS is estimated by using the dataset of Table 1 in off-line. The window sizes are set to four and eight to observe the changes depending on various sizes.

Experiment for Prefetching

The prefetching is executed on an Android 2.2 emulator with Linux kernel 2.6.29. The size of a target application is 2.31 Mbytes. The android platform informs the time from the starting activity to the first exposed window at *windowsVisible()*. Hence, the time notified by the android is employed as the startup time. We use the average time of the multiple runs in cold startup for the prefetching and the no-prefetching.

5.3 Experimental Evaluation

WWS on the Dataset

We analyze the applications whose number of elements is greater than ten ($n(T(App)) > 10$). Fig. 8 (a) and Fig. 9 (a) are each user's application access pattern. Fig. 8 (b) and

Fig. 9 (b) depict the accuracy of WWS scheme according to the window size of four and eight. As shown in Fig. 8, the accuracy of WWS is at the rate of 54~59% if the access ratio concentrates on a few applications. The WWS is effective at the similar cases to Fig. 8 (a). Although a user launches various applications like Fig. 9 (a), we can succeed to predict the next application correctly about 20% as in Fig. 9 (b). A window size influences on the result of WWS slightly. Higher accuracy is shown in the narrow range when only partial elements of $T(App)$ are repeatedly launched in the small set. Otherwise, if a few elements start repeatedly in the large set, better accuracy is resulted in. The optimal window size strongly relies on the usage patterns. For that reason, we leave finding on optimal window size as our future work.

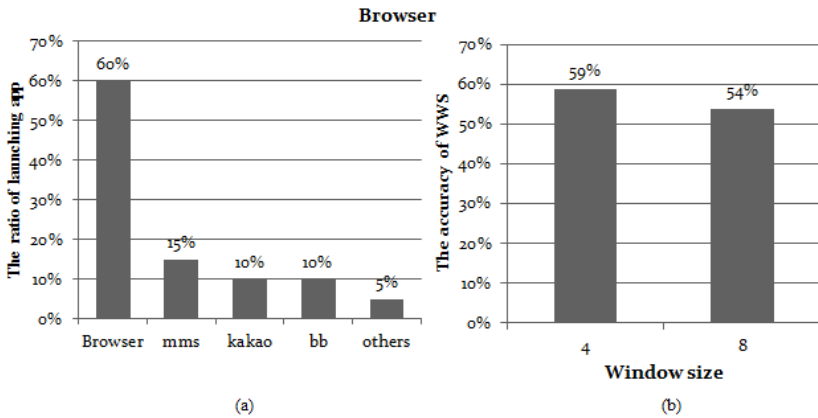


Fig. 8. (a) Usage pattern for Browser (b) WWS accuracy of Browser's candidate

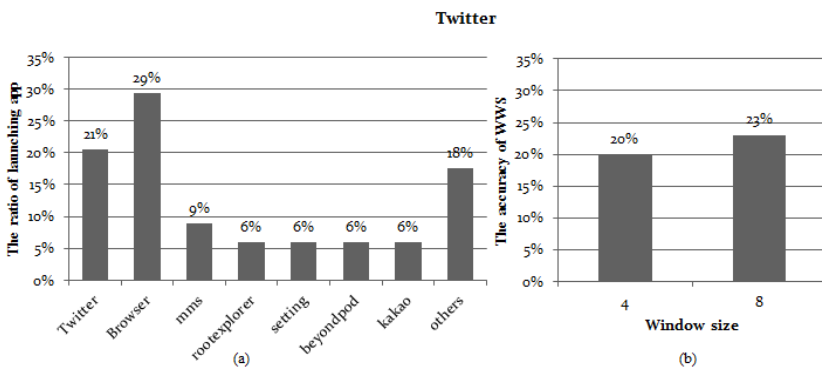


Fig. 9. (a) Usage pattern for Twitter (b) WWS accuracy of Twitter's candidate

Experiment for Prefetching

As shown in Fig. 10, the startup time with the prefetching is reduced by about 5% compared to the no-prefetching.

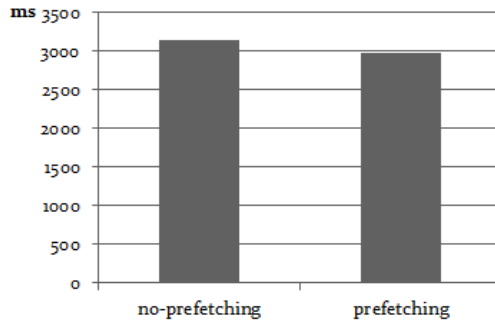


Fig. 10. Performance evaluation of the prefetching

6 Conclusion

Previously, we comment that the application startup time is important as the user perspective performance. If page faults occur frequently in the startup time, the user experience is subjected to an adverse effect.

Our study shows that users have their own usage patterns for each application. We propose a WWS scheme as a prediction model for mobile devices. WWS which reflects the frequency and the recency shows meaningful accuracy for the datasets. The scheme is low run-time overhead and low memory consumption. Thus, it is acceptable to mobile devices.

The improvement of the startup time by the prefetching component of UPP is smaller than our expectation since page faults still remain.

As our future work, we will study to find the best prediction model and optimize UPP.

Acknowledgments. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2012-(H0301-12-3001))supervised by the NIPA(National IT Industry Promotion Agency).

References

1. Belogolov, S.A., Park, J., Hong, S.: Scheduler-Assisted Prefetching: Efficient Demand Paging for Embedded Systems. In: Proc. the 14th IEEE International Conference on Embedded and Real-Time Computation Systems and Applications (RTCSA), pp. 111–119 (2008)

2. Chiou, D., et al.: Scheduler-Based Prefetching for Multilevel Memories. MIT Computation Structures Group Memo 444 (2001)
3. Microsoft. Windows PC Accelerators, <http://msdn.microsoft.com/en-us/windows/hardware/gg463388.aspx> (updated: October 8, 2010)
4. Kim, J., Lee, D., Lee, C., Kim, K.: RT-PLRU: A New Paging Scheme for Real-Time Execution of Program Codes on NAND Flash Memory for Portable Media Players. *IEEE Transactions on Computers* 60(8) (2011)
5. Kim, J., Lee, D., Kim, K., Ha, E.: Real-Time Program Execution on NAND Flash Memory for Portable Media Players. In: *Real-Time Systems Symposium*, pp. 244–255 (2008)
6. Falaki, H., Mahajan, R., Kandula, S.: Diversity in Smartphone Usage. In: *MobiSys* (2010)
7. Compressed Caching for Linux: <http://code.google.com/p/compcache/>
8. Bovet, D.P., Cesati, M.: *Understanding the Linux Kernel*, 3rd edn. O'Reilly (2005)
9. Esfahbod, B.: Preload-An adaptive prefetching daemon. Master's thesis, Graduate Department of Computer Science, University of Toronto, Canada (2006)
10. Park, D., Du, D.H.C.: Mass Storage Systems and Technologies (MSST). In: *2011 IEEE 27th Symposium*, pp. 1–11 (2011)
11. Joo, Y., Ryu, J., Park, S., Shin, K.G.: FAST: Quick Application Launch on Solid-State Drives. In: *Proc. FAST 2011 Proceedings of the 9th USENIX Conference on File and Storage Technologies* (2011)

EIMOS: Enhancing Interactivity in Mobile Operating Systems

Sunwook Bae¹, Hokwon Song¹, Changwoo Min¹, Jeehong Kim²,
and Young Ik Eom²

¹ Samsung Electronics Co., Ltd., Suwon, Korea

^{1,2} School of Information and Communication Engineering
Sungkyunkwan University, Suwon, Korea

{swbae98, hokwon, multics69, jjilong, yieom}@ece.skku.ac.kr

Abstract. Interactivity is one of the most important factors in the computing systems. There has been a lot of research to improve the interactivity in traditional desktop environments. However, few research studies have been done for interactivity enhancement in mobile systems like smart phones and tablet PCs. Therefore, different approaches are required to improve the interactivity of these systems. Even if multiple processes are running in a mobile system, there is only one topmost process which interacts with the user due to the resource constraints like small screen sizes and limited input methods. In this paper, we propose EIMOS, a system which identifies the topmost process and enhances the interactivity. Our system improves the CPU process scheduler and I/O prefetcher in the mobile operating system. We also implement EIMOS in the Android mobile platform and performed several experiments. The experimental results show that the performance is increased up to 16% compared to that of the existing platform.

Keywords: Interactivity, Topmost process, Mobile system, Operating system.

1 Introduction

The computing hardware technology has been rapidly developed, but the issue of interactivity still remains due to multitasking and software bloat. Many processes execute concurrently in the system; some interact with users in the foreground, and the others run just in the background. These processes need sufficient system resources like CPU, memory, and I/O to run smoothly. The traditional operating systems allocate those resources to the processes that are more important to the user first, e.g., interactive processes. In the mobile systems, multitasking is also supported, but the performance of them is lower than desktop systems due to the limited system resources. Therefore it is more important to improve the interactivity in the mobile devices.

There has been lots of research to improve the interactivity in the operating systems of desktop environments [3, 6, 8, 9, 10, 11, 12]. However, few research studies exist for that issue in the mobile systems. In desktop environments, it is possible for the user to work with multiple foreground processes on a single screen. For example, it is feasible to run a

translator or a calculator while working on a word processing program. In mobile systems, the multitasking feature is also supported but they use different methods to interact with users due to small screen sizes and touch-based input methods. With those restrictions, only one topmost process is used to interact with the user in most cases despite of supporting the multitasking feature [1]. The interactivity experienced by mobile system users depends on the topmost process as in Fig. 1.

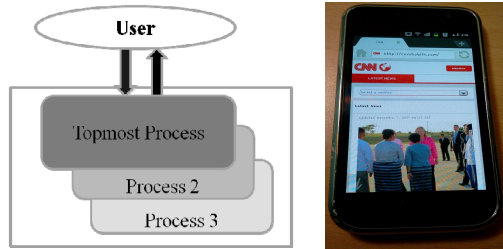


Fig. 1. Topmost process in mobile system

In this paper, we present EIMOS, a system to improve the interactivity by considering the mobile system characteristics. EIMOS identifies the topmost process and favors it to enhance interactivity of the system. It improves the process scheduler and prefetching mechanism of the I/O data. EIMOS also has a small runtime overhead and small amount of modifications to the existing operating systems. Specific contributions are as follows:

- Identify the interactive process with a small runtime overhead
- Adaptively apply the process scheduling and I/O prefetching for interactivity
- As far as we have known, this is the first research to improve the interactivity of the mobile operating systems.

The rest of the paper is organized as follows: we review related work in Section 2 and describe the key ideas of the paper and the implementation details in Section 3. Section 4 gives experimental evaluation, and we present conclusion and future work in Section 5.

2 Related Work

This section describes representative related work. We start by reviewing the Linux kernel scheduler. Moreover, we review typical research activities that are proposed to improve the interactivity on the desktop environments.

2.1 Process Scheduler in the Linux Systems

In the Linux kernel, the $O(1)$ scheduler was introduced in the early version 2.6 by Ingo Molnar [2, 3]. The scheduler improved the interactivity of the interactive processes in two ways. The first was to give a dynamic priority [2] to the interactive

processes by analyzing the average sleep time of the processes. The second was to let the interactive processes remain in the active queue [2] even if the time slice expired. It therefore removed the delay by other processes. From the Linux kernel 2.6.23, Completely Fair Scheduler (CFS) [4, 5] scheduler has replaced by the old O(1) scheduler. The design goal of the CFS scheduler is to provide fair CPU resource allocation for executing processes. However, it also optimized the interactivity by adjusting the virtual runtime value of the processes that slept for a period of time [6].

Lo et al. [7] proposed Modified Interactive Oriented Scheduler (MIOS), a scheduler that improves the interactivity by eliminating unnecessary overhead from the Linux scheduler. MIOS achieved the improvement by removing the overhead caused by maintaining two queues, active and expired, in the O(1) scheduler.

Above three studies used the limited information, such as the average sleep time, to identify the interactive processes. Because of the lack of information, they cannot detect the right interactive processes even if they improved the performance and the scalability of the scheduler.

2.2 Research in the Desktop Environment

2.2.1 Identification of the Interactive Processes

Etsion et al. [8, 9] suggested that multimedia processes should be treated in a special way as well as interactive processes. They defined these as Human Centered (HuC) processes which are detected based on the display output production. They also showed that the performance of the HuC processes was not degraded even if heavy background processes were running. However, the approach is not suitable in the mobile system because there are still many interactive processes which are not related to multimedia jobs, such as messenger and calendar applications.

Zheng and Nieh [10] presented RSIO, an approach improving the response time of interactive latency-sensitive processes. RSIO identified the interactive processes dynamically by monitoring the I/O channels usage for user interactions and then boosted the priority of interactive processes when they handled latency-sensitive activities. The problem of this approach is that some I/O channels, which are suggested in RSIO such as tty and mouse devices, cannot be applied to the mobile system and it is not enough to improve the interactivity by adjusting priorities in the processor scheduler.

2.2.2 OS Support for Improving the Interactivity

The previous research was focused on identifying the interactive processes and prioritizing them in the process scheduler to improve the interactivity. Yan [11], however, presented a holistic approach addressing process scheduling, memory management, and I/O scheduling. He proposed a new process scheduling policy, LRU memory management system, and disk I/O scheduling policy. He showed the improvement of computer responsiveness by modifying on the existing Linux/X desktop system.

Yang proposed Redline [12], a system designed to support interactive and resource-intensive modern applications in commodity operating systems. It maximized the

responsiveness of interactive applications by orchestrating memory and disk I/O management with the CPU scheduler.

Above two studies focused on the desktop environment and modified the operating systems entirely. The suggested policies of memory management and disk I/O scheduling also have a large runtime overhead. Therefore, it is hard to apply them in the mobile operating systems.

3 Design and Implementation of Eimos

In this section, we introduce a design and implementation of EIMOS in detail. EIMOS is a system that improves the interactivity in the mobile systems. EIMOS consists of two steps. The first step is to identify the topmost process which is an important process for interactivity. The second step is to improve the interactivity by allocating the CPU/I/O resources to the topmost process first.

3.1 Identification of the Topmost Process

The method to identify the topmost process depends on the mobile operating systems. It is also impossible to detect the process in the OS kernel layer alone because the information can be managed by the UI or windows framework. In case of the Android mobile platform, we can use a low memory killer module which was newly added to the Linux kernel to address the out-of-memory problem [13]. The main role of the low memory killer module is to kill the less important processes depending on the information of Table I. The `oomkilladj` variable in the `task_struct` structure which handles the process information is used to classify the processes and the variable is updated from the ActivityManager which is a component to handle the UI information in the Android mobile platform. According to the `oomkilladj` variable, the process groups are divided into seven. The low memory killer module finds the less important processes from the groups of high `oomkilladj` values, such as `EMPTY_APP` and `HIDDEN_APP`. EIMOS, however, identifies the topmost process by using the information of the `FOREGROUND_APP` group. This approach is simple and has less runtime overhead comparing with previous studies.

Table 1. `oomkilladj` values for the classes of processes

Group	Oomkilladj
SYSTEM	-16
FOREGROUND_APP	0
VISIBLE_APP	1
SECONDARY_SERVER	2
HIDDEN_APP	7
CONTENT_PROVIDER	14
EMPTY_APP	15

3.2 Scheduler Support

There can be many methods to improve the interactivity of the topmost process. We first present to allocate the CPU resource to the topmost process by giving an additional bonus to the process in the OS scheduler. Fig. 2 shows the workflow of the scheduler support in EIMOS. First, we add a topmost flag to all `task_struct` structures and initialize it to a value of `false` on device booting time. Second, we identify the topmost process by monitoring the low memory killer module and set the topmost flag of the process to a value of `true`. Third, we adjust the priority of the topmost process in the OS scheduler. Finally, if the topmost process is changed, we modify the topmost flag and the priority of the previous process to an old value. The Linux O(1) scheduler and CFS scheduler are the priority-based scheduler and the processes with higher priorities get better response time. We give a dynamic priority bonus of 19 to the topmost process in EIMOS. This approach has a little impact on other processes and the interactivity of the topmost process does not significantly decrease in situation of running lots of background processes.

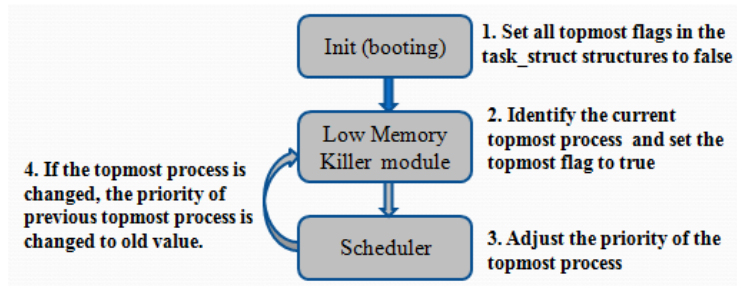


Fig. 2. Workflow of process scheduler support in EIMOS

3.3 I/O Prefetch Support

The I/O performance is always a bottleneck of the computer systems. Recently, there are many applications to handle big data like music videos and movies. In these cases, the I/O prefetch technique which loads the I/O data to memory in advance is very effective due to hiding the I/O latency. We therefore propose the interactivity aware adaptive prefetch scheme here. The previous prefetch scheme in the Linux kernel used the information based on the sequentiality, but we consider the interactivity additionally. We implement the read-ahead policy to read more adjacent pages of data when the topmost process requests the system to read the I/O data. Fig. 3 shows the workflow of the I/O prefetch support in EIMOS. In the Linux kernel, the maximum size of the read-ahead buffer is fixed but we also extend it eight times in case of the topmost process. This approach can improve the I/O performance of the topmost process with a little modification.

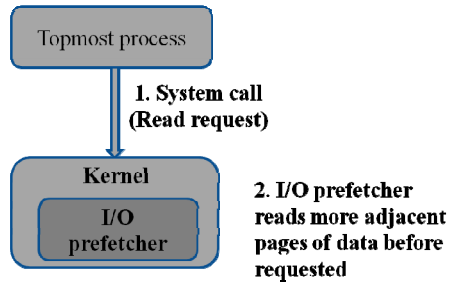


Fig. 3. Workflow of I/O prefetch support in EIMOS

4 Evaluation

The hardware environment for the experiments is shown in Table II. We implemented EIMOS by modifying the Linux kernel sources in the Android 2.2 froyo emulator. We developed two micro benchmarks and used one more realistic workload for evaluation: CPU and IO bound micro-benchmarks and a Linpack benchmark [14] which is a measure of a system's floating point computing power. Linpack has been used for years on all types of computers and it shows the performance of the topmost process. We can evaluate the improvement of the interactivity by using these benchmarks and by comparing between EIMOS and original Android emulator.

Table 2. Experiment environment

CPU	Intel(R) Core(TM) i5 CPU 2.40 GHz
Memory	4 GB RAM
OS	Ubuntu 10.10 (Linux)

We first made a simple micro-benchmark for measuring the CPU performance of EIMOS. Fig. 4 shows the source code of the benchmark. We ran this micro-benchmark 50 times and measured the average time of them as in Fig. 5. The first graph shows the result by running this micro-benchmark alone and the second graph is the result by running the micro-benchmark with 10 stresses of the same benchmark application running on the background. Without any background loads, the performance of EIMOS improves by 4%. As the load on the system increases, the performance of EIMOS improves by 16%.

```

For (i=0; i<1000000; i++)
    value += i;
    For (j=0; j<1000000; j++)
        value += j;

```

Fig. 4. Source code of CPU bound micro-benchmark

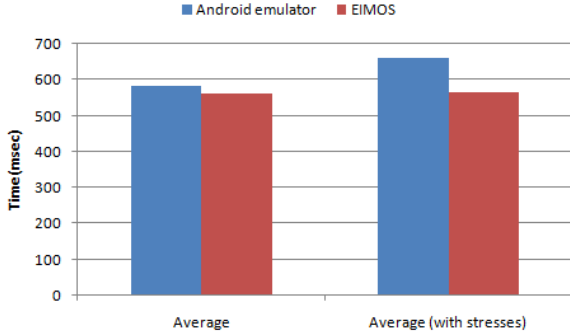


Fig. 5. Average time of CPU bound micro-benchmark

Fig. 6 shows the code of our micro-benchmark for measuring I/O performance. The benchmark reads 48 bytes data sequentially from the file of total 30MB size and finishes when it reaches the end of the file. We ran the micro-benchmark by changing the maximum size of the read-ahead buffer by 8 times and 16 times additionally. Fig. 7 shows the result of the average time of 10 trials. We can see that the I/O performance improved by 5% and 7% respectively. Although this I/O operation performs the sequential read requests, there was not much improvement from the experiments than we expected. We should consider the page cache of the Linux kernel for I/O operation. Fig.8 shows the result of taking time to read a same file of 30MB size repeatedly. We can see little improvement after first trial because the page cache contains the previous I/O data. Therefore, we have a future research plan to improve the I/O performance in EIMOS considering the page cache.

```
byte[] buffer = new byte[48];
FileInputStream fis
= openFileInput("test.data");
    BufferedInputStream buf
= new BufferedInputStream(fis);
while (true) {
    num = buf.read(buffer);
    if (num < 0)
        break;
}
```

Fig. 6. Source code of I/O bound micro-benchmark

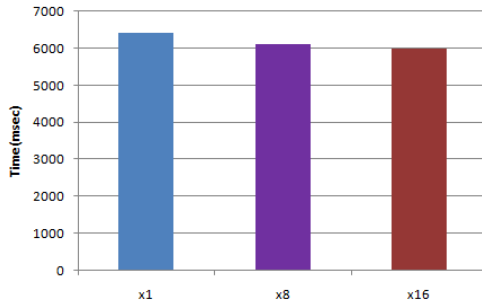


Fig. 7. Average time of I/O bound micro-benchmark

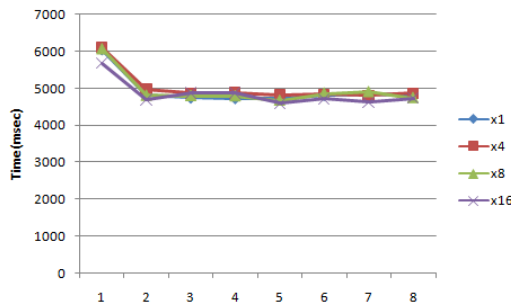


Fig. 8. Time to read same data repeatedly

Fig. 9 shows the result by running the Linpack 1.2.8 benchmark with and without background stresses. EIMOS also improves the performance maximum 18 times in the background stresses of the 10 micro-benchmark. In computing systems, if the background workloads increase, the performance of the topmost process should be degraded. However EIMOS shows the consistent performance in the heavy background stresses because EIMOS identifies the topmost process and boosts the priority of the process.

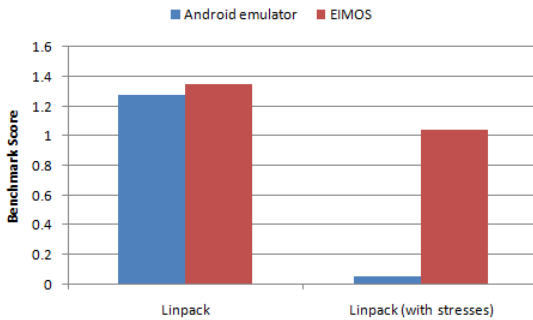


Fig. 9. Linpack 1.2.8 benchmark score

5 Conclusions and Future Work

This paper presents EIMOS, a new approach to enhance the interactivity on the mobile operating systems. EIMOS identifies the topmost process which has the biggest impact on user responsiveness and supports the process by adjusting the CPU scheduling and using the I/O prefetching technique. We implemented EIMOS in the Android mobile platform and the experiment results showed that the CPU performance of the topmost process was improved by 16% and the I/O performance by 7% in the heavy background stresses.

In future work, we will consider not only memory management, but also I/O scheduling to improve the interactivity of the topmost process. We are also interested in other resource management algorithms which are best suited for the mobile environments.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0025971).

References

1. Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., Estrin, D.: Diversity in Smartphone Usage. In: Mobile Systems, Applications and Services (MobiSys) (June 2010)
2. Bovet, D.P., Cesati, M.: Understanding the Linux Kernel, 3rd edn. O'Reilly (2006)
3. Molnar, I., Kolivas, C.: Interactivity in Linux 2.6 Scheduler (2003), <http://www.kerneltrap.org/node/780>
4. Molnar, I.: Linux CFS Scheduler (2007), <http://kerneltrap.org/node/11737>
5. Molnar, I.: A description of CFS design, <http://people.redhat.com/mingo/cfs-scheduler/sched-design-CFS.txt>
6. Wong, C.S., Tan, I.K.T., Kumari, R.D., Lam, J.W., Fun, W.: Fairness and Interactive Performance of O(1) and CFS Linux Kernel Schedulers. In: Information Technology, ITSim 2008 (2008)
7. Lo, L., Lee, L.T., Chang, H.Y.: A Modified Interactive Oriented Scheduler for GUI-based Embedded Systems. In: Computer and Information Technology (July 2008)
8. Etsion, Y., Tsafirir, D., Feitelson, D.G.: Desktop scheduling: How Can We Know What the User Wants? In: Proc. of the 14th International Workshop on Network and Operating Systems Support for Digital Audio and Video. ACM Press (2004)
9. Etsion, Y., Tsafirir, D., Feitelson, D.G.: Process Prioritization Using Output Production: Scheduling for Multimedia. ACM Transactions on Multimedia Computing, Communications and Applications (November 2006)
10. Zheng, H., Nieh, J.: RSIO: Automatic User Interaction Detection and Scheduling. In: Proc. of the 2010 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (June 2010)

11. Yan, L., Zhong, L., Jha, N.K.: Towards a Responsive, Yet Power-Efficient, Operating System: A Holistic Approach. In: Proc. of the 13th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (September 2005)
12. Yang, T., Liu, T., Berger, E.D., Kaplan, S.F., Moss, J.E.B.: Redline: First Class Support for Interactivity in Commodity Operating Systems. In: Proc. of the 8th Symposium on Operating Systems Design and Implementation (December 2008)
13. Linux/drivers/staging/android/lowmemorykiller.c,
[http://lxr.free-electrons.com/source/
drivers/staging/android/lowmemorykiller.c?v=2.6.29](http://lxr.free-electrons.com/source/drivers/staging/android/lowmemorykiller.c?v=2.6.29)
14. Linpack for Android, <http://www.greenecomputing.com/apps/linpack>

Development of Mobile Hybrid MedIntegraWeb App for Interoperation between u-RPMS and HIS

Young-Hyuk Kim, Il-Kwon Lim, Jae-Pil Lee, Jae-Gwang Lee,
and Jae-Kwang Lee*

Hannam Univ., Department of Computer Engineering,
Daejeon, Korea

{yhykim, iklim, jplee, leejk, jklee}@netwk.hannam.ac.kr

Abstract. This study was intended to extend the existing research on u-RPMS(USN Remote Patient Monitoring System), which is used to send patients' biometric information to HIS (Hospital Information System) through smartphones, and thereby make u-RPMS interoperate with HIS, and develop hybrid web/app in order to use integrated medical service and patient monitoring service on the mobile environment. In the case of u-RPMS, the existing study results were used. And, hybrid web/app named MedIntegra Web(Medical Integra Web) project was implemented with the use of Sencha Touch framework and HTML5/CSS3. The finally implemented MedIntegra Web was able to provide remote patients' information and status on the mobile environment, and, especially, to help the location of an emergent patient.

Keywords: u-RPMS, SenchaTouch, Hybrid Web/App, MedIntegraWeb.

1 Introduction

The desire for health and relevant services, starting from the wellbeing, extended to happy life, wellbeing life as well as longevity. In the past, the conditions of happiness were physical abundance and capital power, but, now people have changed their fundamental thought about the meaning of happiness, realizing that happiness can't be achieved through the physical richness. Accordingly, people have been interested in how to achieve good health and longevity and enjoy happy life, and, consequently, have strongly longed to search for and access information on treatment conveniently and use them efficiently [1].

Medical service use for health has been affected by many factors, including demographically environmental change, supply of medical resources, emergence of new disease, development of medical technology, and medical security system. As a result, people have required healthy life and services more in the current era than in the previous eras. Unfortunately, the national medical policy and service fail to satisfy people's desire for relevant medical services and health life. With the rapid aging society, in particular, the national finance hasn't met the public's expectation for

* Corresponding author.

social medical support, and therefore various issues have arisen in the society. Above all, the aged and patients in high danger, who are isolated from medical benefits and convenience, have spent much money receiving medical service. Therefore, to solve the problem, a study on u-RPMS (USN Remote Patient Monitoring System) was conducted [2].

u-RPMS is u-Healthcare system in which the aged and patients in high danger are given information on their health status while doing activities ‘at any time, and at any place, and without restraint’. Originally, the u-RPMS was aimed at remotely monitoring the aged and patients in high danger, who should constantly visit hospitals to receive medical checkup, and thereby reducing the number of their visits to hospitals and preventing the cost caused by their hospitalization, and eventually reducing the national health and welfare budget. To implement the u-RPMS, it is necessary to gather patients’ biometric information through the USN as shown in the [Fig. 1] and send the information to smartphone on the WBAN (Wireless Body Area Network) environment. The smartphone encrypt the collected biometric information and then make the information saved into the medical centers’ server.

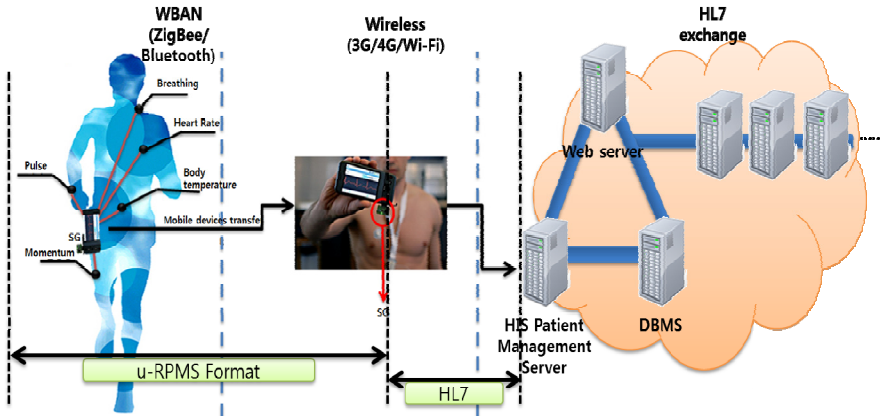


Fig. 1. u-RPMS service

To communicate between medical centers’ server and HIS (Hospital Information System) and exchange information between them, HL7 (Health Level 7) studied in the previous research [3] was used to design MII (Medical Integra Interface), as shown in [Fig. 2], necessary to implement MedIntegraWeb in this thesis.

With the help of previous research [3], MII was implemented with Java based TCP/IP Socket communications. And, a test proved that HL7 standard based communications were smoothly made. Regarding the MII system, the existing u-RPMS’s module and MII necessary to convert into HL7 format are built in a smartphone. At the beginning, biometric information is converted into a string type through the ‘conversion’ belonging to SGM [2], and the data is sent to MII’s HL7 Generator to be changed into HL7 format [3]. After that, HIGHT encryption is executed

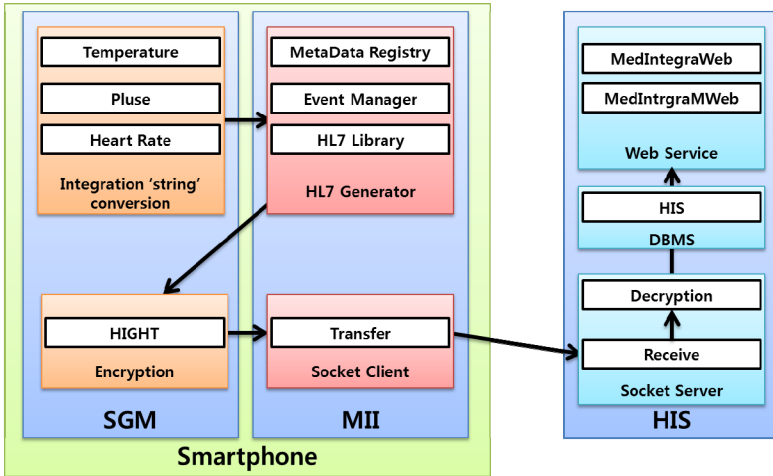


Fig. 2. MII (Medical Integra Interface)

by reason of security, and eventually the encrypted data is delivered to medical centers’ HIS via MII’s socket communications module. During all communications, ‘AA (Response)’, ‘AE (Error)’, or ‘AR (Reject)’ must be transmitted through ACK to see if the data receipt is successfully made.

On the basis of the previously studied system [2, 3], hybrid web/app, which helps use integrated medical service and patient monitoring service on the mobile environment, was designed and implemented in this thesis. These researchers call the study MedIntegraWeb (Medical Integra Web) project, which was implemented for the hybrid web/app with the use of Sencha Touch framework, a mobile framework, and with the use of HTML5/CSS3. In addition, to authenticate users who have authority to open patients’ biometric information, these researchers applied positioning based access control technology. Therefore, in chapter 2, positioning based access technology and context based access technology are analyzed for user authentication. In chapter 3, MedIntegraWeb is designed. In chapter 4, MedIntegraWeb designed in the chapter 3 is actually implemented to establish a server, and hybrid web/app for android OS is developed to be installed into actual smartphone and smartpad and to be tested. In chapter 5, the meaning and issues of this research and future studies are proposed.

2 Related Studies

As MedIntegraWeb provides patients’ personal information, any data about their life, and medical doctors’ information, it is necessary to prevent unauthorized users from accessing the information. For the reason, context based access control and positioning based access control were analyzed.

2.1 Context Based Access Control

Since u-healthcare, based on the wired and wireless internet, processes massive amount of medical information, it can cause ethical issues, including user privacy and information security, and information-divide issue. In particular, medical information directly related to patients' personality, life and privacy should meet the following security requirements [4, 5, 6].

Strong Authentication: the current healthcare authentication technology uses bio information, such as a footprint, an iris, a signature, and a voice, going beyond a password.

Electronic Signature: Electronic signature guarantees four security requirements-authentication, integrity, confidentiality, and non-repudiation. When u-healthcare is applied, it guarantees a user's identification, prevention of forgery and alternation of medical records, non-repudiation of the creation of medical information.

Context Aware Control Infrastructure: Adaptive security service to change and manage security method, authority, and security level dynamically in accordance with context should be made possible.

No other service systems have complicated access rules than u-healthcare system in terms of access right interacting with many users and roles. And, regarding users, roles and permission of objects, the access control should be achieved through the context awareness.

The access control between a client object and a server object through ACL (Access Control Lists) helps maintain the right of access to information, although a user's context, such as a location, working hours and changes in roles, change. As a result, many studies based on RBAC (Role Based Access Control) have been applied to u-healthcare system. Among the studies, studies on dynamic access control technology to limit access information depending on a system's context [7, 8] and studies on u-healthcare security system to improve the existing context-aware based RBAC (xoRBAC) [9, 10] provide the service similar to one proposed in this thesis.

The studies were all related to context based access control technology for u-healthcare system, having something in common in terms of the offering of flexible access control. But, their main study environment was not the mobile environment, so the frequent communication between clients and servers caused a system's loads. In this respect, simpler access technology should be applied to the mobile environment.

2.2 Positioning Based Access Control

As the existing role based access control method was not able to perform access control based on context-aware like time and location, GRBAC model [8] was proposed to address the issue. GRBAC model uses subject role, object role, and environment role in making an access control decision, and thereby expanded the existing role based access control. In other words, due to the limit of RBAC, relevant studies have been expanded to context-based access control.

As users, objects and environmental elements were structuralized as roles, the model provided simplicity and flexibility of access control policy technology. But, the access control policy, which is operated on the basis of a variety of information, caused an issue in the user environment that requires short use time and simple, fast access. As a result, positioning based access control technology, through which a user's location is determined on the basis of the information on mobile' GPS and 3G/4G base station, was used.

Positioning information based service is divided into two types, one of which is Multi-Layer System that puts each user's positioning coordinates on a map. The other type is Search Based System that calls map information placed within the scope of a user's coordinates and then search for proper information in the map information [11].

3 u-RPMS MedIntegraWeb Design

[Fig. 3] presents the overall framework structure of u-RPMS MedIntegraWeb. MedIntegraWeb should be equipped with an interface module to interoperate with each service system. Therefore, in this study, MII to which HL7, standardized information exchange format between systems is applied was implemented and loaded as the interface module.

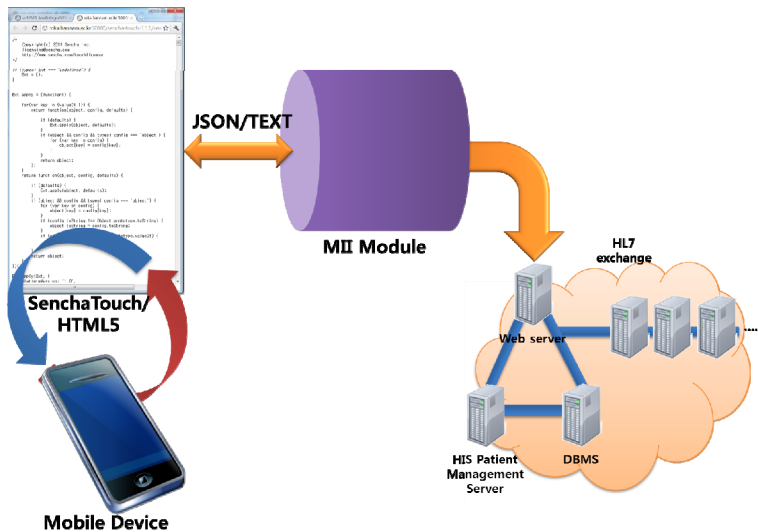


Fig. 3. Overall framework structure of u-RPMS MedIntegraWeb

Each system must have TCP socket and support web service, and a mobile web application server exchanges JSON typed data with Sencha Touch through HTTP. A mobile app supports a web screen, which was implemented with Sencha Touch, by using webview, and provides biometric information chart graphs, patients' status record, patients' location, and physicians in charge by using javascript.

Sencha Touch provides various functions related to the interoperation with a server, and its basic model is presented in [Fig. 4-(a)] [12]. But, MedIntegraWeb proposed in this thesis interoperates with server as shown in [Fig. 4-(b)]. Each interoperation function is shown as follows.

Model: defines data structure.

Proxy: is charge of direct connection to interoperate with a server.

Store: is charge of processing of retrieved data at a local storage.

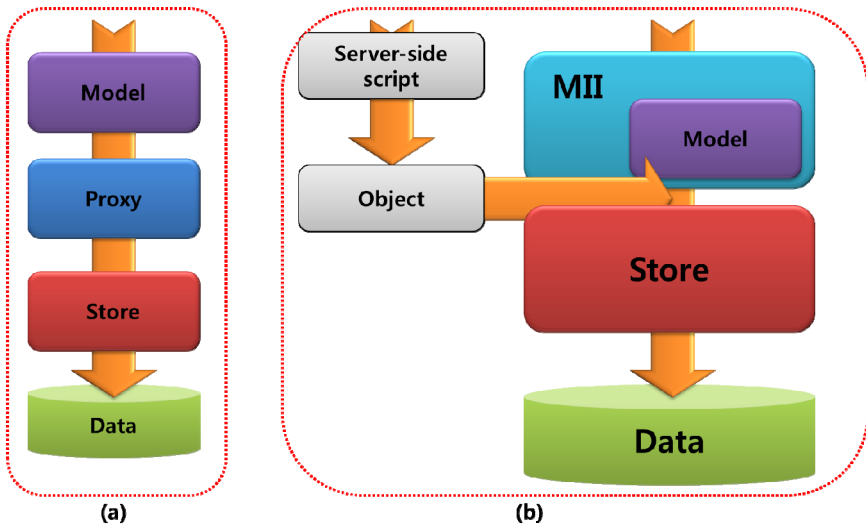


Fig. 4. Structure of MedIntegraWeb's interoperation with server

MedIntegraWeb is implemented as hybrid web/app using Sencha Touch framework. Therefore, app. and web divides operation process of the same work and execute each operation. In other words, a mobile device utilizes resources to display relevant images on a users' screen, and a server separately executes an operation to display relevant images on a screen through a user's app. [Fig. 5] presents such division and process.

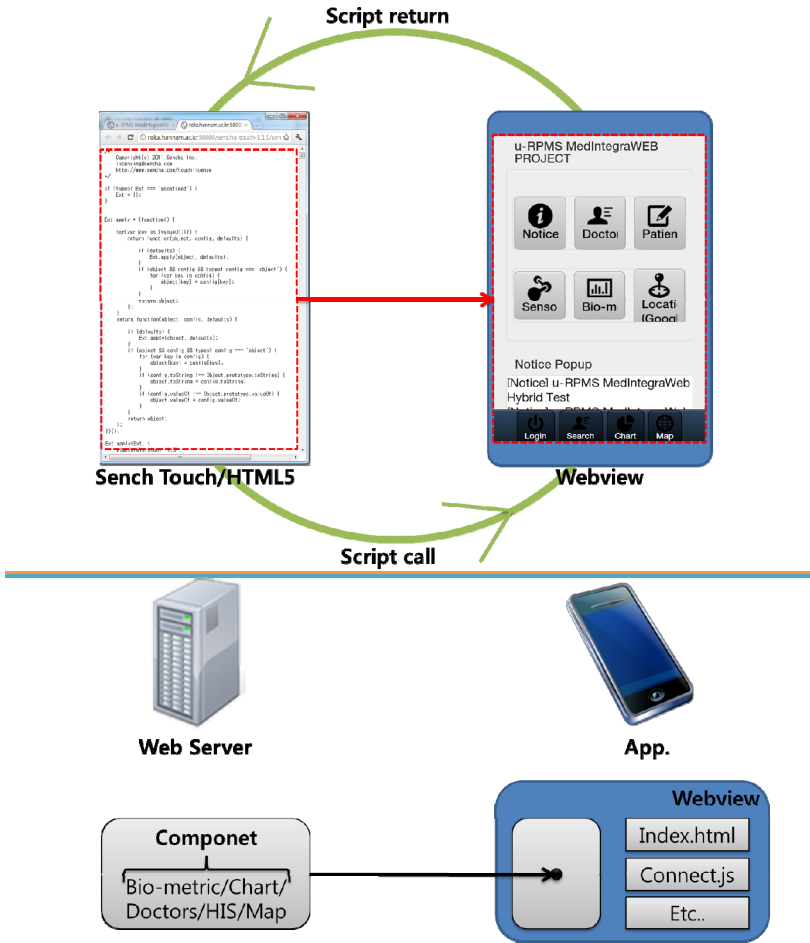


Fig. 5. Structure of each role of Web & App

4 Implementation Results

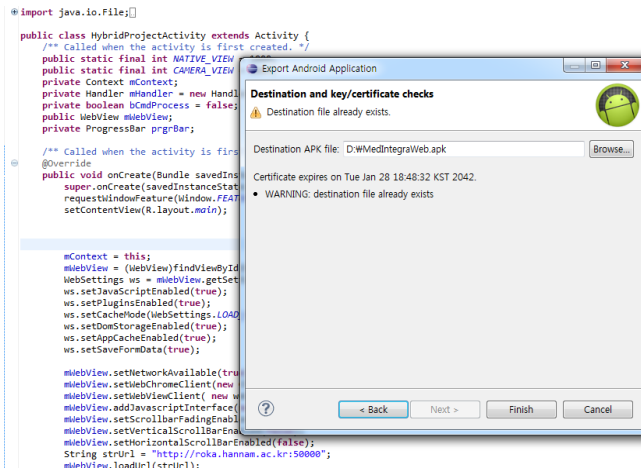
4.1 Implementation Environment

Table 1 shows the implementation environment of MedIntegraWeb for test service and operation. Just as the previous researches described earlier used Java language, these researchers implemented a test-bed on the android OS environment to make the use of the results of the previous studies.

Table 1. Implementation Environment

Item	Contents
Language	Java, HTML5, JSP, Android SDK 2.3
Mobile framework	Sencha Touch framework 1.1
HL7 API	HAPI 1.2
Development Tool	Eclipse
Web Server	Intel Core2 Quad Q9400 2.66GHz, 4GB
HIS Server	Intel Core i5 3.2GHz, 4GB
Client	Samsung Galaxy S2 ODROID-7 Windows Server 2008
OS	Windows 7 64bit Android OS 2.3
DB	MySQL55

In this study, Sencha Touch framework Ver 1.1 and JSP as a server-side script language were used. Server was divided into HIS server that has medical centers' DB and information and into Web server that provides MedIntegraWeb service. As client devices, commercial product Galaxy S2 and ODROID-7 tablet for developers were used to access relevant services. [Fig. 6] illustrates creation of .apk file with MedIntegraWeb in eclipse.

**Fig. 6.** Creation of MedIntegraWeb.apk

4.2 Image of Service Implementation

[Fig. 7] shows captured images of each running service in the implemented MedIntegraWeb, which is largely categorized into 6 items in providing information.

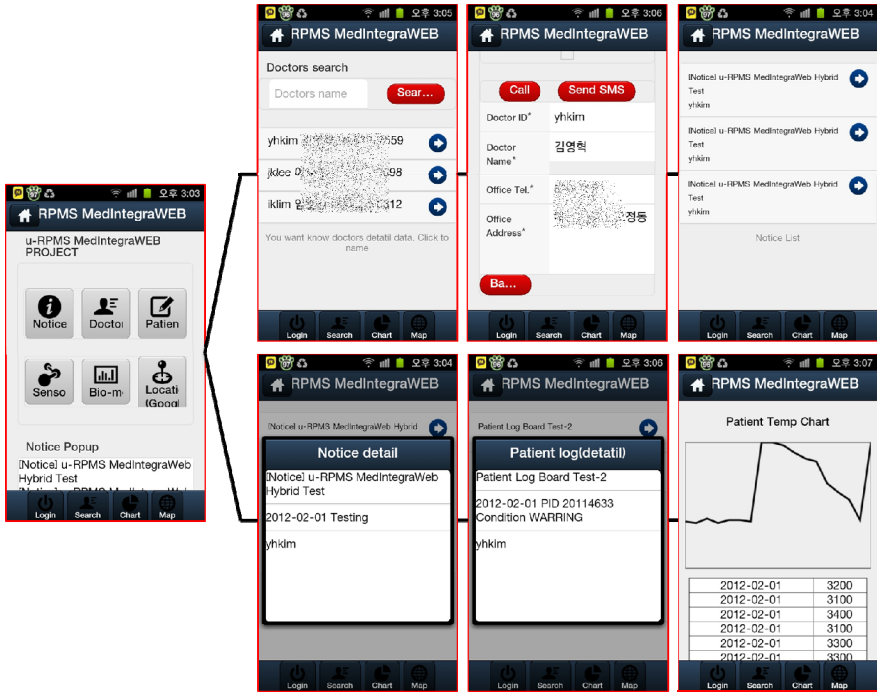


Fig. 7. MedIntegraWeb’s overall service structure

The first item [Notice] is a space where notices to be sent to users are displayed. The second item [Doctors] is in charge of providing information for medical doctors registered into HIS and of making a direct connection via telephone or text messages. The third item [Patient Log] displays patients’ changing medical records on a user (doctor)’s screen. The fourth item [Sensor Node State] shows the status of sensors attached to patients. So, any sensor causing a problem can be treated properly. The fifth item [Bio-metric Chart] currently displays a patient’s body temperature as a graph. In the future, heart rate, pulse, and other biometric information are expected to be provided. The last item [Location] is used as any emergent status changes in a patient occur in [Patient Log], showing the patient’s location.

[Fig. 8] presents a patient’s body temperature graph displayed on Galaxy S2 and ODROID-7, with which a user accessed MedIntegraWeb.



Fig. 8. Image of execution on Galaxy S2 (left) and ODR0ID-7 (right)

5 Conclusion

This thesis extended the existing research on u-RPMS (USN Remote Patient Monitoring System), which is used to send patients' biometric information to HIS (Hospital Information System) through smartphone, to make u-RPMS interoperate with HIS, and developed hybrid web/app in order to use integrated medical service and patient monitoring service on the mobile environment. u-RPMS was based on the results of the previous studies, and hybrid web/app named MedIntegraWeb project was implemented with Sencha Touch framework and HTML5/CSS3.

Android app was used in operating a test in order to make the most of the existing studies in which HL7 was applied to u-RPMS sending data to medical centers, and MII, an interface for compatibility with HIS, was implemented in Java language. The implemented MedIntegraWeb includes various functions, such as notice, search of a doctor in charge and direct connection, patients' medical records, information on sensor nodes attached to patients, charts based on patients' biometric information, and positioning tracking function to find a patient's location.

The test operation of the implemented application proved the possibility of its use for commercial service. But, for the implementation and test operation in this thesis, a patient's information was neither encrypted nor processed in security aspect. In

addition, positioning tracking of a patient's location can violate the patient's privacy, an issue which goes beside the point of this thesis.

In future research, it is necessary to study technologies and policies to solve the issues described earlier-security and privacy- and apply them to systems. Also relevant studies and analysis should come to be followed.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0013029).

References

1. Kim, Y.-R.: Comprehensive Study on Social and Legal Issues of Internet Medical Information. *Media Science Research* 10(2) (June 2010)
2. Kim, Y.-H., Lim, I.-K., Lee, J.-K.: Mobile Based HIGHT Encryption for Secure Biometric Information Transfer of USN Remote Patient Monitoring System. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part V. LNCS, vol. 6786, pp. 83–95. Springer, Heidelberg (2011)
3. Kim, Y.-H., Lim, I.-K., Lee, J.-G., Lee, J.-P., Lee, J.-K.: Designing Medical Integra Interface with u-RPMS and HIS. In: *FutureTech 2012* (2011)
4. Weaver, A.C., Dwyer III, S.J., Snyder, A.M., et al.: Federated, Secure Trust Networks for Distributed Healthcare IT Services. In: *IEEE International Conference on Industrial Informatics* (2003)
5. Wilikens, M., Feriti, S., Sanna, A., Masera, M.: A Context-Related Authorization and Access Control Method Based on RBAC: A case study from the health care domain. In: *Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies* (2002)
6. Covington, M.J., Long, W., Srinivasan, S.: Secure Context-Aware Applications Using Environment Roles. In: *Proceedings of the Sixth ACM Symposium on Access Control Models and Technologies* (2001)
7. Won, J.C., Ho, K.D., Chong, J.S.: Context-based Dynamic Access Control Model for u-healthcare and its Application. *KIPS Journal C* 15-C(6), 493–506 (2008)
8. Moyer, M.J., Ahamad, M.: Generalized Role-Based Access Control. In: *IEEE International Conference on Distributed Computing Systems (ICDCS 2001)*, pp. 391–398 (2001)
9. Kim, C.-B., Lee, S.-S., Lee, B.-S.: A Study on U-healthcare Security Framework based Context-Aware. *KIIT Journal* 6(4), 37–46 (2008)
10. Wilikens, M., Feriti, S., Sanna, A., Masera, M.: A Context-Related Authorization and Access Control Method Based on RBAC: A case study from the health care domain. In: *Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies* (2002)
11. Jang, W.-J., Lee, H.-W.: Development of Secure Access Control System for Location Information on Smart Phone. *KIISC Journal* 21(2), 139–147 (2011)
12. Sencha touch mobile programming, pp. 283–285. Acorn publishing (2011)

A Distributed Lifetime-Maximizing Scheme for Connected Target Coverage in WSNs

Duc Tai Le¹, Thang Le Duc¹, and Hyunseung Choo^{2,*}

¹ College of Information and Communication Engineering
Sungkyunkwan University, Korea
{ldtai, ldthang}@skku.edu

² Department of Interaction Science
Sungkyunkwan University, Korea
choo@ece.skku.ac.kr

Abstract. In this paper, we consider the problem of scheduling sensor activity to prolong the network lifetime while guaranteeing both discrete target coverage and connectivity among all the active sensors and the sink, called *connected target coverage* (CTC) problem. We proposed a distributed scheme called Distributed Lifetime-Maximizing Scheme (DLMS) to solve the CTC problem. In our proposed scheme, at first the source nodes are selected to ensure the target coverage. After that, energy-efficient paths to transmit the sensory data from source nodes to the sink will be built. The cost of the construction of the connected cover graphs is significantly reduced in comparison with the some conventional schemes since the number of targets (i.e., the necessary number of source nodes) is much smaller than the number of sensor nodes in the practical environment. In addition, the energy consumption is more balanced so that the network lifetime will be increased. Our simulation results show that DLMS scheme performs much better than the conventional schemes in terms of the network lifetime.

Keywords: Distributed scheme, coverage, connectivity, network lifetime, sensor activity scheduling, wireless sensor networks.

1 Introduction

In CTC problem, a number of targets with fixed locations are required to be continuously monitored (covered) in the field by a (large) number of randomly scattered sensors. A sensor, which is selected to be active for performing the monitoring task, is called a *source* sensor. The source sensors generate sensed data messages and send these messages to a sink node. In many assumptions used by the prior work, the transmission range of sensor does not enable all nodes to communicate with the sink directly. So, the sensed data could reach the sink via single-hop or multi-hop communication. A sensor node which does not perform monitoring task but needs to be activated to relay data is called a *relay* node. A sensor is called an *active* node if it is selected either as a source or as a relay or both. A sensor that is not active goes into an energy saving sleep state.

* Corresponding author.

In [1], Q. Zhao et al. proposed an algorithm for solving the connected target coverage (CTC) problem by scheduling sensors into multiple sets, each of which can maintain both target coverage and connectivity among all the active sensors and the sink. In this paper, scheduling sensor activity refers to determining the state of the deployed sensors to be either active (as source or relay or both) or sleep as well as their state durations. The main objective is to maximize the network lifetime by scheduling only a subset of sensor necessary to be active for guaranteeing both certain coverage and connectivity requirements. However, in their proposed schemes, a spanning tree covering the entire network must be rebuilt whenever they want to use another subset to ensure certain targets coverage and connectivity. This approach is not efficient, especially for the networks with a large number of nodes, because the cost for construction the whole network spanning tree is very high. Moreover, due to the utilization of static cover tree, the scheme could not achieve a good energy balancing between nodes in some cases. In these cases, some nodes may be selected as both relay and source node, or some nodes can be bottleneck nodes if they have many child nodes selected as source nodes.

In this paper, we proposed a new approach to reduce the protocol costs by selecting some source nodes to ensure the coverage first, and then finding energy efficient paths to transmit the sensing data from these selected source nodes to the sink. The cost of the construction of the connected cover graph is reduced significantly because the number of targets (i.e., the necessary number of source nodes) is much smaller than the number of sensor nodes. Moreover, our proposed scheme tries to balance the energy consumption of the sensors in the network to maximize the network lifetime.

The remainder of this paper is organized as follows. In the next section, we briefly describe the related works. The system model and our assumptions are provided in section 3. Our proposed schemes are presented in section 4. Section 5 evaluates the performance of our scheme. Finally, we conclude our work and give some future research directions in the last section.

2 Related Works

Scheduling sensor activity while guaranteeing a certain coverage requirement to prolong the network lifetime has been studied in the literature (see e.g., [5] for a survey and references therein). The coverage requirements that are commonly considered include (complete or partial) area coverage and complete target coverage. Barrier coverage is another type of coverage problem but the objective is to minimize the probability of undetected intrusion through the barrier [6], [7]. The problem of scheduling sensor activities for complete area coverage is addressed in [8]- [14]. Maintaining partial (but high) area coverage is discussed in [15]- [17]. Here, we briefly review some recent advances on scheduling sensor activity to cover discrete targets [1], [18]- [22].

In [18], an algorithm to find discrete sensor covers each providing complete area coverage is proposed without considering network connectivity. In [19], M. Cardei et al. modeled the discrete target coverage problem as a disjoint set cover problem which is proven to be NP-Complete. In [20], M. Cardei et al. extended their work in [19] and argued that the network lifetime can be further improved without the constraint that the selected set covers are disjoint, i.e., a sensor may appear in different covers. In [21],

M. Cardei further extended their work on discrete target coverage by assuming that sensors have adjustable sensing ranges. However, connectivity is not considered in [19]-[21]. In [22], M. Lu et al. schedule sensor activity by self-configuring sensing range, in the environment where both discrete target coverage and connectivity are satisfied. However, only sensing power is taken into account in their energy consumption model. Further, the heuristic proposed in [22] maintains network-wide connectivity which may not be necessary for target coverage. In fact, only those sensors along the routes carrying the sensed data are required to be active.

In [1], Q. Zhao et al. model the CTC problem, as a maximum cover tree (MCT) problem and prove that the MCT problem is NP-Complete. They then proposed a heuristic algorithm, called Communication Weighted Greedy Cover (CWGC) algorithm, and presented a distributed implementation of the heuristic scheme. The heuristic scheme has a high cost for rebuilding the network-wide spanning tree and a low network-lifetime for not balancing the energy consumptions. We are motivated by these above remaining problems. Our proposed scheme tries to reduce the protocol overhead and balances the energy between the sensors.

3 Preliminaries

3.1 Assumption and Definitions

The sensor field consists of a set of discrete targets with fixed locations, a number of randomly deployed sensors and a sink node. We assume that sensors are equipped with power controlled transceivers and non-rechargeable batteries with limited energy. Each sensor covers a disk centered at itself with a fixed sensing range as the disk radius. All sensors are assumed to have the same sensing range and the same maximum communication range.

The application requirements are to cover all the targets all the time and to send all the sensory data to the sink by a subset of the deployed sensors. All the sensors deployed in the WSN can reach the sink via single-hop or multi-hop communication. Each sensor is assumed to cover a fixed area and any target located in the area could be monitored by the sensor. It is also assumed that each target is covered by at least one sensor.

The data that are sensed and transmitted by the sensors are collected and processed by a sink node. If a sensor is selected to be active for performing the monitoring task, it generates data messages (e.g., quantized measurements) at a certain rate. Such a sensor is called a source sensor. Sensory data messages are transmitted to the sink via radio communication. Multiple-hop communication may be needed from a source to the sink. A sensor node which does not perform monitoring task but needs to be activated to relay data is called a relay node. A sensor is called an active node if it is selected either as a source or as a relay or both. A sensor that is not active goes into an energy saving sleep state. In this paper, scheduling sensor activity refers to determining the state of the deployed sensors to be either active (as source or relay or both) or sleep as well as their state durations.

The network lifetime is defined as the time period from the time when the network was set up until 1) one or more targets cannot be covered, or 2) a route cannot be found to send the sensory data to the sink.

3.2 Problem Statement

The CTC problem requires that all the targets in the sensor field are covered by a subset of sensors (coverage requirement) and all the targets are connected to the sink node through a subset of sensors by multi-hop paths (connectivity requirement). If any of the above requirements cannot be satisfied, we say that the deployed WSN reaches its lifetime. Our objective is to maximize the network lifetime of such a WSN.

The network lifetime can be increased by scheduling only a subset of sensors necessary to be active for meeting the application requirements. We found that the network can be divided into a number of sensor sets each of which can cover all the targets and can send all the sensory data to the sink. These sensor sets need not be disjoint, and are activated successively one by one. Based on these analyses, the CTC problem can be stated as follows. Given targets with known locations and an energy constrained WSN with sensors, it is required to schedule sensor activity so as to maximize the network lifetime subject to the conditions: 1) each target is covered by at least one source and 2) from each source to the sink, there must exist a route traversing through only the active sensors.

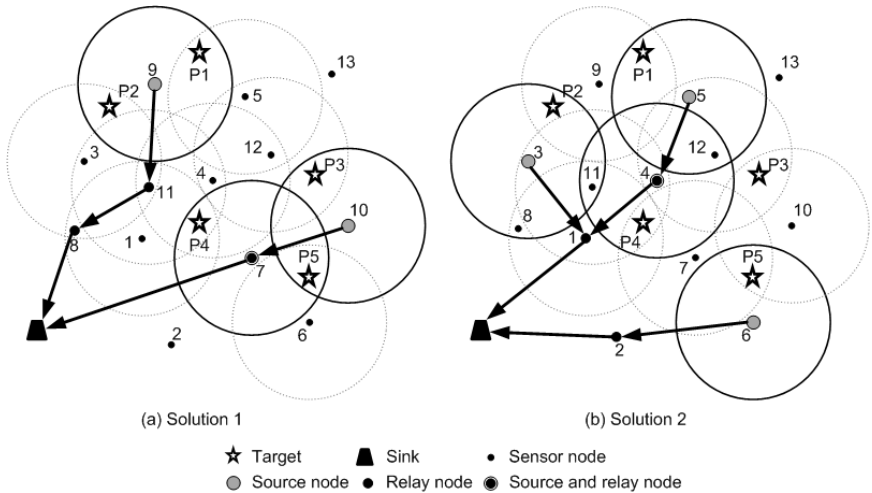


Fig. 1. Connected Target Coverage problem

We illustrate the CTC problem in Fig. 1. There are thirteen sensors, five targets and one sink in the sensor field. The sensors that can cover one or more targets are indicated by their circlessolid circles for active source sensors and others for sleep or relay sensors. Arrowed lines are used to denote the routes used to relay data from sources to the sink. Two possible solutions are illustrated in Fig. 1(a) and Fig. 1(b). This figure illustrates that only a subset of the deployed sensors is sufficient to carry out the functionalities of sensing targets and forwarding sensory data to the sink in the WSN. Different subsets can be used in different time intervals, call operational time interval (OTI).

3.3 Energy Consumption Model

We consider the model of sensor energy consumption which mainly considers the energy consumption for sensing and relaying data. A sensor consumes the energy depending on how much data is generated, transmitted and received [2]. In case of the target coverage scenario, each target needs to be monitored by the sensors continuously and the sensors generate data about each target. We assume that all source sensors have the same data generation rate for a target. In other words, all source sensors use the same sampling frequency, quantization, modulation and coding scheme for each target. Therefore, a fixed amount of bits, denoted by $BR(\tau)$, is generated by each source sensor for a target in an OTI τ . It also indicates that each sensor consumes different amount of energy according to the number of targets which the sensor covers. Let e_s and e_r denote the energy consumed for sensing and receiving one bit of data, respectively. Let e_{ij}^t denote the energy consumed by the sender s_i for transmitting one bit to the receiver s_j , it is followed as

$$e_{ij}^t = e_t + b \times d_{ij}^\alpha$$

where e_t and b are constants, d_{ij} is the Euclidean distance between sensor s_i and s_j and α is the path loss factor. For simplicity, we omit the sender and receiver id and use e_{trans} to represent e_{ij}^t .

Assuming that a graph $G(\tau)$ is constructed including a set of active sensors and a set of edges used to connect the selected active sensors and the sink for an OTI. The graph $G(\tau)$ has the following properties: 1) All the edges converged at the sink; 2) Each most outer node of the tree is a source sensor; 3) Each target can directly connect to at least one source in the graph. Such a graph is called as cover graph since it covers all the targets and, by definition, the graph is connected. Note that a sensor can act as a source node or relay node or both. In a cover graph, we call a sensor s_i a *descendant* of another sensor s_j if sensor s_i needs s_j to relay its data to the sink; and s_j is called the *ancestor* of s_i . Let $D(s, G(\tau))$ denote the number of sources among the descendants of sensor s in a given cover graph $G(\tau)$. Since all the sensed data should be relayed to the sink, a sensor s in the graph needs $(e_{trans} + e_r)BR(\tau)D(s, G(\tau))$ units of energy to relay the data from its descendants in an OTI.

Based on the above discussion, for the constructed cover graph $G(\tau)$ with set of sources $S_s(\tau)$ and set of relays $S_r(\tau)$, the energy consumption model for a sensor s in the sensor field is given by

$$E(s, G(\tau)) = \begin{cases} e_s BR(\tau) + e_{trans} BR(\tau), & \text{if } s \in S_s(\tau) \text{ and } s \notin S_r(\tau); \\ (e_{trans} + e_r) BR(\tau) D(s, G(\tau)), & \text{if } s \notin S_s(\tau) \text{ and } s \in S_r(\tau); \\ e_s BR(\tau) + e_{trans} BR(\tau) \\ + (e_{trans} + e_r) BR(\tau) D(s, G(\tau)), & \text{if } s \in S_s(\tau) \cap S_r(\tau); \\ 0, & \text{if } s \notin S_s(\tau) \text{ and } s \notin S_r(\tau); \end{cases}$$

4 Proposed Scheme

In our proposed scheme, first the source nodes are selected to ensure the target coverage. After that, energy-efficient paths to transmit the sensory data from source nodes to the sink will be built. The cost of the construction of the connected cover tree is significantly reduced in comparison with the original scheme in the target paper since the number of targets (i.e., the necessary number of source nodes) is much smaller than the number of sensor nodes. In addition, the energy consumption is more balanced so that the network lifetime will be increased. We take an example to illustrate the proposed scheme. The network consists of a sink, thirteen sensor nodes and five targets as shown in Fig. 2. The sensors that can cover one or more targets are indicated by their circles. This figure illustrates that there are several nodes can cover a common target.

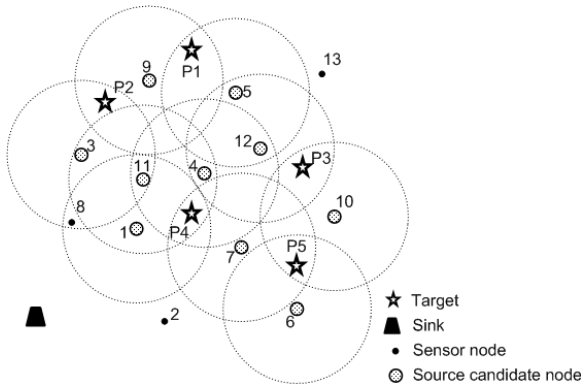


Fig. 2. Initial network

The proposed scheme composes of two stages. The initialization stage is executed once and the operation stage is executed after each operation time duration. In the initialization stage, a configuration information process is initiated at the sink and executed through all sensor nodes in the network. One INITIAL message is created at the sink and broadcast to all other nodes. At each node s_i , the path weight, the minimum weight of the route originated from itself to the sink through its neighbors, is computed by using this formula:

$$w_{s_i} = \min_{s_j \in S_n(i)} (w_{ij} + w_{s_j})$$

where $w_{ij} = \alpha \frac{e_{ij}^t}{E_r(s_i)} + \beta \frac{e_r}{E_r(s_j)}$ denotes the weight of link between two nodes, which considers both the distance and the residual energy of two vertices of link, $E_r(s)$ is the residual energy of sensor s , α and β are coefficients; w_{s_j} is the path weight of node s_j . The node then attaches this information to the INITIAL message and broadcasts it to other neighboring nodes which are further from the sink. As a result, the INITIAL message is propagated to all nodes in the network. Thanks to this, all nodes are assigned

a level and provided necessary information of its lower level neighboring nodes such as the level, the residual energy, targets in the sensing range. The INITIAL message is also used for synchronizing the clock locally among a node and its neighbors. After the initialization stage, each node knows its location, its level, its residual energy, its 1-hop neighbors information, location of targets within its sensing range, and the minimum weight of path originated from itself to the sink.

In the operation stage, a cover graph is built. The proposed scheme iteratively executes the process of building cover graphs and this process stops only when no new cover graph can be built (i.e., the network lifetime is reached). Each sensor can estimate its residual energy using the energy consumption model (given in session 3.3) to adjust the building process balancing the energy consumptions. There are three phases in each iteration of building a new cover graph.

In phase 1, each node which has uncovered targets in the sensing area computes a backoff time using this formula

$$T(s_i) = \frac{w_{s_i}}{|P_{s_i} - P_{s_i} \cap P'|}$$

where P_s is the set of targets that can be covered by s ; and P' is the set of covered targets. The rationale of this formula is to give higher priority (smaller $T(s_i)$) to sensors that have smaller communication cost and cover a larger number of uncovered targets. When $T(s_i)$ expires, if there are still any uncovered target in the sensing area, node s_i declares itself as a source node. The source node estimates its residual energy and updates its minimum weight of path originated from itself to the sink. It then broadcasts an advertising message to their neighbors which have higher level. During the backoff time, if a node receives the advertising messages from its neighbors, it updates $T(s_i)$ by removing the targets covered by advertising sensors and updating its minimum path weight.

Fig. 3 shows the result of this phase. Since node 7 is close to the sink and can cover two targets; it is selected as a source node. Although node 9 is far from the sink, but it can cover two targets; so it is also selected as a source node. After node 7 is selected as a source node, node 10 should use the path through node 6 and node 2 as it shortest

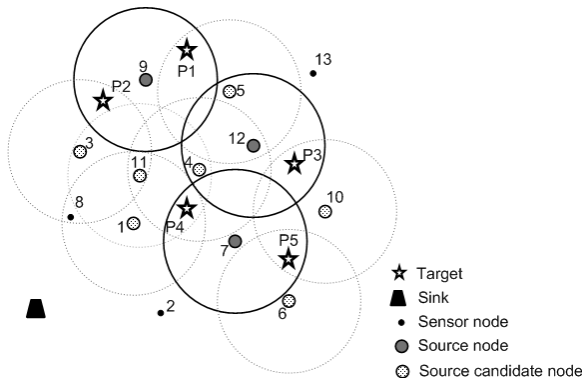


Fig. 3. Source selection phase

path toward to the sink. Because node 12 has a shorter path, through node 4 and node 1, it will be selected as a source node.

In phase 2, after the source nodes are selected, source nodes recursively notify the nodes in their shortest paths towards the sink to become the relay nodes. Each relay node then updates its minimum path weight by updating the residual energy of its neighbors and broadcasts its status, including estimated residual energy, to its neighbors. The result of this phase is shown in Fig. 4. The cover graph is dynamically built because a node can update its minimum path weight by updating the residual energy of its neighbors.

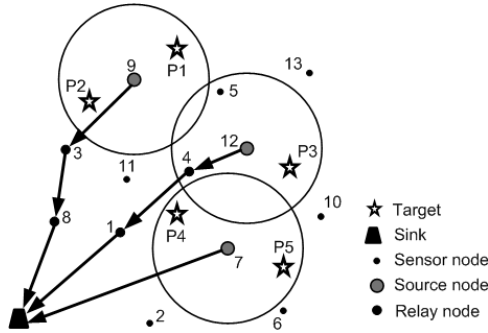


Fig. 4. Cover graph construction phase

Finally, in the last phase, the operation time duration for this cover graph will be estimated. Each cover graph operates during this fixed time duration if no sensor in the cover graph die (i.e. out of battery) before the time duration expires. Otherwise, the operation time duration of the cover graph is determined based on the sensor which has the least residual energy using this formula

$$\tau_x = \frac{E_r(s)}{E(s, G_x(\tau))} \tau$$

where $E(s, G_x(\tau))$ is the energy consumption of sensor s in a given cover graph G in the default operation time duration τ . A sensor which runs out of battery is called a dead sensor. An alive sensor which cannot find any route towards the sink without traversing a dead sensor is considered as an isolated sensor. The network will be updated by removing the dead and isolated sensor nodes.

5 Performance Evaluation

In this section, we majorly evaluate the performance of our proposed scheme and the CWGC (Communication Weighted Greedy Cover) scheme. A network is given with sensors and targets randomly deployed in a $100m \times 100m$ area. The sink node is placed in the middle of the area (at the point $(50m, 50m)$). The initial energy of each sensor is set to be $20J$; the value of various parameters are chosen to be $e_t = 50nJ/bit$,

$b = 100pJ/bit/m^4$, $\alpha = 4$, $e_r = 150nJ/bit$ and $e_s = 150nJ/bit$ [3]; and data is generated by each source node at the rate of 10 kbps. In the simulation, we assume that each sensor covers a disk centered at itself with a fixed sensing range as the disk radius. All sensors are assumed to have a similar sensing range $R_s = 20m$ and the communication range $R_c = 40m$.

To show the dominance of our scheme, we compare the performance of our proposed scheme with the CWGC scheme in our target paper. The CWGC is the only scheme that tries to minimize the energy consumption to maintain both target coverage and connectivity among all active sensors and the sink. However, this scheme has to rebuild the spanning tree for the entire network whenever another subset is needed to ensure coverage and connectivity. This approach is not efficient when the network has a large number of nodes since the costs for the whole network spanning tree construction is huge. Moreover, it could not achieve a good energy balancing between nodes for cases in which some nodes can be selected as both relay and source node, or some nodes can be bottleneck nodes if they have many children selected as source nodes. As a result, it cannot achieve a good approximation network lifetime since some nodes will deplete their energy more quickly than the other nodes.

Each value plotted on the curves is obtained from the results of one hundred random topologies.

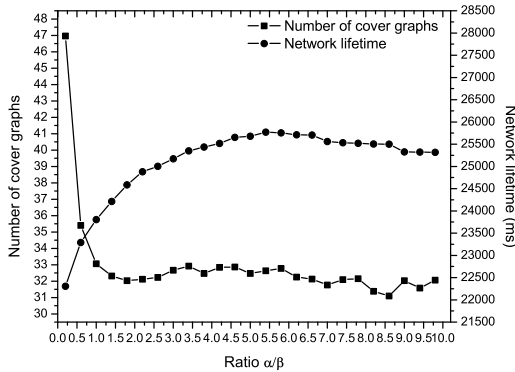


Fig. 5. Effect of coefficient factors α and β

5.1 Impact of Algorithm Parameters

First we study the impact of two parameters α and β . In which, α and β . In our simulation, to determine the effect of the coefficient factors, we measured the performance of the DLMS by varying the value of the coefficient factors with the same topology. One hundred sensor nodes and 20 targets are randomly scattered in the area. Fig. 5 shows the average lifetime and number of cover graphs achieved by DLMS with the initial operational time interval is set to 1000s. It can be seen that the network lifetime increased and the number of cover graphs decreased when the ratio of α and β increased. Because

sender node usually consumes energy much more than receiver node in a transmission, the energy consumption will more balance when we consider the sender higher than the receiver in the link-weight calculation.

Next we study the impact of operation duration on the performance of the DLMS scheme. The same scenario as chosen for the above simulation is used. Fig. 6 shows the average lifetime and number of cover graphs achieved by DLMS. As expected, although the average lifetime achieved by the scheme depends on the network topology more significantly than the initial operation time, it decreases slightly. The average number of cover graphs decreases when τ increases because the activated sensors have to consume more energy in each working interval.

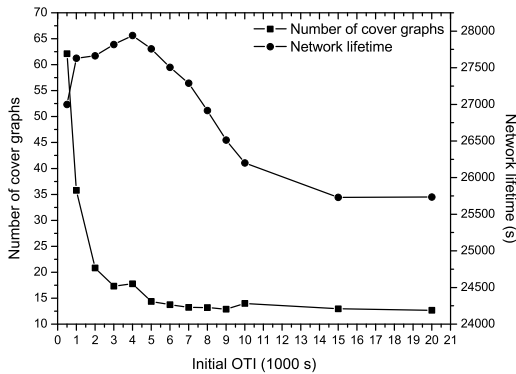
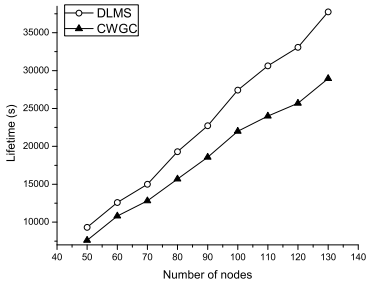


Fig. 6. Effect of initial operation time interval

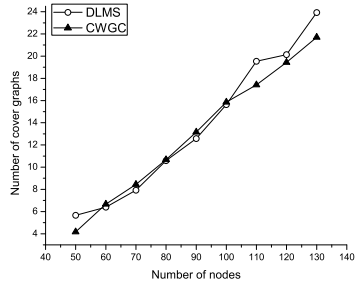
The more number of cover graphs are built, the more cost the scheme take for re-building the cover graph. In the following simulations, we fix the value of $\alpha = 2.7$, $\beta = 0.5$ and the initial OTI $\tau = 5000s$ which are a reasonable balance between performance and computation complexity.

5.2 Impact of Network Parameters

Next, we study the performance variation of the schemes under different sensor node densities. The number of targets is fixed at 20, and the number of nodes increases from 50 to 130. From Fig. 7(a) and Fig. 7(b), we can see that, the lifetime achieved by the DLMS scheme outperforms CWGC scheme while taking the same overhead (the same number of cover graphs). As the number of nodes increases, the network lifetime achieved by both schemes increase. It can also be observed that the DLMS scheme performs better than CWGC in the density network. The reason can be attributed to the non-balance in energy consumption in CWGC, in that case the sensors, which are closer to the sink, will deplete their energy earlier than the other ones. Thus, the hole will occur and the alive sensors can not connect to the sink.

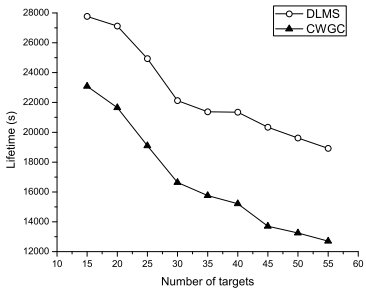


(a) Node's density vs. performance.

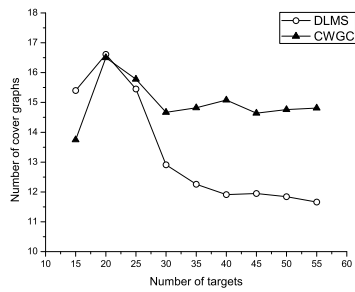


(b) Node's density vs. cost.

Fig. 7. Effect of the number of nodes



(a) Target's density vs. performance.



(b) Target's density vs. cost.

Fig. 8. Effect of the number of targets

Finally, we study the impact of varying the number of targets on the performance of the schemes. The number of targets is increased from 15 to 55 and the number of sensors is fixed at 100. The simulation results in Fig. 8 showed that DLMS scheme performs better than CWGC scheme in all cases. The performance of the schemes decreases as the number of targets increases, which is due to the reason that the number of source nodes should be increased and the amount of generated messages also increased. It is also observed in Fig. 8 that the overhead of the DLMS is smaller than CWGC in a network which has many targets.

6 Conclusion

In this paper, we have proposed a scheme to schedule the active time of sensors such that the active sensors can cover all the targets and formulate the efficient routes from each monitor node to the sink. From the simulation study, we also prove that the performance of the proposed scheme is better than the exiting scheme in terms of network lifetime.

The performance gain comes from balancing the energy consumption between all the nodes in the network. As part of our future work, we will study the CTC problem in the duty-cycled wireless sensor networks. In another research direction, we will extend our ideas to the other types of coverage, e.g., area coverage or barrier coverage.

Acknowledgment. This research was supported in part by MKE and MEST, Korean government, under ITRC NIPA-2012-(H0301-12-3001), WCU NRF (No. R31-2010-000-10062-0) and PRCP(2011-0018397) through NRF, respectively.

References

1. Zhao, Q., Gurusamy, M.: Lifetime Maximization for Connected Target Coverage in Wireless Sensor Networks. *IEEE/ACM Transactions on Networking* 16(6) (2008)
2. Hu, Y., Li, D., Wong, K., Sayeed, A.: Detection, Classification, Tracking of Targets. *IEEE Signal Processing Magazine* 19(2), 17–29 (2002)
3. Chang, J.-H., Tassiulas, L.: Maximum lifetime routing in wireless sensor networks. *IEEE/ACM Trans. Networking* 12(4), 609–619 (2004)
4. Cardei, M., Thai, M.T., Li, Y., Wu, W.: Energy-efficient target coverage in wireless sensor networks. In: *Proc. IEEE INFOCOM*, pp. 1976–1984 (2005)
5. Cardei, M., Wu, J.: Energy-efficient coverage problems in wireless ad hoc sensor networks. *Computer Commun.* 29(4), 413–420 (2006)
6. Meguerdichian, S., Koushanfar, F., Potkonjak, M., Srivastava, M.: Coverage problems in wireless ad hoc sensor networks. In: *IEEE INFOCOM* (2001)
7. Meguerdichian, S., Koushanfar, F., Qu, G., Potkonjak, M.: Exposure in wireless ad hoc sensor networks. In: *MOBICOM* (2001)
8. Wang, X., Xing, G., Zhang, Y., Lu, C., Pless, R., Gill, C.: Integrated coverage and connectivity configuration in wireless sensor networks. In: *Proc. ACM Int. Conf. Embedded Networked Sensor Systems (SenSys)*, pp. 28–39 (2003)
9. Zhang, H., Hou, J.C.: Maintaining Sensing Coverage and Connectivity in Large Sensor Networks. *Dept. Computer Sci., Univ. of Illinois at Urbana-Champaign, Tech. Rep.* (2003)
10. Berman, P., Calinescu, G., Shah, C., Zelikovsky, A.: Power efficient monitoring management in sensor networks. In: *WCNC* (2004)
11. Cardei, M., MacCallum, D., Cheng, X., Min, M., Jia, X., Li, D., Du, D.-Z.: Wireless sensor networks with energy efficient organization. *J. Interconnection Netw.* 3(3), 213–229 (2002)
12. Zhou, Z., Das, S., Gupta, H.: Connected k-coverage problem in sensor networks. In: *Proc. 13th Int. Conf. Computer Communications and Networks (ICCCN)*, pp. 373–378 (2004)
13. Tian, D., Georganas, N.D.: A coverage preserving node scheduling scheme for large wireless sensor networks. In: *Proc. 1st ACM Workshop on Wireless Sensor Networks and Applications* (2002)
14. Gupta, H., Zhou, Z., Das, S.R., Gu, Q.: Connected sensor cover: Self-organization of sensor networks for efficient query execution. *IEEE/ACM Trans. Networking* 14(1), 55–67 (2006)
15. Liu, Y., Liang, W.: Approximate coverage in wireless sensor networks. In: *Proc. IEEE Conf. Local Computer Networks, 30th Anniversary (LCN 2005)* (2005)
16. Wang, L., Kulkarni, S.S.: Pcover: Partial coverage for long-lived surveillance sensor networks. *Dept. Computer Sci. Eng., Michigan State Univ. Tech. Rep. MSC-CSE-0530* (2005)
17. Abrams, Z., Goel, A., Plotkin, S.: Set k-cover algorithms for energy efficient monitoring in wireless sensor networks. In: *IPSN, Berkeley, CA* (2004)

18. Slijepcevic, S., Potkonjak, M.: Power efficient organization of wireless sensor networks. In: Proc. IEEE Int. Conf. Communications (ICC), vol. 2, pp. 472–476 (June 2001)
19. Cardei, M., Du, D.-Z.: Improving wireless sensor network lifetime through power aware organization. *Wireless Networks* 11(3), 333–340 (2005)
20. Cardei, I., Cardei, M.: Energy-efficient connected-coverage in wireless sensor networks. *International Journal of Sensor Networks* 3(3), 201–210 (2008)
21. Cardei, M., Wu, J., Lu, M., Pervaiz, M.O.: Maximum network lifetime in wireless sensor networks with adjustable sensing ranges. In: IEEE Int. Conf. Wireless and Mobile Computing, Networking and Communications (WiMob) (2005)
22. Lu, M., Wu, J., Cardei, M., Li, M.: Energy-Efficient Connected Coverage of Discrete Targets in Wireless Sensor Networks. In: Lu, X., Zhao, W. (eds.) ICCNMC 2005. LNCS, vol. 3619, pp. 43–52. Springer, Heidelberg (2005)

Reducing Last Level Cache Pollution in NUMA Multicore Systems for Improving Cache Performance

Deukhyeon An¹, Jeehong Kim¹, JungHyun Han², and Young Ik Eom¹

¹ College of Information and Communication Eng., Sungkyunkwan University,
2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeong gi-do, 440-746, Korea
{novum21, jjilong, yieom}@ece.skku.ac.kr

² College of Information and Communication, Korea University, 145, Anam-ro,
Sungbuk-gu, Seoul, 136-701, Korea
jhan@korea.ac.kr

Abstract. Non-uniform memory architecture (NUMA) system has numerous nodes with shared last level cache (LLC). Their shared LLC has brought many benefits in the cache utilization. However, LLC can be seriously polluted by tasks that cause huge I/O traffic for a long time since inclusive cache architecture of LLC replaces valid cache line by *back-invalidate*. Many research on the page coloring, partitioning, and pollute buffer mechanism handled this cache pollution. But, there are no scheduling approaches considering I/O-intensive tasks in NUMA systems. To address the above problem, OS scheduling that reduces cache pollution is highly needed in NUMA systems.

In this paper, we propose a software-based mechanism that reduces shared LLC miss in NUMA systems. Our mechanism includes I/O traffic measurement and devil conscious scheduling. The experimental results show that LLC miss rate can be reduced up to 37.6%, and our approach improves execution time to 1.48%.

Keywords: Cache Pollution, Cache Performance, Last Level Cache, NUMA Scheduling, Task Characteristics, I/O Intensive Task.

1 Introduction

The studies about multicore scheduling in operating system have been researched since multicore processors are recently becoming more common. One of the main issues is reducing resources contention by using OS scheduling. Especially, cache contention problem is unceasingly considered by many researchers [1], [2], [3], [4], [5], [6], [7]. Advance of processor architecture makes more efficient cache access with shared last level cache (LLC). However, this causes cache pollution which can lead to performance degradation [6]. In particular, cache pollution can be seriously happened by tasks which raise huge I/O traffic for a long time [7]. Also, tasks whose I/O traffic is over max size of LLC lead to eviction of cache lines. Consequently, it makes other tasks difficult to utilize LLC efficiently.

In inclusive cache architecture included in modern processors, cache coherence mechanism is activated when cache pollution occurs. This lead to L1 and L2 cache pollution by *back-invalidate* [8]. L1, L2, and shared LLC (L3) latency for accesses to cache line is different among them. The upper level cache has a less latency than lower level cache [9]. If LLC which has high latency was being polluted, processor cannot use even L1 and L2 which has low latency. LLC pollution makes vicious cycle with loss of cache access cycles. LLC pollution has been treated by previous researches such as the page coloring, cache partitioning, cache pollute buffer [1], [3], [10]. However, there is none of approach in regard of OS scheduling to reduce shared LLC pollution considering I/O-intensive tasks in NUMA systems. In general, that system has more than two nodes which included shared LLC. On this account, it is very meaningful and important to reduce share LLC pollution for performance improvement in NUMA systems. In this paper, we deal with how to classify these I/O-intensive tasks which make serious LLC pollution. Also, we treat the scheduling that tasks can efficiently use cache.

The contributions of the paper are twofold. First, we suggested classification method for tasks which cause huge I/O traffic. Second, for reducing LLC pollution, we proposed a scheduling mechanism that such tasks can be dynamically migrated to designated node in NUMA systems.

The rest of this paper is organized as follows: Section 2 looks over background for clear understanding of this paper and motivation which inspired the proposed method. Section 3 explains I/O traffic measurement and devil conscious scheduling to solve the problem as mentioned above. Section 4 shows detailed implementation and evaluations resulted in each workload. Finally, in Section 5 we conclude this paper and discuss our plans for future work.

2 Background and Motivation

2.1 Cache Pollution in Shared LLC

Cache pollution is defined as the displacement of cache data by a useless one [6]. Cache pollution is commonly happened on shared LLC, which cause cache miss and performance degradation by cache line eviction and request operation of core [11]. Especially, this can be caused by I/O-intensive tasks that lead to huge I/O traffic for a long time such as multimedia application or downloading from web. We called these as *big I/O* tasks.

Below Fig. 1 illustrates the LLC miss rate of **Merge** when co-running with **PostMark** which is I/O-intensive task on the Dell PowerEdge T610¹: One in the same node and the other in another node. The LLC miss rate of the *same* graph higher than another one. On the contrary, the *another* graph shows lower than the *same* since they did not share the LLC. So, we can believe that cache pollution can be happened when co-running with *big I/O* task in the same node because of shared LLC. Ding, X., et al. drew similar result with **Merge** and **Grep** [7]. Therefore, we make sure that *big I/O* tasks can be isolated to designated node using OS scheduling in NUMA.

¹ Experimental Setup minutely explained in the chapter 4. Performance Evaluation.

Below Fig. 2 illustrates the LLC miss rate when co-running with two **Tars**, which put the Linux kernel source, respectively: One in the same node and the other in another node as the previous experiment. It represents similar miss rate between both. Unlike task which can efficiently use cache, *big I/O* task did not perform well in view of cache utilization. *Big I/O* task occur cache pollution when even executed alone. So, it is hard to reuse cache for tasks.

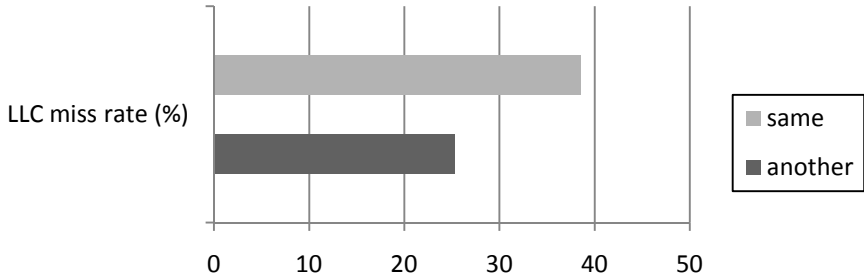


Fig. 1. LLC miss rate of mergesort when co-running with PostMark at the same time

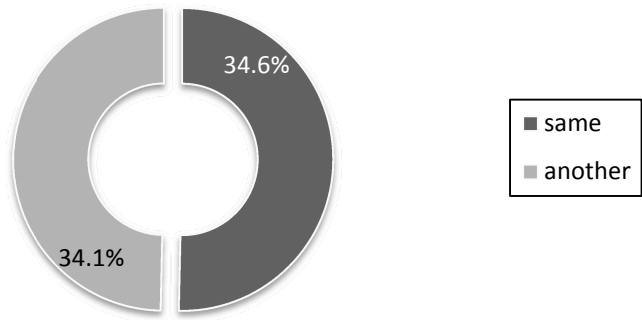


Fig. 2. LLC miss rate of tar in two other cases

2.2 Inclusive Cache Coherence

Inclusive cache architecture included in general processors. An inclusive cache means the lower cache has all the data that is available in the upper cache. In the Intel Nehalem microarchitecture, for example, it appears as the L3 cache has all the cache data in the L1 and L2 cache [11]. Thus, if cache line was evicted by LLC pollution, L1 and L2 cache line is also evicted by *back-invalidate* to maintain cache coherence as Fig. 3. Eventually, since LLC pollution causes the upper cache pollution, reducing LLC pollution is very important in NUMA systems. L1, L2, and L3 (LLC) latency for accesses to cache line is different among them. In the Nehalem, the latency to local of L1, L2, and L3 are 4, 10, and 38 cycles, respectively [9]. Therefore, shared LLC pollution even irritates fast cache access to L1 and L2 on processor.

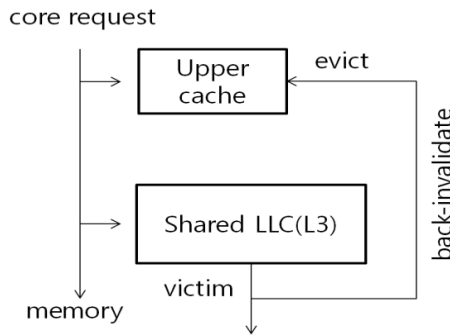


Fig. 3. Operation of back-invalidate in inclusive cache

3 LLC Pollution Reduction Mechanism

3.1 I/O Traffic Measurement

When co-running several tasks on multicore processors with shared LLC, it can cause cache pollution. To solve this problem, there are many researches how task affects the shared cache [4], [12], [13], [14]. Xie, Y., et al. showed an animalistic taxonomy which classify tasks by cache access and cache miss rate using the hardware profiling. There are four kinds of type in the taxonomy such as Turtle, Sheep, Rabbit, and Devil [14]. In this paper, we suggest a classification method with I/O traffic. As illustrated in Fig. 4, the task whose I/O traffic overflows max size of LLC causing serious cache pollution is named a devil, or which I/O traffic is less than max size of LLC is named a normal. The reason why the tasks classify only two kinds is as follows. First, as we already seen in Fig. 2, there is no significance of find-grained classification in terms of reducing miss.

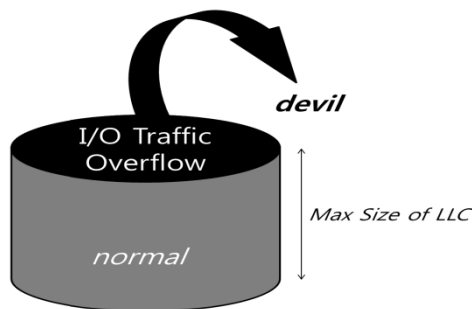


Fig. 4. I/O traffic measurement with max size of LLC

Second, the proposed scheduling algorithm has only to decide whether the scheduler migrate task to designated node or not. The scheduling algorithm will be explained in 3.2. Fig. 4 illustrates this I/O traffic measurement. There are various I/O

devices like storage, network, etc., which can be the factor of cache pollution. Many OS manage these devices by its own way. The Linux manages all the devices as files and leaves log file per task in */proc* directory when I/O traffic of each device occurs in the kernel. Even though system call such as *read()*, *write()*, *send()*, and *receive()* can be used in I/O operation, we limit storage-related I/O *read()* and *write()* to help easily explain the mechanism². *do_io_accounting()* is a internal kernel function to leave storage-related I/O log file in */proc/PID/IO*, which counts I/O traffic related to *read()*, *write()* system call by byte [15]. I/O traffic can be measured by them in real time.

$$MA = (\text{Previous IO traffic} \times (1 - \alpha)) + (\text{Current IO traffic} \times \alpha). \quad (1)$$

We can periodically obtain current I/O traffic at regular time interval using *rchar*, *wchar* variables of task through I/O accounting scheme provided by the kernel. In Formula (1), is a weighted value to reflect current I/O traffic into MA (Moving Average). The MA is the key that decides whether a task is devil or normal. The bigger a value α , the more recent I/O traffic size is sensitively reflected. On the contrary, the smaller a value α , the more previous I/O traffic size is reflected. Therefore, de-pending on how a value α sets, it can be adjusted in I/O traffic fluctuation. With the moving average, the MA moves exponentially preventing itself from a sudden fluctuation when rapid I/O traffic change occurs in a short time. Also, this prevents tasks from migrating frequently between nodes, which causes performance degradation by overhead. In Fig. 4, task is classified as devil if the value of MA excesses max size of LLC, or else it is classified normal when the value of MA is less than max size of LLC.

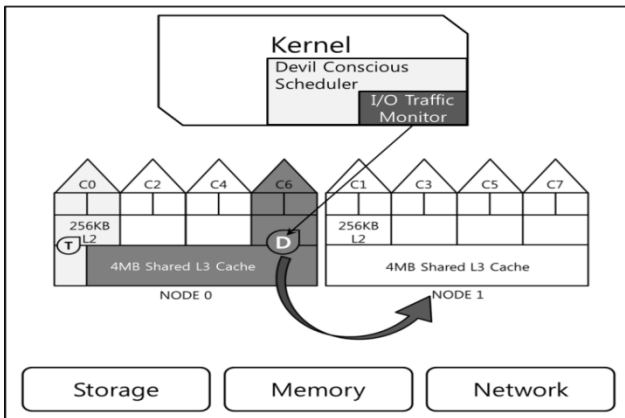


Fig. 5. Overall system architecture of devil conscious scheduler

² In this paper, we employ *rchar* and *wchar* defined in */include/linux/task_io_accounting.h* to measure all of I/O traffic from page cache and physical storage.

3.2 Devil Conscious Scheduling Method

All of tasks have a *task_struct* in the Linux operating system. In our method, it has a flag additionally, which called *Is_Devil* flag. Once a certain task is turned out as a devil by I/O traffic measurement, *Is_Devil* flag of corresponding task is changed to 'ON' state. When state changed, this task is regarded as enough to disturb other task which can efficiently use shared LLC by serious cache pollution. For this reason, it is needed for migration to designated node to reduce cache pollution. We called this node as a devil pool. Designated node can be selected as one of nodes in NUMA system. Above Fig. 5 illustrates overall system architecture of devil conscious scheduler. Since shared LLC exists generally only one at each node, it takes the identical effect on shared LLC no matter where the task runs on which core.

I/O traffic can be decreased under max size of LLC or stopped according to task characteristics while a task is running. In this case, the task is reclassified by normal and *Is_Devil* flag is changed to 'OFF' state. So, corresponding task can be migrated to normal pool from devil pool. The normal pool is all nodes except a designated node as a devil pool. Like this, proposed method in this paper can do dynamic task migration through current I/O traffic of task. Below Algorithm 1 shows pseudo-code of devil conscious scheduling.

Algorithm 1. Devil Conscious Scheduler(DCS)

While continuously monitoring **do**

1. I/O traffic measuring of by regular period

 Get Current I/O traffic

if MA > LLC max size **then**

Is_Devil flag of current task ← ON

else

Is_Devil flag of current task ← OFF

end if

2. Migrating the task to designated node

if *Is_Devil* flag of current task is ON

 task migration to designated node in NUMA (devil pool)

 migration pages of previous node to designated node

else *Is_Devil* flag of current task is OFF

 task migration to normal node in NUMA (normal pool)

 migration pages of designated node to current node

end if

end while

4 Performance Evaluation

We implemented prototype of I/O traffic measurement and devil conscious scheduling in user space to evaluate our suggested method. As described in 3.1, we used */proc/PID/io* to obtain I/O traffic related storage of various I/O devices.

To reflect I/O traffic quickly, I/O traffic measuring monitor checks traffic by one second as *iostat* which is well-known physical storage I/O monitoring tool. The α set 0.8 in formula (1) because it is optimal value by 4.2.2 experiment. Completely Fair Scheduler (CFS) which is the modern Linux scheduler runs by default [16]. In addition, we set devil conscious scheduler that the task has to be migrated to any core in a designated node when a certain task is classified as devil by I/O monitor.

The evaluation of our prototype was carried out in a set of micro-benchmarks. All of measurement was evaluated by *perf* which is the Linux kernel-based performance analysis tool using the Hardware Performance Counter (HPC). The measurement was performed by ten time and the results were averaged out. The system used the evaluation is a Dell PowerEdge T610 which have two nodes. Its specification was minutely listed in Table 1.

Table 1. Specification of the Dell PowerEdge T610

Component	Specification
L1Data Cache	4 x 32KB, 8-way
L1 Instruction Cache	4 x 32KB, 4-way
L2 Cache	4 x 256KB, 8-way
L3 shared Last Level Cache	4MB, 16-way
Memory	2GB in each node

4.1 α Sensitivity in MA

Optimal α in formula (1) is a meaningful factor in performance of devil conscious scheduling. So, we identified the result of cache miss rate when changing α value. To find out this, we adjusted α value by 0.2, 0.4, 0.6, and 0.8, respectively with **merge** and **PostMark**. As a result, we obtained lowest cache miss rate when set α by 0.8 as Fig. 6.

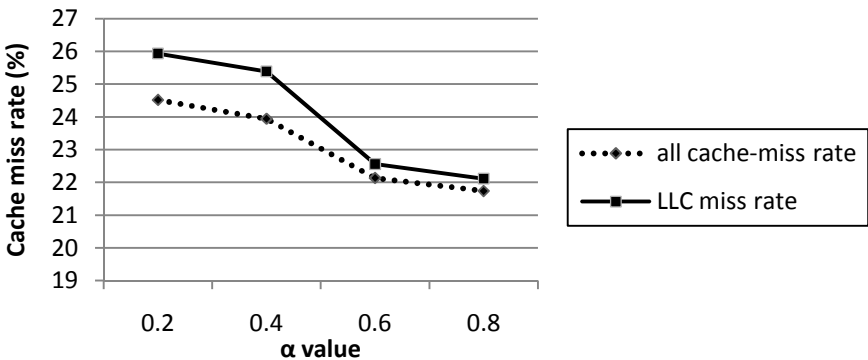


Fig. 6. Cache miss rate of mergesort

4.2 Micro-benchmarks

Micro-benchmarks were divided by two groups. We identified how corresponding workloads can efficiently utilize shared LLC by suggested methods. One group is comprised of memory-intensive workloads; the other is comprised of *big I/O* workloads [17], [18].

- **Gobmk** is an artificial intelligence workload related with game playing. It is included in a set of SPEC CPU2006.
- **Hmmer** is statistical models of multiple sequence alignments, which are used in computational biology to search for patterns in DNA sequences. It is also included in a set of SPEC CPU2006.
- **Merge** sorts an array size of 2MB with recursive 2-way mergesort algorithm.
- **PostMark** is I/O-intensive workload. It conducts file access and operation with four hundred thousand files whose sizes 128KB.
- **Tar** is also I/O-intensive workload. We run it to bind the Linux kernel source with bzip2 compression option.

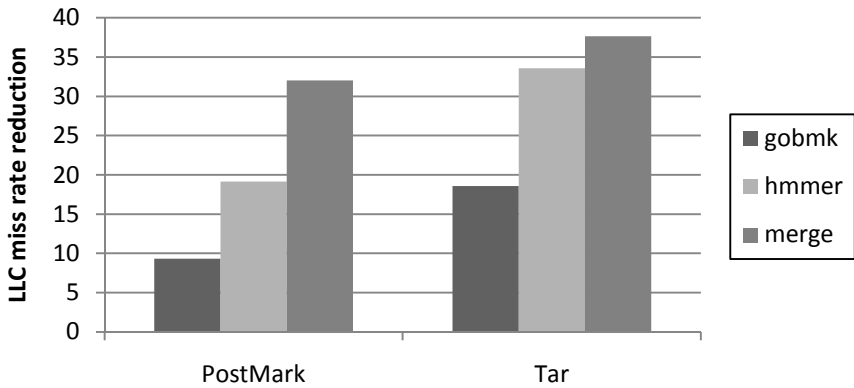


Fig. 7. LLC miss rate reduction of each workload

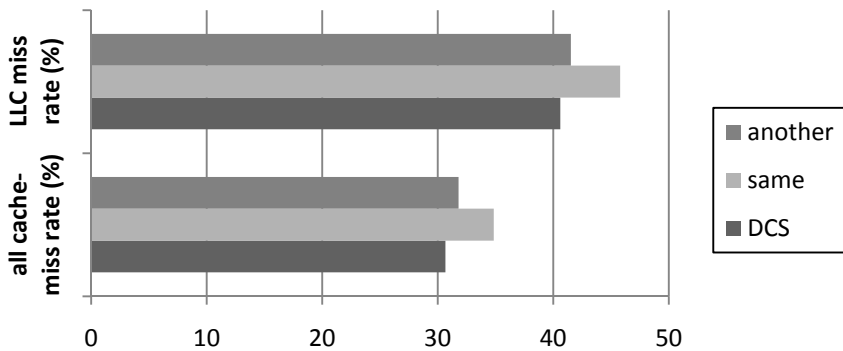


Fig. 8. Negative influence of LLC miss rate on the upper caches

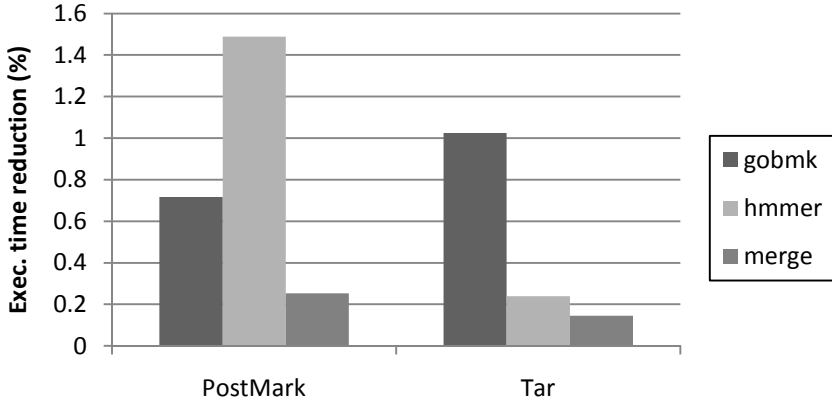


Fig. 9. Execution time reduction of each workload

For co-running with **PostMark** and **Tar**, respectively, Fig. 7 illustrates the LLC miss rate of all memory-intensive workloads. The *another* is the case that each memory-intensive workloads run on another node with I/O-intensive workloads. The *same* is when run them on the same node. *DCS* represents when devil conscious scheduler was applied. In the result of our experiment, *DCS* reduced the LLC miss rate of **Hmmer** by 19.1% when co-running with **PostMark**. This improved the execution time reduction by 1.48% as illustrated in Fig. 9. Especially, The **Merge** reduced by 37.6% when co-running with **Tar**. *DCS* reduced average LLC miss rate and cache miss rate of all references about 20.7%. Also, it improved average execution time about 0.87%

Fig. 8 illustrates the LLC miss rate and miss rate of all cache references of **Gobmk** when co-running **PostMark**. When *DCS* was applied, all cache miss rate reduced in proportion to the LLC miss rate. Since *DCS* reduces LLC miss rate, *back-invalidate* do not occur in the upper caches.

5 Conclusion

In this paper, we identified why the shared LLC is important in NUMA system environment and how *big I/O* tasks causes serious LLC pollution. To resolve this problem, we proposed I/O traffic measurement and devil conscious scheduling. *Big I/O* task, which also called devil, can be classified by I/O traffic measurement. Devil conscious scheduling can assign tasks into designated node to take advantage of cache utilization. Our experimental results showed that LLC miss rate are reduced up to 37.6%, and it improves execution time reduction up to 1.48%. We identified our proposed mechanism can improve cache performance as only software-based approach. As future work, we can also combine other cache improvement scheme for better performance.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0025971).

References

1. Azimi, R., Tam, D., Soares, L., Stumm, M.: Enhancing operating system support for multicore processors by using hardware performance monitoring. In: ACM Special Interest Group on Operating System, pp. 56–65 (2009)
2. Blagodurov, S., Zhuravlev, S., Fedorova, A., Kamali, A.: A case for NUMA system-aware contention management on multicore systems. In: 19th Parallel Architectures and Compilation Techniques, pp. 557–558 (2010)
3. Kim, J., Kim, J., Ahn, D., Eom, Y.: Page Coloring Synchronization for Improving Cache Performance in Virtualization Environment. In: 11th Computational Science and its Applications, pp. 495–505 (2011)
4. Dey, T., Wang, W., Davidson, J.W., Soffa, M.L.: Characterizing multi-threaded applications based on shared-resource contention. IEEE Performance Analysis of Systems and Software, 76–86 (2011)
5. Chandra, D., Guo, F., Kim, S., Solihin, Y.: Predicting Inter-Thread Cache Contention on a Chip Multi-Processor Architecture. In: 11th IEEE High-Performance Computer Architecture, pp. 340–351 (2005)
6. Soares, L., Tam, D., Stumm, M.: Reducing the Harmful Effects of Last-Level Cache Polluters with an OS-Level, Software-Only Pollute Buffer. IEEE MICRO Architecture, 258–269 (2008)
7. Ding, X., Wang, K., Zhang, X.: SRM-buffer: An OS buffer management technique to prevent last level cache from thrashing in multicores. In: 6th ACM European Conference on Computer Systems, pp. 243–256 (2011)
8. Jaleel, A., Borch, E., Bhandaru, M., Simon, C.S., Emer, J.: Achieving Non-Inclusive Cache Performance with Inclusive Caches: Temporal Locality Aware (TLA) Cache Management Policies. In: 43rd IEEE MICRO Architecture, pp. 151–162 (2010)
9. Molka, D., Hackenberg, D., Schöne, R., Müller, M.S.: Memory Performance and Coherency Effects on an Intel Nehalem Multiprocessor System. In: 18th IEEE Parallel Architectures and Compilation Techniques, pp. 261–270 (2009)
10. Zhuravlev, S., Blagodurov, S., Fedorova, A.: Addressing shared resource contention in multicore processors via scheduling. In: 15th ACM Architectural Support for Programming Languages and Operating Systems, pp. 129–142 (2010)
11. Qian, B., Yan, L.: The Research of the Inclusive Cache used in Multi-Core Processor. IEEE Electronic Packaging Technology & High Density Packaging, 1–4 (2008)
12. Tam, D., Azimi, R., Soares, L., Stumm, M.: RapidMRC: approximating L2 miss rate curves on commodity systems for online optimizations. In: 14th ACM Architectural Support for Programming Languages and Operating Systems, pp. 121–132 (2009)
13. Knauerhase, R., Brett, P., Hohlt, B., Li, T., Hahn, S.: Using OS Observations to Improve Performance in Multicore Systems. IEEE MICRO Architecture, 54–66 (2008)
14. Xie, Y., Loh, G.H.: Dynamic Classification of Program Memory Behaviors in CMPs. In: 2nd Workshop on Chip Multiprocessor Memory Systems and Interconnects (2008)

15. The Linux Kernel Archives: THE proc FILESYSTEM, <http://www.kernel.org/doc/Documentation/filesystems/proc.txt>
16. Blagodurov, S., Fedorova, A.: User-level scheduling on NUMA system multicore systems under Linux. In: 13th Annual Linux Symposium (2011)
17. Jaleel, A.: Memory Characterization of Workloads Using Instrumentation-Driven Simulation, <http://www.jaleels.org/ajaleel/workload/SPECanalysis.pdf>
18. SPEC CPU2006 Documentation, <http://www.spec.org/cpu2006/Docs/>

The Fast Handover Scheme for Mobile Nodes in NEMO-Enabled PMIPv6

Changyong Park¹, Junbeom Park², Hao Wang¹, and Hyunseung Choo^{1,*}

¹ College of Information and Communication Engineering
Sungkyunkwan University, Korea
{gspcy, wanghao}@skku.edu, choo@ece.skku.ac.kr

² Department of Interaction Science
Sungkyunkwan University, Korea
Junbeompark43@gmail.com

Abstract. In Proxy Mobile IPv6 (PMIPv6), the serving network provides mobility management on behalf of the mobile node (MN). Moreover, the MN need not have a mobility support stack and it can handover faster in PMIPv6 than in MIPv6 with less packet loss. However, when several nodes in the same MAG (Mobile Access Gateway) handover to another MAG at the same time, a lot of signaling costs occur as each cost is incurred on each node. NEMO-BSP (Network Mobility-Basic Support Protocol) developed by the IETF (Internet Engineering Task force), enables the mobile network to which MNs are connected to perform only one handover signaling for all the MNs, reducing handover signaling cost, but NEMO-BSP cannot be immediately applied to PMIPv6. Therefore, several schemes have been proposed to combine PMIPv6 and NEMO. These schemes suggest the way the mobile network handovers in the PMIPv6 domain and focuses on the reduction of the signaling cost. However, they are often unconcerned about the signaling cost of mobile nodes that move between the mobile network and the PMIPv6 domain. In this paper, we propose fast handover scheme to enhance MN's handover performance, when the MN handovers between the mobile network and the PMIPv6 domain

Keywords: PMIPv6, Proxy Mobile IPv6, Fast Handover, NEMO, Network Mobility.

1 Introduction

Due to the rapid evolution of mobile devices and wireless network access technologies, there is a rapid increase in the number of mobile device users. Moreover, people demand Internet access anytime and anywhere. They spend a lot of time on the way, by car or subway using wireless Internet and this in turn needs a protocol which supports mobility management.

* Corresponding author.

Mobile IPv6 (MIPv6) [1], the host-based mobility management protocol developed by the Internet Engineering Task Force(IETF) enables mobile nodes handover to maintain a session in a wireless environment. However, a host-based mobility management protocol needs the mobility protocol stack in the MN [2]. On the other hand, Proxy Mobile IPv6 (PMIPv6) [3], developed by the IETF NETLMM working group, provides network-based mobility for the MNs. Local Mobility Anchor (LMA) and Mobile Access Gateway (MAG), newly introduced in PMIPv6, perform a handover signaling on behalf of an MN. The MN do not change its IP address during handover in the same PMIPv6 domain as it creates its own IP addresses by using the same Home Network Prefix (HNP) which it has received from the LMA. Therefore, MNs need not have mobility protocol stack and they do not have to consume energy for handover signaling. Moreover, PMIPv6 reduces the packet loss as the handover delay is shorter than MIPv6 [2][4].

In PMIPv6, a handover is performed for each MN that moves from an MAG domain to another. If many nodes within the same MAG domain move to another MAG at the same time then, the handovers occur for each node. It causes the handover signaling cost to increase and to degrade the overall quality of the network services by consuming network resources. In order to prevent this, a scheme has been proposed to enable handover for all MNs that handover to the same MAG at the same time [5], but it also causes an additional cost to maintain a group of nodes.

The Network Mobility Basic Support Protocol (NEMO-BSP) [6] developed by the IETF is a protocol that is able to provide mobility for a mobile network and this includes a set of MNs which were developed ahead of PMIPv6. When the mobile network accesses to an Access Router (AR), Mobility Router (MR) performs handover signaling on behalf of all the MNs in the mobile network. The MNs in the mobile network can maintain their session with their correspondent node that is outside during handover. Each node creates their own address by using Mobile Network Prefix (MNP) received from MR. If the mobile network takes handover then, the MNs do not realize that they have taken handover as their addresses do not change.

In this paper, we propose a scheme that enhances the handover performance when MNs take handover between the mobile network and PMIPv6 domain. Our proposed scheme takes fast handover, reduces unnecessary signaling procedures and this leads to a reduction in the handover signaling cost and delay.

The remainder of this paper is organized as follows. In Section 2, we introduce a detailed explanation of NEMO-BSP and PMIPv6 as well as the scheme that combines these two protocols. Our proposed scheme that improves the scheme is introduced in section 2 and it is presented in detail in Section 3. Section 4 discusses the performance evaluation and results. Finally, we conclude this paper.

2 Related Work

2.1 Proxy Mobile IPv6

PMIPv6, network-based mobility support protocol, enables MNs handover without the participation in any mobility signaling [2][3]. The handover signaling for the MN

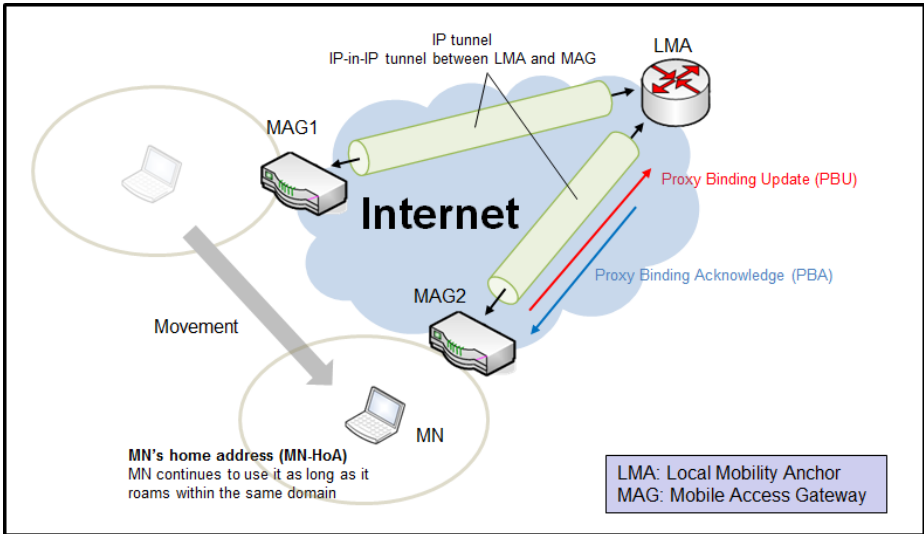


Fig. 1. PMIPv6 Architecture

is performed by LMA and MAG, newly introduced in PMIPv6 on behalf of the MN, and the MN need not have the mobility support protocol stack as the MN has no hand in the handover signaling.

The architecture of PMIPv6 is shown in Figure 1[2]. The MN connected to MAG1 creates its own IP address by using HNP that is received from LMA. This HNP is the MN's own address space which is never changed while the MN stays in the PMIPv6 domain, and the MN does not change its IP address even though it handovers to the MAG2 domain. Therefore, in PMIPv6 the handover delay is reduced without Duplicate Address Detection (DAD) process when an MN handovers in the PMIPv6 domain.

The handover procedure in PMIPv6 is a signaling procedure on the scenario where an MN takes handover. If many MNs take handover at the same time then, a hand over procedure occurs for each MN. In a case when many MNs move together in a group, this is inefficient. It wastes more network resources and as a result, it causes the quality of the entire Internet to decline.

2.2 Network Mobility

In NEMO-BSP, a handover procedure is performed for the entire network when the mobile network takes handover. When the mobile network is in its home network, MR creates its own Home Address (HoA) by using the prefix that is provided by the Home Agent (HA). HoA is the MR's own address space and it does not change even though the mobile network moves to another network. All the MNs in the mobile network create their own IP address by using the MNP that is provided by the MR. If the mobile network leaves its home and accesses to the foreign network then, the MR

creates its Care-of Address (CoA) by using the prefix that is provided by the foreign network and sends Binding Update (BU) message to the HA. The HA received the BU message sends Binding Acknowledgement (BA) message to the MR after creating the cache entry to bind the HoA to the CoA of the MR. After the bidirectional tunnel is created, all the packets head for an MN in the mobile network and they are transmitted through the tunnel. MNs in the mobile network do not recognize that they have moved as their IP addresses have not changed. The handover signaling procedure of NEMO-BSP seems similar with the handover signaling procedure of MIPv6 [12].

2.3 Network Mobility in PMIPv6

NEMO-BSP has the advantage of reducing the signaling cost as it performs a handover for all MNs in the mobile network with just a single handover signaling. The handover of the network is transparent to the MNs as the MNs do not participate in handover signaling. However, it is impossible to combine NEMO-BSP and PMIPv6 to be intact. The reason is that MNs create their own IP addresses by using MNP in the mobile network and by using HNP which they have received from LMA in the PMIPv6 domain. This is contrary to the policies of PMIPv6 where, IP address of an MN does not change when the MN takes handover in the PMIPv6 domain.

[7] proposes NEMO-enabled PMIPv6 (N-PMIPv6) to combine NEMO-BSP and PMIPv6. N-PMIPv6 supports the complete network-based mobility for an MN about the MN's handover as well as the mobile network's handover. In N-PMIPv6, the MR and MNs create their IP addresses by using HNP which is received from LMA. When the mobile network moves, the MR is treated as an MN. The MAG performs handover signaling with the LMA on behalf of the MR. When an MN handovers between the mobile network and the PMIPv6 domain, the MR acts as an MAG and it performs handover signaling with the LMA on behalf of the MN. Both the MR and MN create their own IP address by using HNP that is received from LMA.

In N-PMIPv6, the BCE is extended. A new field, M flag, which shows whether the AR that the MN is connected with is a fixed MAG or an MR. Figure 2 shows the handover signaling procedure in N-PMIPv6. If the mobile network attaches to the MAG domain, the MAG announces the new access of entity by sending PBU message to the LMA. The LMA which has received the PBU makes a binding cache entry for the entity, and the M flag value becomes 'NO'. This is due to the MAG that the MR has accessed is fixed. The LMA transmits the PBA message in response to the PBU message that includes HNP for the MR. The MAG sends the RA message to the MR. Then, a bidirectional tunnel is established between the LMA and MAG.

The MR performs handover signaling with the LMA on behalf of the MN. The PBU and PBA messages are transmitted through the MAG. The LMA which has received the PBU message updates its BCE and the M flag value becomes 'YES' as the MN is connected to an MR. The LMA transmits the PBA message to the MR that includes HNP for the MN. Then, the bidirectional tunnel between the LMA and the MR is established. The MR completes the handover by sending the RA message to the MN that contains the HNP of the MN.

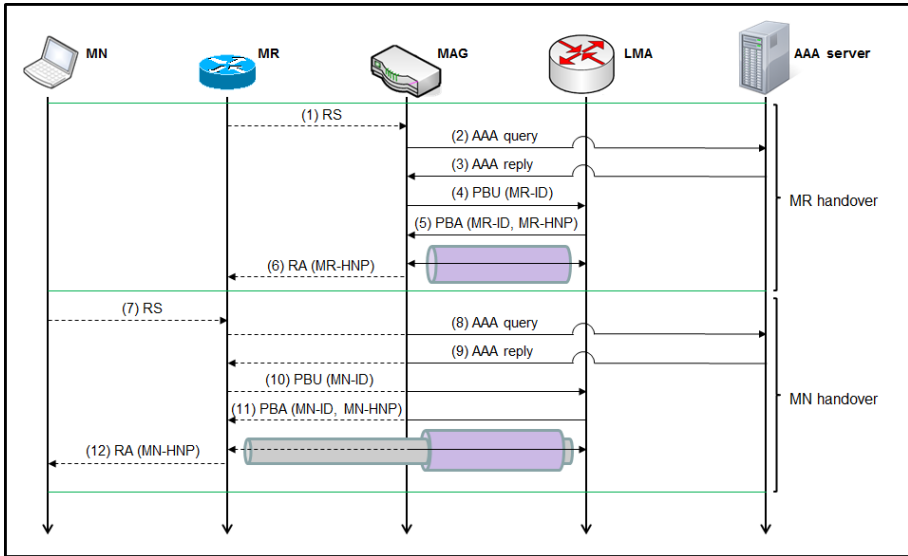


Fig. 2. N-PMIPv6 handover procedure

3 Proposed Scheme

3.1 Overview of the Proposed Scheme

N-PMIPv6 gives support to MNs to handover between the mobile network and PMIPv6 domain by applying NEMO-BSP to PMIPv6. The handover follows the procedure of the standard PMIPv6. So, it still has the PMIPv6 handover latency and packet loss problem. [13] proposes a fast handover scheme in PMIPv6 domain where it reduces the handover latency by omitting AAA authentication procedure between MAG and AAA server when an MN accesses to MAG.

Our proposed scheme reduces the signaling cost and handover latency by omitting AAA authentication procedure when an MN moves between a mobile network and PMIPv6 domain. In order to achieve this, MAGs need to know information on MNs in the mobile network. Figure 3 shows handover signaling procedure in our proposed scheme.

3.2 Fast Handover Scheme for Mobile Nodes

In Figure 3, the MR passes Node Information (NI) message which contains information on all MNs in it after it handovers to the MAG. The MAG stores the information in the table. Since then, when one of the MNs in the mobile network takes handover to the MAG, the MAG refers to the table and it figures out that the MN was in the mobile network. The MAG omits the authentication procedure with the AAA server and it performs handover signaling with the LMA right away. When one of the MNs in the MAG domain takes handover the mobile network, the AAA authentication procedure is not required as the mobile network is in the MAG domain

with the MN and the authentication on the MN is already done. After the handover is completed, the MR passes the information on the MN to the MAG. The MAG in turn stores the information in the table in case MN moves out to omit the authentication procedure.

The information of MN passed from the mobile network to the MAG includes the ID of the MN. When the mobile network takes handover, the MR sends IDs of all the MNs to the MAG to which the MR has accessed. When an MN takes handover from the MAG domain to the mobile network, the MR sends an ID of the unisonous MN to the MAG. Only MN's ID is required as MAGs identify MNs by their ID when they access to an MAG.

When a mobile network takes handover in the PMIPv6 domain and an MN takes handover from the MAG domain to the mobile network, the node information may increase signaling the cost. However, the signaling cost due to the AAA authentication procedure is reduced. Particularly, it is more efficient when the movement of the MNs is frequent between the mobile network and the MAG domain. Another advantage is to reduce the packet loss through fast handover. The NI message for the updation of the table contains the information of MNs in MAG and it is irrelevant to the handover time as it is transmitted after the handover procedure.

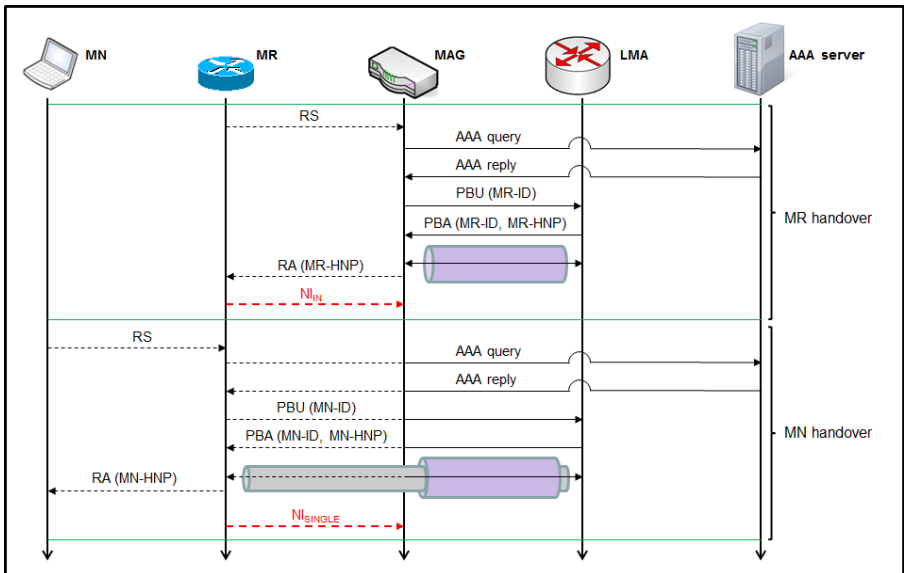


Fig. 3. Handover procedure of the proposed scheme

4 Performance Evaluation

4.1 Analysis Scenario and Notations

Figure 4 is a scenario for performance evaluation. The mobile network takes the handover between MAG domains at a distance of time. In each MAG domain, at this

point, MNs take handover. Some of the MNs in the mobile network take handover to an MAG domain while some of the MNs in the MAG domain take handover to the mobile network. The notations that are used in the performance evaluation are tabulated in Table 1.

Table 1. Notations

$H_{MAG-LMA}$	The number of hops between MAG and LMA	5	n_i	The number of nodes attached to MR	
$H_{MAG-AAA}$	The number of hops between MAG and AAA	5	n_f	The number of message failures over the wireless link	
H_{MR-MAG}	The number of hops between MR and MAG	1	p_f	The wireless link failure probability	
H_{MN-MAG}	The number of hops between MN and MAG	1	$D_{wireless}$	The propagation delay for the wireless link	10ms
H_{MN-MR}	The number of hops between MN and MR	1	D_{wired}	The propagation delay for the wired link	2ms
S_{RS}	The RS size	70 bytes	T_{RS}	The arrival delay of RS message	
S_{RA}	The RA size	80 bytes	T_{AAA}	The delay of the AAA procedure	
S_{AAA}	The AAA size	100 bytes	$T_{PBU-OUT}$	The arrival delay of PBU message from MAG to LMA	
S_{PBU}	The PBU size	96 bytes	T_{PBU-IN}	The arrival delay of PBU message from MR to LMA	
S_{PBA}	The PBA size	96 bytes	$T_{PBA-OUT}$	The arrival delay of PBA message from LMA to MAG	
S_{NI_HEADER}	The size of NI header	40 bytes	T_{PBA-IN}	The arrival delay of PBA message from LMA to MR	
S_{MN_ID}	The size of MN_ID	8 bytes			

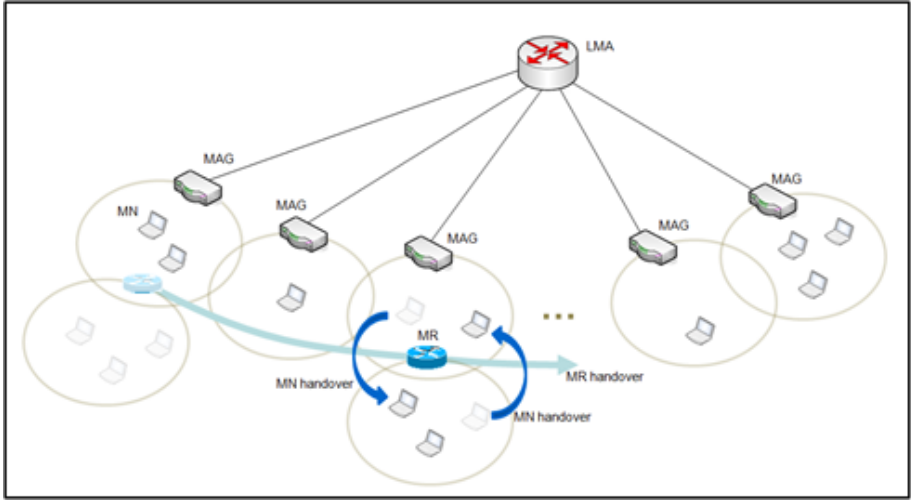


Fig. 4. Analysis scenario

4.2 Traffic Cost Analysis

The Signaling cost is the sum of the message sizes that is required for one handover. First, the mobile network handover signaling cost in PMIPv6 domain is given as follows,

$$C_{MR}^{N-PMIP} = \sum_{n_f}^{\infty} n_f \beta(n_f) \times RS_{MR-MAG} + 2AAA_{MAG-AAA} + BU_{MAG-LMA} + BA_{LMA-MAG} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times RA_{MAG-MR} \quad (1)$$

$AAA_{MAG-AAA}$ is the cost of transmission of an AAA authentication message. The cost is doubled as the AAA query and AAA acknowledgement are represented. $BU_{MAG-LMA}$ is the transmission cost of the PBU message that is sent from the MAG to the LMA, and $BA_{LMA-MAG}$ is the transmission cost of the PBA message which is sent from the LMA to the MAG. RS_{MR-MAG} and RA_{MAG-MR} are transmission costs of the RS message and RA message. These messages are transmitted over wireless links. When these messages are transmitted over a wireless link, they may be failed to be transmitted with a failure probability p_f and the number of failures, n_f .

$$\sum_{n_f}^{\infty} n_f \beta(n_f) = \left(1 + \frac{p_f}{1-p_f}\right) \quad (2)$$

Now, equation (1) is rewritten as equation (3).

$$C_{MR}^{N-PMIP} = \left(\left(1 + \frac{p_f}{1-p_f}\right) \times H_{MR-MAG} \times S_{RS} \right) + 2(H_{MAG-AAA} \times S_{AAA}) + (H_{MAG-LMA} \times (S_{PBU} + S_{PBA})) + \left(\left(1 + \frac{p_f}{1-p_f}\right) \times H_{MR-MAG} \times S_{RA} \right) \quad (3)$$

In the case when MN moves, following is the signaling cost. When MN takes the handover from the mobile network to an MAG domain, the signaling cost is obtained as follows.

$$C_{MN_OUT}^{N-PMIP} = \sum_{n_f}^{\infty} n_f \beta(n_f) \times RS_{MN-MAG} + 2AAA_{MAG-AAA} + BU_{MAG-LMA} + BA_{LMA-MAG} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times RA_{MAG-MN} \quad (4)$$

The RS message and RA message are exchanged between the MAG and MN while in equation (1), they are exchanged between the MR and MAG. Equation (4) is represented as follows.

$$C_{MN_OUT}^{N-PMIP} = \left(\left(1 + \frac{p_f}{1-p_f}\right) \times H_{MN-MAG} \times S_{RS} \right) + 2(H_{MAG-AAA} \times S_{AAA}) + (H_{MAG-LMA} \times (S_{PBU} + S_{PBA})) + \left(\left(1 + \frac{p_f}{1-p_f}\right) \times H_{MN-MAG} \times S_{RA} \right) \quad (5)$$

Signaling cost gets higher when MN moves to a mobile network from the MAG domain and then, MN moves in the opposite direction. The reason is that the AAA authentication and PBU, PBA message for the MN is transmitted over a wireless link between MR and MAG and also through the wired link. The signaling cost is as follows.

$$C_{MN_IN}^{N-PMIP} = \sum_{n_f}^{\infty} n_f \beta(n_f) \times RS_{MN-MR} + 2 \left(\sum_{n_f}^{\infty} n_f \beta(n_f) \times AAA_{MR-MAG} + AAA_{MAG-AAA} \right) + \left(\sum_{n_f}^{\infty} n_f \beta(n_f) \times (BU_{MR-MAG} + BA_{MAG-MR}) \right) + BU_{MAG-LMA} + BA_{LMA-MAG} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times RA_{MR-MN} \quad (6)$$

The RS and RA messages are exchanged between the MN and MR. AAA_{MR-MAG} is the transmission cost of an AAA authentication message that is transmitted between the MR and MAG. BU_{MR-MAG} and BA_{MAG-MR} are the transmission costs of the PBU, PBA message between the MR and MAG. Equation (6) is represented as follows.

$$C_{MN_IN}^{N-PMIP} = \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MR} \times S_{RS} \right) + 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times S_{AAA} + H_{MAG-AAA} \times S_{AAA} \right) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times (S_{PBU} + S_{PBA}) + H_{MAG-LMA} \times (S_{PBU} + S_{PBA}) \right) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MR} \times S_{RA} \right) \quad (7)$$

Now, we deduct the signaling costs of the proposed scheme. First, when the mobile network takes a handover in the PMIPv6 domain, the signaling cost is as follows.

$$C_{MR}^{proposal} = \sum_{n_f}^{\infty} n_f \beta(n_f) \times RS_{MR-MAG} + 2AAA_{MAG-AAA} + BU_{MAG-LMA} + BA_{LMA-MAG} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times RA_{MAG-MR} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times NI_{IN} \quad (8)$$

NI_{IN} is the signaling cost that is required to transmit from the mobile network to the MAG and the MAG contains the information of all the MNs. Equation (8) is represented as follows.

$$C_{MR}^{proposal} = \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times S_{RS} \right) + 2(H_{MAG-AAA} \times S_{AAA}) + (H_{MAG-LMA} \times (S_{PBU} + S_{PBA})) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times S_{RA} \right) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times (S_{NI_HEADER} + n_i \times S_{MN_ID}) \right) \quad (9)$$

Compared to N-PMIPv6, the signaling cost of the MR that moves in the proposed scheme is higher. This is due to the MR which sends the MNs' information to the MAG.

In the proposed scheme, the signaling cost of an MN which moves to an MAG domain from the mobile network is represented as given in equation (10).

$$C_{MN_OUT}^{proposal} = \sum_{n_f}^{\infty} n_f \beta(n_f) \times RS_{MN-MAG} + BU_{MAG-LMA} + BA_{LMA-MAG} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times RA_{MAG-MN} \quad (10)$$

$$C_{MN_OUT}^{N-PMIP} = \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MAG} \times S_{RS} \right) + (H_{MAG-LMA} \times (S_{PBU} + S_{PBA})) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MAG} \times S_{RA} \right) \quad (11)$$

The signaling cost for the MN that moves to the mobile network from an MAG domain in the proposed scheme is represented as shown in equation (12). After the completion of the handover, the MR sends information of the MN to the MAG and the MAG adds the information to the table. NI_{SINGLE} is the cost of the message that contains the information of an MN sent to an MAG.

$$C_{MN_IN}^{proposal} = \sum_{n_f}^{\infty} n_f \beta(n_f) \times RS_{MN-MR} + \left(\sum_{n_f}^{\infty} n_f \beta(n_f) \times (BU_{MR-MAG} + BA_{MAG-MR}) + BU_{MAG-LMA} + BA_{LMA-MAG} \right) + \sum_{n_f}^{\infty} n_f \beta(n_f) \times RA_{MR-MN} + \sum_{n_f}^{\infty} n_f \beta(n_f) \times NI_{SINGLE} \quad (12)$$

The AAA procedure is removed and the procedure to pass information of the MN is added. As the cost required for the AAA procedure is higher, the signaling cost for the MN that moves to the mobile network is reduced compared to N-PMIPv6. Equation (12) is represented as follows.

$$C_{MN_IN}^{proposal} = \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MR} \times S_{RS} \right) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times (S_{PBU} + S_{PBA}) + H_{MAG-LMA} \times (S_{PBU} + S_{PBA}) \right) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MR} \times S_{RA} \right) + \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times (S_{NI_HEADER} + S_{MN_ID}) \right) \quad (13)$$

4.3 Handover Latency

It takes the same time for the mobile network to handover in PMIPv6 domain in our proposed scheme as the N-PMIPv6. Therefore, we will only compare the MN's handover latency between the mobile network and MAG domain. In N-PMIPv6, when an MN takes handover between the mobile networks, handover latency is as follows.

$$L_{MN_OUT}^{N-PMIP} = T_{RS} + T_{AAA_OUT} + T_{PBU_OUT} + T_{PBA_OUT} + T_{RA} \\ = 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MAG} \times D_{wireless} \right) + 2(H_{MAG-AAA} \times D_{wired}) + 2(H_{MAG-LMA} \times D_{wired}) \quad (14)$$

$$L_{MN_IN}^{N-PMIP} = T_{RS} + T_{AAA_IN} + T_{PBU_IN} + T_{PBA_IN} + T_{RA} \\ = 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MR} \times D_{wireless} \right) + 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times D_{wireless} + H_{MAG-AAA} \times D_{wired} \right) + \\ 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times D_{wireless} + H_{MAG-LMA} \times D_{wired} \right) \quad (15)$$

Equations (14) and (15) are the handover latencies of the MN which handovers from the mobile network to a MAG domain and from a MAG to the mobile network in N-PMIPv6 respectively. The latency of the MN moving to the mobile network is longer. This is due to the signaling which is performed over a wireless link between the MR and the MAG.

The latency of the MN moving between the mobile network and the MAG domain in the proposed scheme is as follows.

$$\begin{aligned}
 L_{MN_OUT}^{proposal} &= T_{RS} + T_{PBU_OUT} + T_{PBA_OUT} + T_{RA} \\
 &= 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MAG} \times D_{wireless} \right) + \\
 &\quad 2(H_{MAG-LMA} \times D_{wired})
 \end{aligned} \tag{16}$$

$$\begin{aligned}
 L_{MN_IN}^{proposal} &= T_{RS} + T_{PBU_IN} + T_{PBA_IN} + T_{RA} \\
 &= 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MN-MR} \times D_{wireless} \right) + \\
 &\quad 2 \left(\left(1 + \frac{p_f}{1-p_f} \right) \times H_{MR-MAG} \times D_{wireless} + H_{MAG-LMA} \times D_{wired} \right)
 \end{aligned} \tag{17}$$

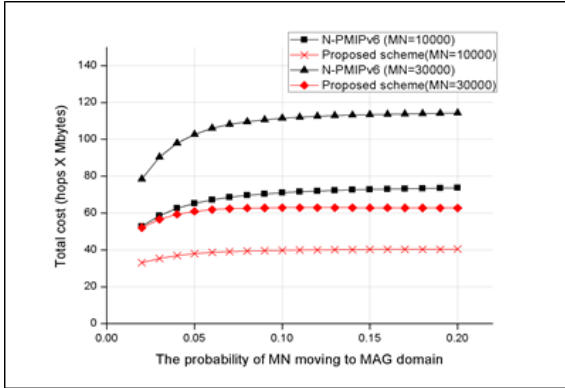
Equations (16) and (17) are the latencies of the MN that take handover from the mobile network to the MAG domain and from the MAG domain to the mobile network in our proposed scheme respectively. In both the cases, by reducing the AAA authentication procedure, the latencies are shorter compared to that in N-PMIPv6.

4.4 Results

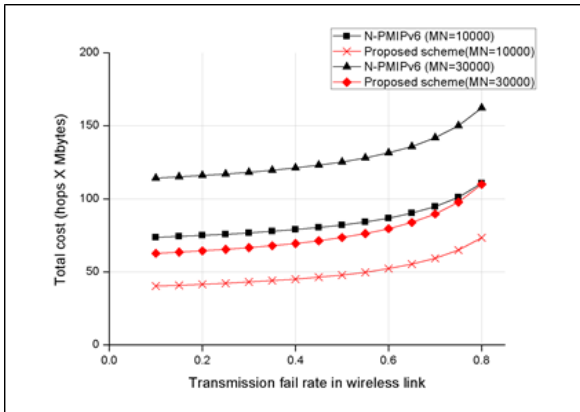
In this subsection, we apply the equations that were deduced in the subsection 4.3 to the analysis scenario and we obtain the signaling costs for the comparison of N-PMIPv6 and the proposed scheme. We assume that the mobile network takes handover for every 2minutes between the MAG domains. In this scenario, there are 50 MAGs and the mobile network goes through all the MAG domains. Mobile nodes handover between the mobile network and a MAG domain causes a change in the number of MNs in the mobile network. Then, the mobile network moves to the next MAG domain. After the mobile has passed through all the MAG domains, we compare the signaling costs. Figure 5(a) shows the change in the signaling costs according to that the change in the probability of the MN’s handover to an MAG domain increases. We maintain a constant number of MNs which handover to the mobile network from a MAG domain. In both the cases, the signaling costs increase as the probability increases. However, not only the signaling cost but also the rate of increase in the proposed scheme is lower compared to that in N-PMIPv6.

Figure 5(b) shows the change in the signaling costs according to the change of the transmission fail rate over a wireless link. We can observe that the signaling costs increase as the transmission fail rate increases.

Figure 6 shows the comparison of packet loss in the cases of N-PMIPv6 and the proposed scheme. Figures 6(a) and (b) represent the packet loss according to the MN's handover to the mobile network from a MAG domain is greater compared to that in the case of the MN handovers to the MAG domain. This is due to the handover latency of the first case which is longer. We can see that the packet loss in the proposed scheme is lower compared to that in N-PMIPv6.

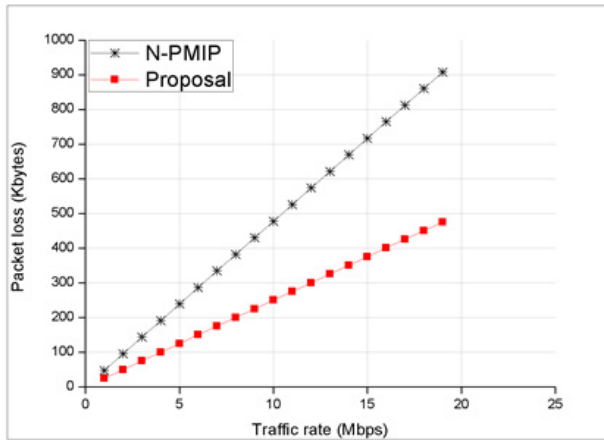


(a). The signaling costs according to the probability of MN handover

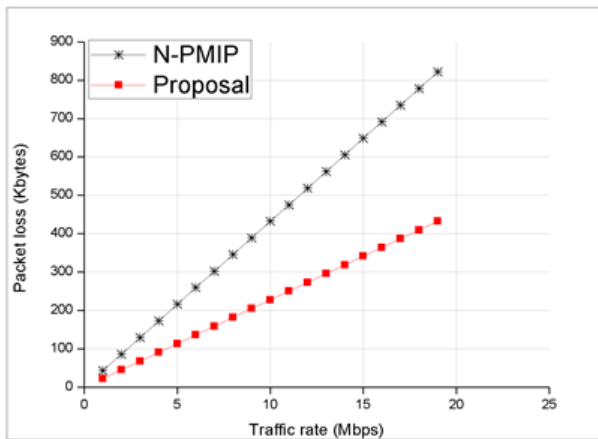


(b). The signaling costs according to the transmission fail rate

Fig. 5. Signaling costs



(a). The packet loss according to the MN's handover to the mobile network



(b). The packet loss according to the MN's handover to an MAG

Fig. 6. Packet loss

5 Conclusion

NEMO-BSP supports the movement of a group of MNs that perform handover signaling only once for all the MNs. However, it cannot be combined with PMIPv6 because this causes a violation of the PMIPv6 policy. In order to solve this problem, several schemes have been proposed. This paper proposes a scheme that improves MN's handover performance between a mobile network and PMIPv6 domain if

NEMO-BSP and PMIPv6 are combined. In the proposed scheme, signaling costs are lower compared to N-PMIPv6 especially MNs handover frequently. Packet loss is also decreased in the proposed scheme and this is due to the handover latency which is shorter than in N-PMIPv6.

Acknowledgement. This research was supported in part by MKE and MEST, Korean government, under ITRC NIPA-2012-(H0301-12-3001), NICDP(2011-0020517) and PRCP(2011-0018397) through NRF of Korea, respectively.

References

1. Johnson, D., Perkins, C., Arkko, J.: Mobility Support in IPv6. IETF RFC 3775 (June 2004)
2. Kong, K., Lee, W., Han, Y., Shin, M., You, H.: Mobility Management for All-IP Mobile Networks: Mobile IPv6 vs. Proxy Mobile IPv6. *IEEE Wireless Communications* (April 2008)
3. Gundavelli, S., et al.: Proxy Mobile IPv6. RFC 5213 (August 2008)
4. Kong, K., Lee, W., Han, Y., Shin, M.: Handover Latency Analysis of a Network-Based Localized Mobility Management Protocol. In: *IEEE International Conference on Communications*, pp. 5838–5843 (May 2008)
5. Li, Y., Jiang, Y., Su, H., Jin, D., Su, L., Zeng, L.: A Group-based Handoff Scheme for Correlated Mobile Nodes in Proxy Mobile IPv6. In: *IEEE Globecom 2009* (November 2009)
6. Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol. RFC 3963 (January 2005)
7. Soto, I., Bernardos, C., Calderon, M., Banchs, A.: NEMO-Enabled Localized Mobility Support for Internet Access in Automotive Scenarios. *IEEE Communication Magazine* (May 2009)
8. Yan, Z., Zhou, H., Zhang, H., Zhang, S., You, I.: Network mobility Support in PMIPv6 Network. In: *IWCNC 2010* (June 2010)
9. Pack, S.: Relay-based Network Mobility Support in Proxy Mobile IPv6 Networks. In: *5th IEEE CCNC 2008* (January 2008)
10. Jeon, S., Kang, N., Kim, Y.: Resource-efficient network mobility support in Proxy Mobile IPv6 domain. *International Journal of Electronics and Communications (AEU)* (October 2011)
11. Lee, H., Han, Y., Min, S.: Network Mobility Support Scheme on PMIPv6 Networks. *IJCNC 2(5)* (September 2010)
12. Lee, J., Ernst, T., Chilamkurti, N.: Performance Analysis of PMIPv6 based Network Mobility for Intelligent Transportation Systems. *IEEE Transactions on Vehicular Technology* (January 2012)
13. Yokota, H., Chowdhury, K., Koodli, R., Patil, B., Xia, F.: Fast Handovers for Proxy Mobile IPv6. IETF RFC 5949 (September 2010)

A Reference Model for Virtual Resource Description and Discovery in Virtual Networks*

Yuemei Xu¹, Yanni Han¹, Wenjia Niu¹, Yang Li¹,
Tao Lin¹, and Song Ci^{1,2,**}

¹ High Performance Network Lab, Institute of Acoustics,
Chinese Academy of Sciences, Beijing, China
{xuym,hany,niuwj,liy,lint}@hpn1.ac.cn

² Department of Computer and Electronics Engineering,
University of Nebraska-Lincoln, NE68182, USA
sci@engr.unl.edu

Abstract. The virtual resource description and provisioning play a key role in virtual resources discovery, selection and binding process. However, there lacks a standard resource description schema for network virtualization. In this paper, we propose a virtual network resource description model, which can give a reference for ISPs (Internet Service Providers) to unify resource management. Furthermore, we extend the WSDL (Web Service Description Language) to specify this model, which is motivated for three reasons. The WSDL supports dynamical update services, which is precisely lacking in the existing network description language. In addition, WSDL is based on XML syntax and is flexible extended for accommodating more properties. Moreover, the resources are essentially services with minimum granularity. Besides the resource definition model and the WSDL-based virtual resource description schema, we also design a virtual resource provisioning framework to confirm the implementation of our proposals. Both theoretical analysis and scenarios demonstration show that the proposed model and framework are effective in dynamic resource discovery and resource composition.

Keywords: Network Virtualization, Resource Description, Resource Provisioning Framework.

1 Introduction

Network virtualization has become a promising solution for overcoming the Internet impasse [1,2,3]. In a network virtualization environment (NVE), multiple heterogeneous network architectures can coexist on a shared substrate network.

* This work has been supported by the National Science and Technology Major Project (NMP) under Grant No. 2010ZX03004-002, the National Natural Science Foundation of China (No. 61103158, No. 11161140319), and Strategic Pilot Project of Chinese Academy of Sciences (No. XDA06010302).

** Corresponding author.

The role of traditional Internet Service Providers (ISPs) is divided into two roles: Infrastructure Providers (InPs) and Virtual Network Provider (VNPs). The InPs manage the substrate networks, response to advertise and register virtual resources, and make them known to VNPs. The VNPs will be able to create various virtual networks to offer customized end-to-end services to the end users. The VNPs can lease shared resources from one or more InPs to deploy different services without considering the physical infrastructure.

The characteristics of resources in virtual networks (VNets) make it challenging for VNPs to select the appropriate resources. Firstly, there are diverse network resources deployed in each InP, and the number of these resources increases over time. Secondly, the resources parameters may change after each allocation. For instance, a link originally has 100Mbps bandwidth. After satisfying a connection request of 10Mbps, its bandwidth turns into 90Mbps, therefore, dynamically updating resources parameters is a crucial aspect. More importantly, users may have the requirements to adjust their used resources. InPs should adjust the virtual resources according to the users real-time demands. The virtual resources description is believed to be a key tool to address these challenges. We need a consensus resource description model to support resources dynamically update and real-time consultation.

However, there lacks a standard resource description for network virtualization. The existing network description specifications (such as cNIS [4], NDL [5], vgDL [6]) are not specified for NVE and cannot describe all the elements in the NVE and adopt to the changing networks. The dynamic resource update is not mentioned in the existing work. Moreover, there is no work concerning on designing a consensus description model for resources in VNets.

In this paper, we abstract the attributes of resources in VNets and utilize the graph theory to model them, which can give a reference for the software developers to unify resources management using common standards and technologies. Then we focus on specifying the model using XML schema.

Web Service Description Language (WSDL) [7] is a W3C standard which provides a model and XML format to describe how and where to get a web service. We tend to describe virtual resources in NVE using WSDL for the following three reasons. Firstly, each VNet request is a service, which is composed of lots of sub-services. The substrate resources are the sub-services with the minimum granularity. For example, a user applies a simplest virtual network consisting of two nodes and a link, then this request is a service and the *link* is a *connection service*, which provides specific bandwidth, latency, etc. Resources are essentially the services of providing some functions. In this sense, WSDL is absolutely able to characterize virtual resources in VNets. The second reason is that, technically, WSDL has a mechanism to support dynamical information update. Once service parameters change, the web service providers can republish a new WSDL document in UDDI (Universal Description, Discovery and Integration) repository to replace the older one, then update the services. Thirdly, WSDL is based on XML syntax, and can be flexibly extended to accommodate more properties [8,9].

Based on the resource definition model and the proposed WSDL-based virtual resource description schema (W-VRDS), VNPs can effectively discover and match resources for user demands. The contributions of our approach are listed as follows:

- We abstract the common characteristics of resources in VNet and propose a mathematical model to define resources.
- We provide a WSDL-based virtual resource description schema to specify our resource model. The WSDL technically supports the dynamic service discovery and update, therefore naturally our proposal also realizes the resources timely and dynamically update.
- We also present a corresponding resource discovery framework and illustrate the resource provisioning process. The analysis shows that our proposed model and W-VRDS can effectively help resource retrieval, discovery and mapping.

The rest of the paper is organized as follows. Section 2 presents a survey of the related work. In section 3, we present a mathematical model to define resources in virtual networks. In section 4, based on the WSDL background, we develop a virtual network resource description schema. Section 5 presents the resource provisioning framework. Then a model and schema analysis is shown in section 6. Finally, conclusions are summarized in section 7.

2 Related Work

In the NVE, InPs have to describe the virtual resources offered to VNPs, while little work concerns about the virtual resources description. Many of the well known specifications such as cNIS [4], NDL [5], are defined to characterize the physical resources in computer of networks. Developed in the GEANT project [1], cNIS (Common Network Information Service) aims to provide a unified repository of all relevant physical network information in a single administrative domain. NDL (Network Description Language) is a semantic Web technique recommended by W3C. Based on the RDF (Resource Description Framework), NDL is mainly used to describe hybrid networks. Furthermore, NDL includes neither fine end resource description nor virtualization constraint specification.

VXDL (Virtual Resources and Interconnection Networks Description Language) [10] is defined for virtual resource interconnection networks, but it is based on data grid applications and only emphasizes the network interconnection, virtualization constraints and the usage timeline description of a resource. The VXDL does not propose a complete schema to describe all aspects of virtual networks. Houidi et al. [11] define a preliminary schema to specify the virtual resource properties and their relationships. This schema, however, is unavailable for dynamic information update, such as the change of user bandwidth or the one-way latency according to the network traffic.

There are numerous researches focused on the WSDL extension. Dai et al. [9] propose a WSDL metamodel extension based on the Model Driven Architecture (MDA) to describe the non-functional aspects of services, hence the WSDL

is able to describe physical objects in the real world. Most of the work extends WSDL for specific applications. The interfaces provided by WSDL are insufficient to model GIS Web services with data-oriented characteristics, therefore Guanhua et al. [12] propose a geographic-Web service description language (G-WSDL) to describe such data-driven services. In order to support the applications with massive datasets, Wiley et al. [13] develop a service invocation mechanism, called WSDL-D, to separate the service invocation messages from their datasets.

Our proposed WSDL-based virtual resource description mainly adopts the *import* mechanism in WSDL schema to accommodate the resources information. We follow the methods in [9] to embed two sub-elements in WSDL schema. So the WSDL extension in this paper is *lightweight*, without altering the original WSDL content.

3 Model Definition

We begin by modeling the virtual resources in VNets. In network virtualization environment (NVE), the network resources include *node*, *link*, *interface* and *path*. Each network element may be virtualized into multiple subelements. For example, a physical node can be virtualized as one or several virtual nodes such as virtual router, or virtual switch; a physical link may contain multiple virtual links with different bandwidth; one or multiple physical/virtual interfaces may connect to a physical/virtual link. An sample model of network virtualization is shown in Fig. 1.

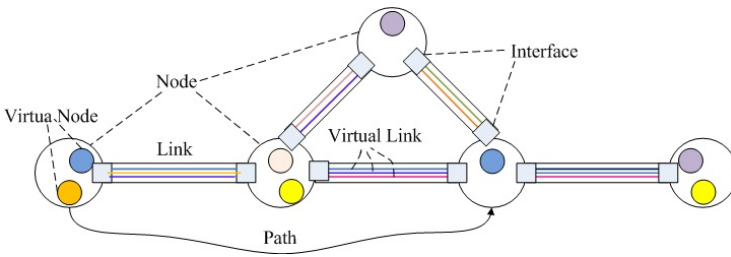


Fig. 1. An Sample Model of Network Virtualization

Let $G = (V, E)$ be a virtual network, node set $V = \{v_1, \dots, v_n\}$, link set $E \subseteq V \times V$. For all $v \in V$ and $e \in E$, interface $i = \chi(v, e) \in I$ represents the the link e is attached directly to node v through interface i . Let *path* P be an ordered set of nodes $P = (v_{P_1}, \dots, v_{P_j})$, such that for all $1 \leq i \leq j$, $(v_{P_i}, v_{P_{i+1}}) \in E$. Considering that each network resource may be virtualized into one or several virtual resources, for all $v_k \in V$, we have $v_k = \{v_{k1}, v_{k2}, \dots, v_{km}\}$, the $v_{ki} (1 \leq i \leq m)$ is the virtual node embedded on node v_k . Note that v_{ki} can be further virtualized. The link $e_k \in E$ and the interface $i_k \in I$ also follow the above node virtualization formulation.

Each network resource has a five-tuple: *name*, *availability* parameter, *timeline*, *required* attributes and *optional* attributes. The *name* is the unique identifier of resources in substrate networks. The *availability* parameter tells the current resources being available or not, which helps InPs control their resources, opening them to users or not, by *True* or *False* value. For example, sometimes InPs may decide to reserve a certain resource, then they can set this parameter as *False*. The *timeline* parameter using the *start* and *for* key words, indicates the moment when the resources are needed and the period for resources reservation. Note that the *availability* parameter is set independently by InPs, while the *timeline* parameter is firstly initiated by users, and is ultimately set in consultation between users and InPs. The *required* attributes define the necessary information of network elements, such as node/link/interface type, bandwidth of a link, OS (operation system), etc. On the other hand, the *optional* attributes characterize the criteria and constraints related to the resources, including performance, QoS, location, etc. The potential attributes of network resource (Node/Link/Interface/Path) are listed in Table 1.

Table 1. Required and Optional Attributes of Network Resources

	<i>Node</i>	<i>Link</i>	<i>Interface</i>	<i>Path</i>
Required attributes	NodeType	LinkType	InterfaceType	BeginNode
	OSType	Bandwidth	...	EndNode
	CPU	ConnectivityType	...	IntermediateNodes
	Memory	IntermediateLinks
	HypervisorType
Optional attributes	Location	Qos	ConnectedNode	Qos
	MacAddress	Location	ConnectedLink	Capacity
	...	ConnectedNode

4 Model Description

In this section, we developed an WSDL-based schema to describe the virtual resource model.

4.1 WSDL2.0 Schema

WSDL2.0 is the latest *Web Service Description Language* which characterizes all the information about the external interfaces of a web service. The web service providers publish the WSDL documents in UDDI registers to tell users what the services can do and how to get the services. On the other side, the users select their favorite services based on the WSDL documents.

The WSDL2.0 schema is illustrated on the left of Fig. 2. The *definition* element is served as the schema container and includes four key elements: *types*, *interface*, *binding* and *service*. The *types* element encloses data type definitions used to define messages. The *interface* element defines the abstract interfaces of a web

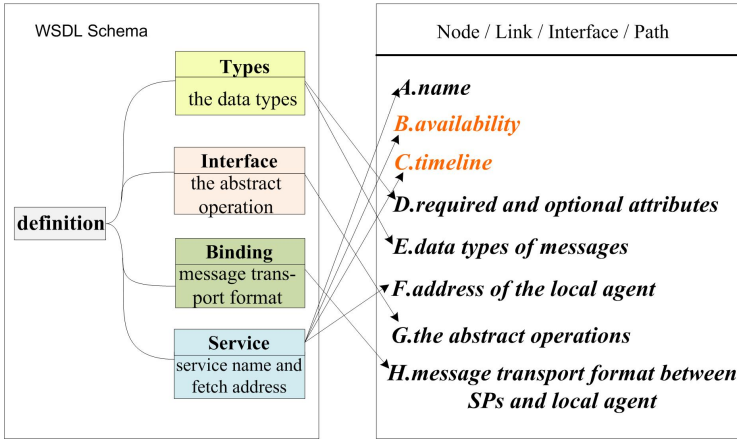


Fig. 2. WSDL2.0 Extension to Describe Virtual Resources

service as a set of abstract operations, each operation representing a simple interaction between the client and the service. The *binding* element defines the underlying format for messages transportation. The *service* element describes a set of endpoints which point out the network address for each *binding* element. To sum up, through the definition of these four key elements, WSDL owns the ability to describe web services' functions and tell how to get services.

4.2 WSDL Extension to Describe the Virtual Resources

A VNet request consisting of many virtual resources, is an applied service requested by users, and finally provided by InPs, which manage infrastructure resources. In this sense, the resources are the sub-services of VNet requests. WSDL is natively used to describe services, therefore it is a natural idea to adopt WSDL to describe virtual resources. Technically, the W3C schema of WSDL allows certain WSDL elements containing extensibility elements. Furthermore, the WS-I Basic Profile 1.1 defines more relaxed extensibility rules. That is, every WSDL element may have extensibility elements and extensibility attributes, which provides theoretical support for our WSDL extension.

In Fig. 2, we map the four key elements of WSDL into describing network resources in VNet. The basic resource information should include two parts: what the resources are and how to get the resources. The A-D elements in Fig. 2 are the five-tuple of network elements, defining what the resources are, while the E-H parts depict how to get these resources, which is consistent with the web service definition rules in WSDL schema.

The E-H parts are the original contents in WSDL schema, and we mainly focus on extending WSDL to characterize the five-tuple of network resources. The *service* element in WSDL contains the *service name* and the *service fetch address*, which serves more likely as the common-sense description part comparing with

```

<?xml version="1.0" encoding="UTF-8"?>
<wsdl:description xmlns:wsdl="http://www.w3.org/ns/wsdl"
  xmlns:msg=
    "http://www.resourcestore.org/virtualresource/xsd"
  ...
  <wsdl:types>
    <xs:import namespace=
      "http://www.resourcestore.org/virtualresource/xsd"
      schemaLocation="resourcelist.xsd"/>
    <complexType name="RequiredAttribute ">
      <attribute name="nodetype " type="msg:NodeType "
        fixed="router"/>
      <attribute name="ostype " type="msg:OsType "
        fixed="linux"/>
      <attribute name="cpu " type="string "
        fixed="1Ghz"/>
      ...
    </complexType>
    <complexType name="OptionalAttribute ">
      ...
    </complexType>
    ... (The other data types definition)
  </wsdl:types>
  <wsdl:Interface> ... </wsdl:Interface>
  <wsdl:Binding > ... </wsdl:Binding>
  <wsdl:service name="host322"
    Availability="True"
    Timeline=Start 2012.1.20 7:00 AM
    For 4 months
    ... >
</wsdl: service>
</wsdl:description>

```

Fig. 3. A Virtual Resources Description Example in WSDL2.0 Schema

the other three key elements. Therefore, on one hand, we use the *service name* to characterize the *name* element of five-tuple; on the other hand, we embed two parameters, *availability*, and *timeline*, in *service* element.

We exploit the *import* mechanism [14] in the *Types* part of WSDL to describe the *required* and *optional* attributes. The definition of these two attributes need accommodate new data types. The *import* mechanism supports reusing the same service description in multiple contexts. The *import* mechanism defines reused types in a target namespace, then the other documents can directly refer to the target namespace without redefining again. The *import* mechanism can save extensive unnecessary labor. We can define all the potential data types of resources' attributes, no matter required or optional, in a target namespace. Then with *import* mechanism, any resource description document can reuse the data type definition in this target namespace. A virtual resource description example is illustrated in Fig. 3. The *import namespace* key word points out the data type definition space, whose URL is "http://www.resourcestore.org/virtualresource/xsd",

```

<?xml version="1.0" encoding="UTF-8"?>
<schema      xmlns="http://www.w3.org/2001/
XMLSchema"
targetNamespace=
"http://www.resourcestore.org/
virtualresource/xsd"
...
<xsd: simpleType name="NodeType">
  <xsd: restriction base="string">
    <xsd: enumeration value="router"/>
    <xsd: enumeration value="switch"/>
    <xsd: enumeration value="baseStation"/>
    <xsd: enumeration value="gateway"/>
    ...
  </xsd: restriction>
</xsd: simpleType>
<xsd: simpleType name="LinkType">
  <xsd: restriction base="string">
    <xsd: enumeration value="VLAN"/>
    <xsd: enumeration value="SONET"/>
    <xsd: enumeration value="802.11"/>
    ...
  </xsd: restriction>
</xsd: simpleType>
...
</schema>

```

Fig. 4. A Segment of the Target Namespace Definition Document

accommodating all the potential types definition. Fig. 4 is a segment of this target namespace definition document.

We present our proposed WSDL-based virtual resource description schema (W-VRDS) in Fig. 5. By adapting WSDL schema to describe the virtual resources in VNet, we do not need to extend WSDL a lot. Our approach totally embeds two elements: *availability* and *timeline* in the *service* element. Then the WSDL is totally capable of characterizing all the network elements of VNet.

5 WSDL-Based VNet Resource Provisioning Framework

In this section, we illustrate a virtual network (VNet) resource provisioning framework based on our proposed W-VRDS. As shown in Fig. 6, InPs own one or several substrate networks, which are composed of many physical and virtual resources. The *local agent* located in substrate networks manages the local network resources and dynamically generates WSDL documents to advertise and register their resources in UDDI registers. Then VNP will know these available

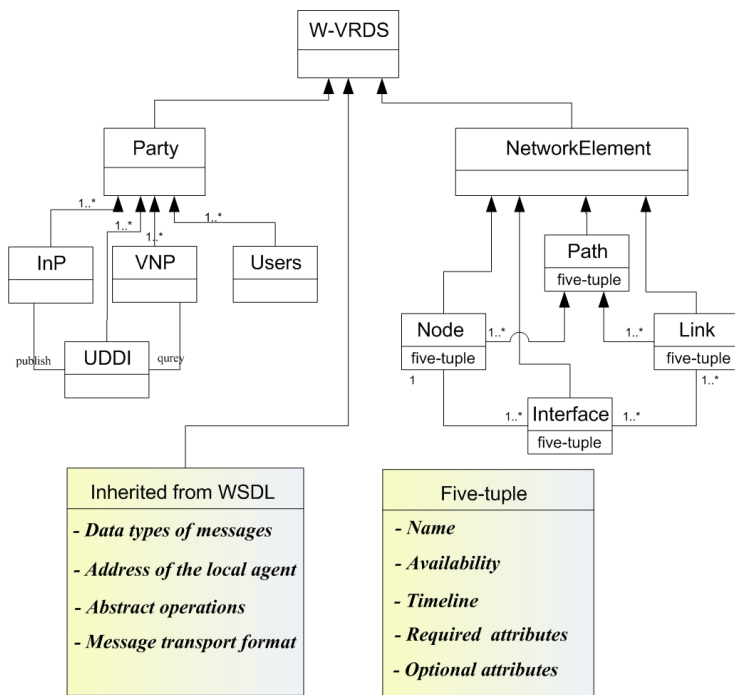


Fig. 5. The WSDL-based Virtual Resource Description Schema

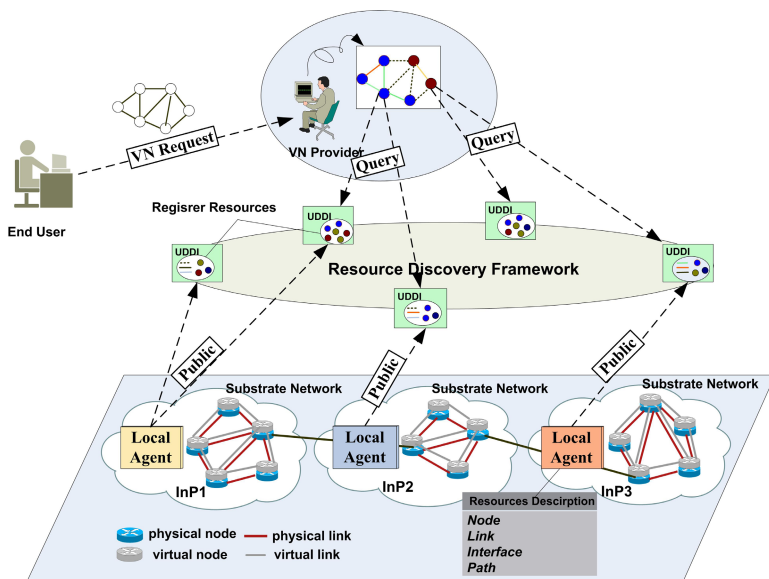


Fig. 6. WSDL-based Resource Provisioning Framework

resources. In addition, UDDI registers will traverse the WSDL documents to abstract the five-tuple of resources to support the resource retrieval, and can use the conceptual clustering algorithm CLUSTER3 [15] to cluster resources, which will enhance the resource retrieval efficiency.

When a user applies for a specific virtual network, he will send a request to VNPs. The VNPs then analyze the user request and search for the corresponding resources in different UDDI registers. The searching process in UDDI is based on the five-tuple of resources. After VNPs selecting their candidate resources, UDDI registers will send VNPs the corresponding WSDL documents, which depict what the resources are and how to get the resources. Then the step forward process is similar to the service fetch mechanism. The E-H contents inherited from WSDL schema describe the resource fetch interfaces. VNPs follow these interface formats, such as message types, and operation types, then communicate with InPs, and finally get consultations on resources allocation. Finally, the *local agents* in InPs assign resources according to the agreements. The above detailed process is illustrated in Fig. 7

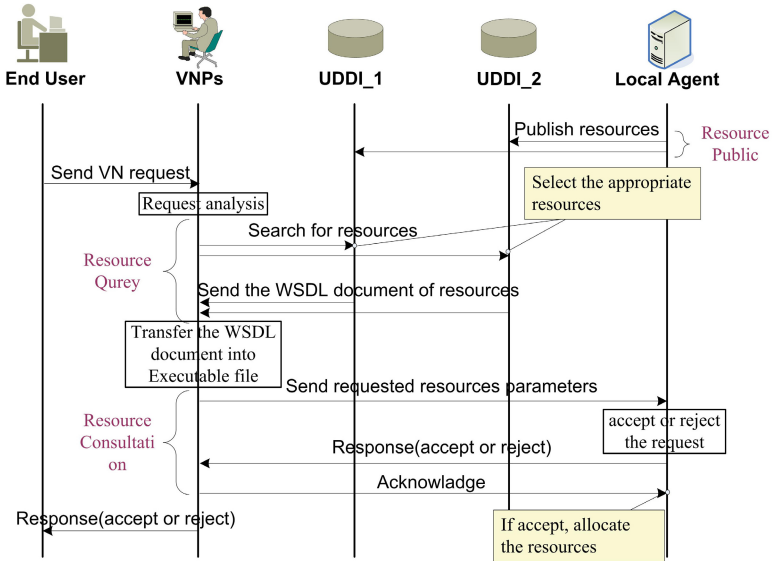


Fig. 7. WSDL-based VNet Resource Provisioning Process

6 Model Analysis

In this section, we illustrate how the resources definition model and the proposed W-VRDS can help retrieve, embed and generate VNet.

Suppose that an InP owns one substrate network, which is assumed to have six nodes. The network topology is shown on the right of Fig. 8. The *local agent* in the InP describes these resources adopting W-VRDS, and publishes them in

Table 2. Required and Optional Attributes of Network Resources

Name	Availability	Timeline	Required attributes	Optional attributes
A	True	Start(2012.1.20,7:00 am) For(3 months)	NT=router;OS=windows; Cpu=2GHz;Memory=3GB	None
B	True	Start(2012.1.20,7:00 am) For(3 months)	NT=router;OS=windows; Cpu=1GHz;Memory=2GB	None
C	True	Start(2012.1.22,8:00 am) For(2 months)	NT=gateway;OS=linux; Cpu=4GHz;Memory=5GB	None
D	True	Start(2012.1.22,8:00 am) For(3 months)	NT=router;OS=linux; Cpu=500MHz;Memory=1GB	None
E	True	Start(2012.1.22,8:00 am) For(2 months)	NT=router;OS=xen; Cpu=2GHz;Memory=4GB	None
F	True	Start(2012.1.22,8:00 am) For(2 months)	NT=router;OS=linux; Cpu=2GHz;Memory=3GB	None

UDDI registers. Due to space limitation, we do not give out the complete WSDL description documents, but we abstract the five-tuple of node resources in the Table 2. Note that the *timeline* parameter of unused resources should be *NULL*, and is written after the resource reservation. The InPs will release the assigned resources when time is up. In addition, the links in the substrate network are characterized by bandwidth attribute, and the corresponding values are directly depicted in Fig. 8.

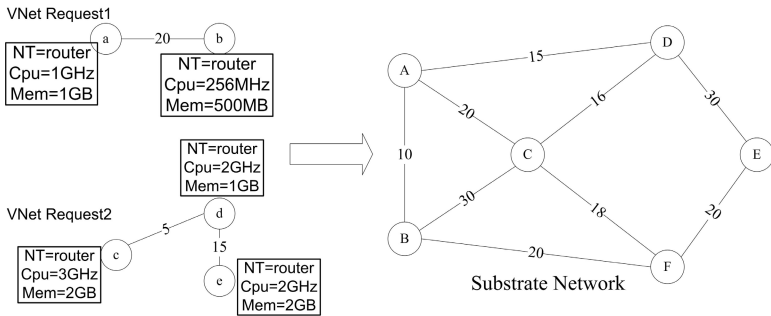


Fig. 8. VNet Requests and Substrate Network Topology

Now we assume user Bob and Alice want to apply for their dedicated virtual networks. The requested network topologies and the corresponding resources parameters are both illustrated on the left of Fig. 8. These two users send their requests to VNPs, then the VNPs will search in UDDI registers to find candidate resources. The resource retrieval, discovery and update process is shown in Alg. 1. In this paper, the UDDI registers use the greedy algorithm and the shortest path algorithm to find and map resources. Adopting both of these two algorithms to select resources in VNets is sufficiently analyzed in reference [16] [17]. Therefore,

we do not provide the detailed algorithms of implementation process. The final mapping results are depicted in Fig. 9. Note that, the available resource parameters of node A, C and the link directly connected them change after mapping the VNet request 1. Specifically, the memory of node A and C reduce 500MB and 1GB, respectively, the available bandwidth of connected link turns into 0, and can not accept other requests. It is interesting to notice that after satisfying the VNet request 1, the node C is still capable to accommodate the requirements of node E. Therefore, two virtual nodes are embedded into node C at the same time.

After selecting suitable resources, the VNPs will fetch the WSDL documents of candidate resources from UDDI registers, and use the described information in WSDL documents to communicate with InPs, asking for the candidate resources reservation.

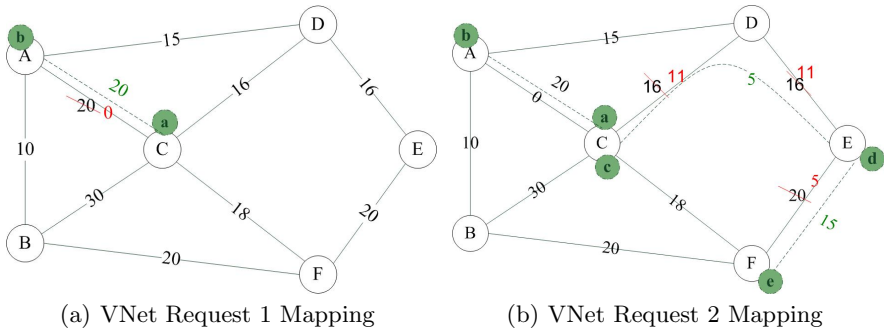


Fig. 9. Mapping the VNet Requests into Substrate Network

Algorithm 1. *Resource discovery and update Algorithm*

1: **INPUTS:**

$VReq_i$: VNet request;

D_j : WSDL documents of all the resources in the substrate network;

$G_s = (V_s, E_s)$: substrate network topology;

2: **begin**

3: UDDI abstracts the five-tuple of the available resources from the WSDL documents.

4: **for** each $VReq_i$ **do**

5: 1. Adopt the greedy algorithm to map the nodes into the substrate network.

2. Adopt the shortest path algorithm to link the candidate nodes.

3. Update the five-tuple information.

4. Update the WSDL documents in the UDDI.

6: **end for**

7: **end**

7 Conclusion

In this paper, we propose a model to define virtual resources in V-Nets, and adopt a WSDL-based virtual resource description schema to specify it. Our proposal is motivated mainly by three actors: 1) resources are essentially services of providing some functions; 2) WSDL is based on XML syntax and is flexible extended to accommodate more properties; 3) WSDL naturally supports dynamical services update, which is exactly what we urgent need in specifying virtual resources. The model and schema analysis shows that our approach has advantages in supporting resources dynamically publication and update, and is effective in resources retrieval, discovery and mapping.

Future work will consist of exploiting resource clustering to enhance the resource discovery process. The clustering algorithms will be discussed. The optimal resources clustering can be an essential research issue.

References

1. Elliott, C.: Geni-global environment for network innovations. In: 33rd IEEE Conference on Local Computer Networks, p. 8 (2008)
2. Ohlman, B., Ahlgren, B., Brunner, M.: et al. First netinf architecture description. 4WARD project D, 6 (2009)
3. Anderson, T., Peterson, L., Shenker, S., Turner, J.: Overcoming the internet impasse through virtualization. *Computer* 38(4), 34–41 (2005)
4. Wolski, M., Osinski, S., Gruszczynski, P., Labeledzki, M., Patil, A., Thomson, I.: Deliverable ds3. 13.1: common network information service schema specification. Information Society and Media, GN2-07-045v4 (2007)
5. Poliac, M.O., Wilcox, G.L.: Ndl: A network description language for generalized back propagation networks (1988)
6. Chien, A., Casanova, H., Kee, Y.S., Huang, R.: The virtual grid description language: vgdL (2004)
7. Chinnici, R., Moreau, J.J., Ryman, A., Weerawarana, S.: Web services description language (wsdl) version 2.0 part 1: Core language. W3C Working Draft, 26 (2004)
8. D'Ambrogio, A.: A model-driven wsdl extension for describing the qos of web services. In: International Conference on Web Services, ICWS 2006, pp. 789–796. IEEE (2006)
9. Dai, C., Wang, Z.: A Flexible Extension of WSDL to Describe Non-Functional Attributes. In: 2010 2nd International Conference on e-Business and Information System Security (EBISS), pp. 1–4. IEEE (2010)
10. Koslovski, G.P., Primet, P.V.-B., Charão, A.S.: VXDL: Virtual Resources and Interconnection Networks Description Language. In: Vicat-Blanc Primet, P., Kudoh, T., Mambretti, J. (eds.) *GridNets 2008*. LNICST, vol. 2, pp. 138–154. Springer, Heidelberg (2009)
11. Houidi, I., Louati, W., Zeglache, D., Baucke, S.: Virtual resource description and clustering for virtual network discovery. In: IEEE International Conference on Communications Workshops, ICC Workshops 2009, pp. 1–6. IEEE (2009)
12. Guanhua, C., Kunqing, X., Xiujun, M., Yanfeng, S., Yuanzhi, Z., Lebin, S.: G-WSDL: a data-oriented approach to model GIS Web services. In: Proceedings of 2005 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2005, vol. 2, p. 4. IEEE (2005)

13. Wiley, M., Wu, A., Su, J.: WSDL-D: A Flexible Web Service Invocation Mechanism for Large Datasets. In: 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, pp. 157–164. IEEE (2008)
14. Chinnici, R., Moreau, J.J., Ryman, A., Weerawarana, S.: Web services description language (wsdl) version 2.0 part 1: Core language. W3C Recommendation, 26 (2007)
15. Seeman, W.D., Michalsk, R.S.: The cluster3 system for goal-oriented conceptual clustering: method and preliminary results. In: Proceedings of The Data Mining and Information Engineering 2006 Conference, Prague, Czech Republic, vol. 37, pp. 81–90 (2006)
16. Zhu, Y., Ammar, M.: Algorithms for assigning substrate network resources to virtual network components. In: Proc. IEEE INFOCOM, vol. 2 (2006)
17. Houidi, I., Louati, W., Zeghlache, D.: A distributed virtual network mapping algorithm. In: IEEE International Conference on Communications, ICC 2008, pp. 5634–5640. IEEE (2008)

TV Remote Control Using Human Hand Motion Based on Optical Flow System*

Soonmook Jeong**, Taehoun Song, Keyho Kwon, and Jae Wook Jeon

School of Information and Communication Engineering, Sungkyunkwan University,
Suwon, South Korea

{kuni80, thsong}@ece.skku.ac.kr,

{kykwon, jwjeon}@yurim.skku.ac.kr

Abstract. Motion recognition systems have been widely developed in the field of human computer interaction. Methods, such as pointing, dynamic gesture and static gesture or hand held devices have been proposed for motion recognition. The motion recognition systems have been gradually adapted to home appliances in our daily. In this paper, we focus on TV interaction, since the device is a recent representative multimedia device applying the motion technique. Most motion recognition systems utilize 3D data, such as horizontal, vertical and depth information by stereo camera or ToF (Time of Flight) camera. However, this paper proposes the different techniques for human-TV interaction. We propose an optical flow based motion recognition system that provides direction and speed, in addition to the position of the moving target in real time. These factors are useful in recognizing human motion more effectively and more dynamically. Therefore, we design the natural interaction for human-TV using these motion data. The calculation process of optical flow is outside the scope of this paper. This real time optical flow calculation is implemented using the FPGA chip supporting parallel processing by a hardware team in our laboratory. We propose a method for human motion recognition based on real time optical flow system.

Keywords: Motion recognition, Human-TV Interaction, Optical flow, Natural interaction, Real-time system.

1 Introduction

TV and game devices are representative consumer devices in our daily lives. These devices are handled by a dedicated controller. The controller is designed to utilize all the device functions. However, as the equipment functions, such as home appliances

* This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2012-(H0301-12-3001))supervised by the NIPA(National IT Industry Promotion Agency), and by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0018397).

** Corresponding author.

closely related to human activity, become diverse and complex, the controller interface becomes complex and dedicated to specific devices. This may be confusing to the user. The user should learn and memorize diverse individual instructions.

Therefore, attention to human-computer interaction or human-machine interaction has increased. Hand motion is one of the most common, natural and intuitive interaction methods among humans. This interaction has been gradually adapted to home appliances in our daily lives, of which the representative appliance is 3D gesture TV. Along with increase of Smart TV, natural interactions, such as hand motion recognition, have actively been developed. The ultimate goal of these interactions is to control TV using only a human's hand motion instead of the conventional remote controller. Motion sensors, such as the 'Kinect' of Microsoft and 'Xtion Pro' or 'WaviXtion' of ASUS, have widely been recently adapted to these gesture TVs. These motion sensors can recognize almost all joints of the human body. This is useful to recognize diverse and complex human motion. Kinect support for Netflix allows the user to watch TV shows and movies streaming from Netflix with just hand gesture and voice sound. The TV gesture control using 'WaviXtion' is also shown in CeBIT 2011. Fig. 1 shows the gesture TV using these motion sensors.

However, even though these motion sensors provide much data, such as hand point, each joint and distance to recognize the diverse and complex human motion, we are not convinced that these data are needed for TV control. Our first doubt is do we really need complex and various motion data to control TV? Recent Smart TV interfaces are similar to the graphical user interface PC environment. Therefore, pointing based interaction is still considered as a general and natural interface to control a device such as a Smart TV. At this point, we wonder if it really provides the user natural and easy interface to control TV. Is the method optimum for human-TV interaction? In an environment similar to a PC, the user has to adjust the pointer to place it on a certain area of the screen. This may require considerable concentration and mental load, since pointing using hand motion is much less elaborate than that using a conventional mouse device. If the interaction is not easy and comfortable to use, even if it is intuitive and natural, the user might avoid using the interaction.

This paper proposes a different motion sensor system to solve the issues we questioned above. In contrast, most recent representative motion sensors are based on the ToF or stereo sensor.



Fig. 1. Gesture TVs using motion sensors

We design a different interaction style to control TV using this system. Our interaction is designed based on important factors for human-machine interaction [1], since the interaction should generally be designed with consideration of its usability and convenience. Our interaction is based on motion tracking. This requires accurate segmentation of objects from the background for effective tracking. It also requires real time operation and accurate tracking. Optical flow can be used to segment incoming images, regardless of background, since optical flow allows multiple moving targets to be separated based on their individual speed. It also provides the direction of the moving target. It is difficult to obtain these features using existing motion segmentation methods, such as blob, edge and skin detection. Despite of these advantages, the huge computation amount of optical flow has been burdensome for the adaption of this technique to effective motion tracking, since it does not assure real time operation. A FPGA chip supporting parallel processing is used in our work to resolve this real time problem. However, this system is not included in the scope of this paper.

Our work here focuses on a new interaction design for human-TV interaction and proposes a new motion sensor system based on a real time system using optical flow at the FPGA level. The remainder of this paper is organized as follows. Section 2 introduces related works. Section 3 and 4 describe the human-machine interaction design to control TV. Section 5 presents the system configuration. In Section 6, we demonstrate experimental results and summarize the results in conclusion section.

2 Related Works

2.1 Pointing Based Interaction

This interaction is similar to that using mouse pointer. The difference is to use the hand or the finger instead of the mouse to control the pointer. This method is generally used for human-machine interaction. William T. F. et al. proposes the method to adjust various graphical controls of television with the hand icon [2]. Andrew W. et al. developed an application that allows users to conduct various window management tasks using feedback, such as hand icon, regarding the user's hand position [3]. However, this interaction requires the more concentration and mental load than conventional device, such as mouse.

2.2 Dynamic Gesture Based Interaction

Michael V. et al. aims to improve a real time hand gesture interaction system by augmenting it with a ToF camera. Masaki T. et al. propose the method which recognizes the various motions accurately. So this work estimates the 3.5D spatiotemporal trajectory features, which contain horizontal, vertical, time and depth information from a ToF camera. However the ToF camera is too expensive to commercialize yet. Another work proposes the method, which allows the user to control the media via a list menu shown on a distant display by drawing circles in the

air with one hand using stereo camera [6]. However, this system only provides two functions, such as the browse and choosing. It is not enough to control the media content. Chen. M. et al. demonstrate an application that recognizes gestures to control TV [7]. Hsieh C. et al. propose a real time hand gesture recognition system based on adaptive skin color model and motion history image [8]. However skin color is always difficult to avoid the influence of light and this method assumes the background is always static. Ryuta Y. et al. proposes the pointing method only using the wrist movements [9]. Ross C. et al. developed the view-based gesture recognition system using optical flow [10]. These Gestures are recognized using a rule-based technique based on characteristics of the motion blobs. However, this system is much simpler comparing with recent motion sensor.

2.3 Static Gesture Based Interaction

Most gestures are similar to the symbols for the corresponding functions in static gesture. Arnaud B. et al. aim at recognizing hand signs and positions using a single webcam [11]. Xiujuan C. et al. focuses on accurate hand segmentation using 3D depth data [12]. He believes this elaborate hand segmentation makes the accurate hand gesture recognition. Xia L. et al. propose the method for recognizing hand gesture by using a sequence of real-time depth image data [13]. This work is shown to be possible to recognize many types of gestures. However, the user should memorize the specific form or the gesture according to each function. As the number of function increases, this method requires more user memory.

2.4 Hand Held Device Based Interaction

Lee D.W. et al. introduces a wristwatch-type of remote that offers a unified way to control various devices and how a user's hand motions can be used in a fast and effective way with the virtual menu [14]. Kim S. et al. presents a hand-held system which is tracking of the full 6 degrees-of-freedom position and orientation for 3D interaction with digital media contents [15]. However the user experiences the onerousness he/she should wear the additional device to use this interaction.

3 Motion Recognition

Our interaction system adopts the optical flow to extract the motion data from the single image. Optical flow is generally used to detect motion using a brightness pattern. It can provide the vector data about the moving object. This section describes how to extract motion data, such as direction, speed, and position from these vector data about the moving object. In our system, vector information of the moving object is obtained by the optical flow system based on FPGA. It is implemented in real time. These vector data correspond to color information, such as RGB. Fig. 2 shows the RGB color information including object's vector information. We use this color information to track the object's motion.

3.1 Generation of Color Table

In this work, the vector information of the moving object is expressed as color information within an image. It means the color information indicates the direction and the speed of the moving object. As the direction and the speed of the moving object change, the color information also changes.

Therefore, the vector information should be extracted from the color information. This is implemented using a color table that matches the color information to the vector information. This color table is generated by referencing the color image 49×49 , as shown in (a) of Fig. 3. It has information about the coordinates (x, y) corresponding to each color pixel of an image. The next section describes how to extract vector data using the color table.

3.2 Vector Extraction

This sub-section describes how to extract the vector data from color image using color table. Suppose that one pixel of the moving object transits from one point to another point. The first step is to read the color information of the transited pixel. Next, find the coordinates corresponding to the color using the color table, as shown in Fig. 3. Then, calculate the direction and speed using the coordinates. It is calculated by equation (1), (2). The third of Fig. 4 shows the vector information representing the direction and the speed of the moving object. The next step is to detect the motion using this vector information.

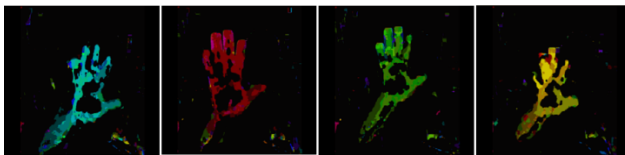


Fig. 2. Color information including vector information

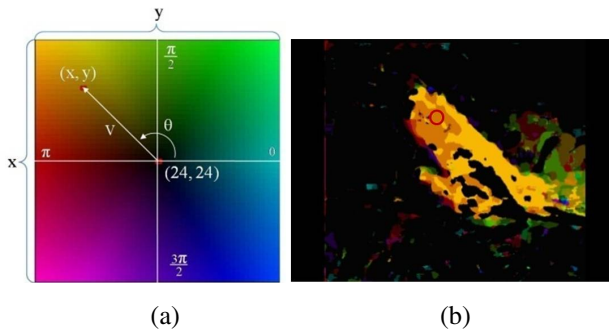


Fig. 3. (a) Reference color image representing the coordinate corresponding to (b) one color pixel of the motion

$$Degree = \left(\tan^{-1} \left(\frac{24-y}{24-x} \right) + 360 \right) \bmod 360 \tag{1}$$

$$Speed = \sqrt{(24-y)^2 + (24-x)^2} \tag{2}$$

3.3 Feature Extraction

However, these extracted vector information are quite coarse and have unnecessary data, such as noise, as shown in third of Fig. 4. That is, these original data make it difficult to estimate the moving object accurately and robustly. We found experimentally that vectors that have the same direction within a moving object have also similar speed. If the direction and speed are converted to a value using feature extraction, the entire vectors within an image are classified according to a similar vector. That is, the specific data set can be extracted from the entire data set.

1) Feature Reduction: We convert these raw data to useful data to facilitate the estimation of the moving object using PCA (Principle Component Analysis). PCA is generally used to reduce the space of D-dimensions to the space of d-dimensions ($D > d$). Dimension reduction can be used for feature extraction and it is implemented by the projection of original vector using transformation matrix. In equation (3), s is the original vector and x is the projected vector using transformation matrix u^T . This transformation matrix is obtained by following equation (4). \bar{s} is the mean vector of s . That is, u is the eigenvector of covariance Σ . In this work, the original vectors are two-dimensional feature vectors that are the direction and the speed of the each pixel within the image. These vectors are multiplied by the transformation matrix u^T respectively. Through this calculation, the two-dimensional original vectors are projected onto one-dimensional vectors we find. These projected vectors are used to extract the hand motion. Fig. 4 shows the whole process of the motion detection.

2) Feature Extraction: Next, we classify these projected vectors according to its value. First these vectors are rearranged in size order by a bubble array. Then these vectors are classified into similar vector groups via slope comparison of vector.

$$x = u^T s \tag{3}$$

$$\begin{aligned} & \frac{\partial \left(\frac{1}{N} \sum_{i=1}^N (u^T s_i - u^T \bar{s})^2 + \lambda (1 - u^T u) \right)}{\partial u} = 0 \\ & = 2u^T \left(\frac{1}{N} \sum_{i=1}^N (s_i - \bar{s})(s_i - \bar{s}) \right) - 2\lambda u = 0 \\ & \quad \leftarrow \\ & = 2\Sigma u - 2\lambda u = 0 \\ & = \Sigma u = \lambda u \end{aligned} \tag{4}$$

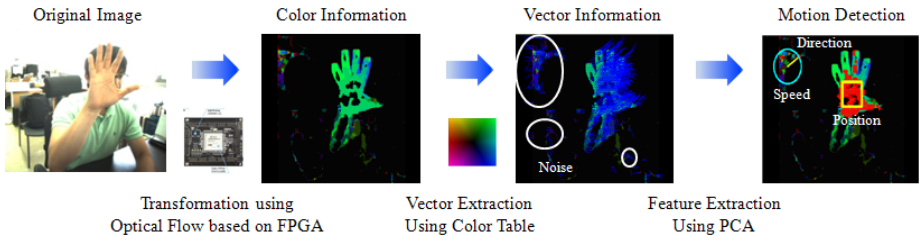


Fig. 4. The whole process of motion detection

Fig. 5 shows these classified vector groups. In these graphs, the candidates for moving object are found in over two places. In this case, the longest range is selected as the motion area among ranges. Therefore, we can accurately extract only the motion from a noisy and coarse image. Fig. 6 shows the extracted hand motion. The size of light blue circle at the left-top shows the speed of the hand motion and the yellow stick means its ongoing direction. The overlaid red points on the object image represent detected range that has similar vector. The yellow square means its central position. We design the interaction method using these extracted motion data.

4 Natural Interaction for TV Control

4.1 Interaction Design

This research focuses on TV control using only hand motion. This section describes the interaction design for TV control from the estimated motion in section 3. The following commands are standard to most multimedia devices, such as Smart TV, mobile phone and mp3 player.

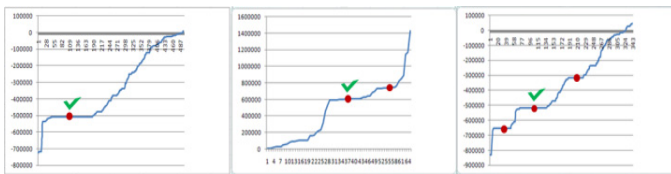


Fig. 5. The graph representing vector group classification

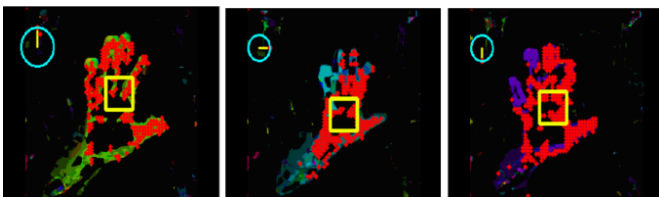


Fig. 6. The extracted direction, speed and position from user hand motion

- 1) Navigation
- 2) Select
- 3) Back (Cancel)

Therefore, these devices are controlled by these commands. Recent motion TV using motion sensors, such as ‘Kinect’ or ‘WaviXtion’, are also controlled by these simple but standard commands. Most motion interaction using these sensors is based on pointing based interaction, similar to mouse device.

In this work, however, we consider the design of a different interaction to resolve the issues we considered above. This research designs a natural and convenient interaction on which our proposed principle is reflected. The proposed interaction should satisfy the following three conditions.

- 1) Movement within a minimum range
- 2) Simple and Consistent
- 3) Easily memorized

The first condition decreases the physical load. The second decreases command complexity. The last improves command learnability.

As shown in Fig. 7, three hand motions are proposed to satisfy the interaction principle of this paper. The navigation command is implemented by spinning the user hand. Its direction can be to the left or right. Putting down the hand generates the selection. The reverse makes the back command. This interaction design is the kind of the hypothesis expected to improve usability for TV control. We evaluate results to test this hypothesis in the next section.

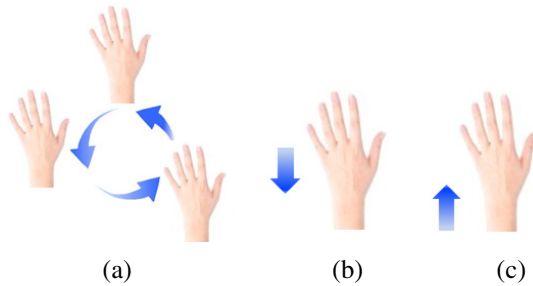


Fig. 7. Natural Interaction Design (a) Navigation (b) Selection (c) Back/Cancel



Fig. 8. The whole process of TV control using the proposed system

4.2 TV Emulation and Visual Feedback

We built a TV emulation environment similar to motion a TV environment to evaluate the proposed interaction. This emulation consisted of three parts. First, the category section is to display diverse genre, such as movie, comedy, music and sports. The user selects one of these categories. Then, content corresponding to the selected category is displayed in the second part. The user navigates the content using hand spin and selects the content he/she wants. The last part is to run the content. In this part, the user controls the volume using hand spin. Fig. 8 shows the entire process of TV control using our proposed interaction.

This architecture is simple but has standard and salient parts. The emulation provides visual feedback to the user. This visual feedback indicates the running command. Fig. 9 shows the four feedbacks corresponding to the hand motion. The first feedback is the holding state. The second indicates the hand spin is running. The third shows the selection by hand down. The last indicates the back command by hand up. The direction of the dark line within the upper circled area corresponds to the direction of hand motion. Its length changes according to the velocity of the motion. This visual feedback is attached to the right side of the TV emulation. Therefore, the user can confirm his/her recognized hand motion state in online.

5 System Configuration

The proposed image processing is based on the optical flow image taken from a Virtex-4 XC4VLX200-10 FPGA from Xilinx. The system interfaces one VCC-8350CL camera from the CIS corporation through the standard camera-link format and takes the 640 x 480 image. The maximum frame rate of the camera is 60 fps. This system was developed by another research team within our laboratory.

We adopt the optical flow image to track the motion in real time, instead of the optical flow system development. The optical flow image is captured by a Meteor-2/CL frame grabber from Matrox. This image is processed on an Intel Core2Duo E7500 (2.94GHz), 3GB DDR2 SDRAM based PC to extract the motion.



Fig. 9. Four feedbacks corresponding to the hand motion

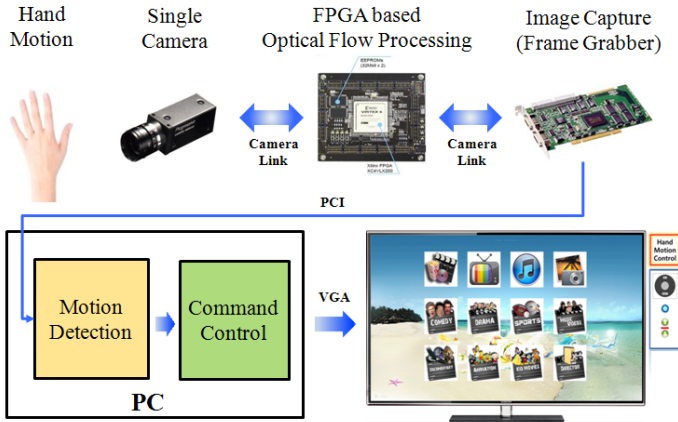


Fig. 10. System configuration

The command control module generates the TV command using the extracted motion. This generated command controls the TV emulation and this is displayed on LCD monitor via VGA cable. Fig. 10 shows the proposed system configuration. As shown figure, this system is consisted of several modules. However this system can be simplified as one ASIC chip. In this work, we show the system configured for experimental purpose.

6 User Evaluation and Results

This system was exhibited at WIS 2011 (World IT Show 2011) held at Seoul this year. Our system was evaluated by many visitors during the exhibition period (four days). Thus, we could obtain diverse opinion and advice in this exhibition. Visitors of various ages, genders and jobs evaluated the proposed motion sensor system.

We briefed the system operation to participants. Then, they participated in the experiment until they were accustomed to the system. They responded with their impressions and thoughts about the system. They evaluated the following factors with this feedback.

1) Accuracy: The proposed system occasionally responded to unintended motion or did not respond to intended motion of the user. We think this is caused by the command type not being separated completely. Accuracy over 80% was estimated, once someone was familiar with the system operation.

2) Learnability: In average, most subjects took 5 minutes to 10 minutes to become familiar with the system. This is a little long to learn some systems. We think it is due to the system recognizing the command when the user motion was similar to the intended motion above 95% compliance. Therefore, subjects spent most time to adjust their motion to the intended motion, despite it being simple motion.

3) Comfort: Most subjects agreed that the hand spin motion to navigate was more convenient than the existing motion sensor. However, they also pointed out the motion of the selection and back were not convenient.

4) Sensitivity: Sensitivity is closely related to the accuracy factor. They have an inverse proportional relationship. As the sensitivity increases, the accuracy decreases. As the sensitivity decreases, the accuracy increases. We tried to decrease the sensitivity of the system. However, an extreme reduction of sensitivity causes an accuracy decrease, as many subjects advised

5) Physical Load: Most subjects did not experience arm fatigue, since our system operates in a small region.

6) Mental Load: Some subjects required to concentrate a lot due to the unfamiliar motion, because they were used to pointer-based interaction, such as 'Kinect'. This caused some mental load.

7) Responsiveness: The proposed system operates in real time, because the system is based on the hardware level using FPGA, which process about 60 frames per second. The load time for motion recognition in the PC was not a problem to control TV emulation. The participants did not take issue with this factor.

8) Intuitiveness: Most subjects agreed the pointing based interaction, such as 'Kinect', was more intuitive than our system, since someone unconsciously behaves by expecting an instant response, according to his/her motion. This was an unexpected result for us.

9) Costs/Benefits: Most subjects agreed that the system is cheaper than the existing motion sensor system, since the system can be built from one CMOS RGB camera and one motion sensor chip at the ASIC level. We estimate the total price is under \$10.

10) User adaptability and Feedback: We provided visual feedback on the right side of the TV emulation. However, subjects advised it divided their concentration for the TV control, because they had to look alternately at the emulation screen and feedback screen to check the running command. They suggested the direct display of visual feedback on the TV emulation would be better than the current system feedback.

Fig. 11 shows the user evaluation scene in the exhibition. The system was evaluated by many participants.



Fig. 11. User evaluation scene in the exhibition at WIS 2011

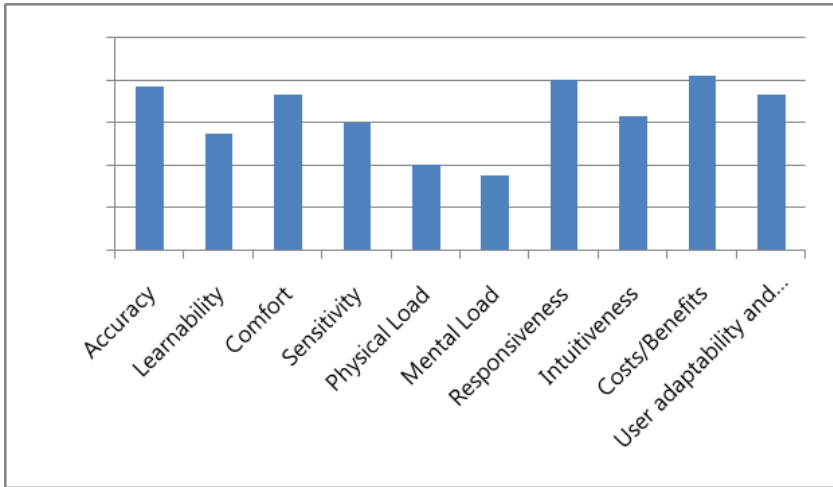


Fig. 12. Questionnaire evaluation about the proposed system

7 Conclusion

This paper proposed a low cost motion sensor system of a different interaction type from that of the interaction of the existing motion sensor. In this research, we focused on TV interaction, since the device is a recent representative multimedia device applying the motion technique. We started this research began with some doubts. 1) Is the representative interaction technique using existing motion sensor is optimum for human-TV interaction? Is the interaction designed with sufficient consideration of the device property? 2) How many functions do we need to control a TV?

As shown in Fig. 12, the proposed system has some issues still to be resolved. The score about learnability and intuitiveness was low among other factors whereas the evaluation about physical load and mental load was good.

It is also true that our system does not support as many functions as other expensive sensors, such as 'Kinect' or 'Xtion Pro', do. Experiments confirmed the proposed system is sufficient for TV control, whereas many participants still consider elaborate pointing based interaction as an important and promising technique for human-TV interaction. The TV Interface does not need as many commands such as those of a complex game. It is mainly controlled via few commands, such as channel searching, volume up/down and turn on/off.

Our system was developed from an understanding of this point. Even if our system does not provide many and complex functions, it provides sufficient to control a device like a TV that can be controlled using a few commands. The proposed system has the merit that it is much cheaper than other motion sensor systems, priced at about \$200. These expensive motion sensors are difficult to commercialize. However, the proposed system can be built for less than \$10. This can compensate for the lack of functions compared to other systems.

References

1. Wachs, J.P., Kolsch, M., Stern, H., Edan, Y.: Vision-Based Hand-Gesture Applications. *Magazine Communication of the ACM* 54(2) (2011)
2. Freeman, A.T., Weissman, C.D.: Television control by hand gesture. In: *IEEE Intl. Workshop on Automatic Face and Gesture Recognition*, Zurich (June 1995)
3. Wilson, A., Oliver, N.: GWindows: Robust Stereo Vision for Gesture-Based Control of Windows. In: *Proceedings of the 5th International Conference on Multimodal Interfaces, ACM ICMI 2003* (2003)
4. Michael, V.D.B., Luc, V.G.: Combining RGB and ToF Cameras for Real-time 3D Hand Gesture Interaction. In: *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 66–72 (2011)
5. Takahashi, M., Fujii, M., Naemura, M., Satoh, S.: Human Gesture Recognition using 3.5-Dimensional Trajectory Features for Hands-Free User Interface. In: *Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS 2010)* (October 2010)
6. Chang, Y.H., Chan, L.W., Ko, J.C., Lee, M.S., Hsu, J., Hung, Y.P.: QPalm: A Gesture Recognition System for Remote Control with List Menu. In: *IEEE International Conference on Ubi-Media Computing*, pp. 20–26 (2008)
7. Chen, M.Y., et al.: Controlling your TV with gestures. In: *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)* (March 2010)
8. Hsieh, C.C., Liou, D.H., Lee, D.: A Real Time Hand Gesture Recognition System Using Motion History Image. In: *Signal Processing Systems (ICSPS)* (2010)
9. Yamada, R., Kuriwa, H., Oka, M., Mori, H.: A Study on Selection Ability in the 3D Space by the Finger. In: *SICE Annual Conference 2010*, pp. 1933–1942 (2010)
10. Cutler, R., Turk, M.: View-based Interpretation of Real-time Optical Flow for Gesture Recognition. In: *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 416–421 (1998)
11. Bernard, A., Bing, B.: Hand Gesture Video Browsing for Broadband-Enabled HDTVs. In: *2010 IEEE Sarnoff Symposium*, pp. 1–5 (2010)
12. Chai, X., Fang, Y., Wang, K.: Robust Hand Gesture Analysis and Application in Gallery Browsing. In: *IEEE International Conference on Multimedia and Expo., ICME 2009*, pp. 938–941 (2009)
13. Xia, L., Fujimura, K.: Hand Gesture Recognition using Depth Data. In: *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 529–534 (2004)
14. Lee, D.W., Lim, J.M., Sunwoo, J., Cho, I.Y., Lee, C.H.: Actual Remote Control: A Universal Remote Control using Hand Motions on a Virtual Menu. *IEEE Transactions on Consumer Electronics* 55(3), 1439–1446 (2009)
15. Kim, S.K., Park, G.Y., Yim, S.H., Choi, S.M., Choi, S.J.: Gesture-Recognizing Hand-Held Interface with Vibrotactile Feedback for 3D Interaction. *IEEE Transactions on Consumer Electronics* 55(3), 1169–1177 (2009)

Fast and Reliable Data Forwarding in Low-Duty-Cycle Wireless Sensor Networks

Junseong Choe¹, Nguyen Phan Khanh Ha¹, Junguye Hong¹, and Hyunseung Choo^{2,*}

¹ College of Information and Communication Engineering,
Sungkyunkwan University, Korea

{cjscmj, npkha, junguye}@skku.edu

² Department of Interaction Science,
Sungkyunkwan University, Korea
choo@ece.skku.ac.kr

Abstract. In this paper, we propose an Enhanced Greedy Forwarding based on low Duty Cycle (GFDC). This novel scheme guarantees reliable and efficient packet transmission by considering a low-duty cycle environment. For the enhancement of the delivery rate and energy efficiency, the existing greedy forwarding schemes forward packets by considering the distance between a destination and the link asymmetry. Energy efficiency is an important problem in Wireless Sensor Networks (WSNs). Most of the energy in WSNs is consumed by radio, and the power consumption for idle listening approximates to the transmission energy. If the radio keeps listening for the incoming packets then, it will cost most of battery energy and the network lifetime decreases. In order to solve this problem, duty-cycle WSNs are developed. However, the high end to end delay may increase due to certain nodes that stay asleep most of the time and wake up asynchronously. This leads to challenges for the development of new data forwarding protocols in low duty-cycle environment. In order to enhance the delivery rate, energy efficiency, and end to end delay, the GFDC uses a path with w (weight) by considering not only the unreliability and asymmetry of wireless links but also the sleep latency problem. Simulation results show that the GFDC improves end to end delay by about 26% and energy efficiency by about 6% compared to MAGF+DC (Duty Cycle).

Keywords: Wireless Sensor Networks, WSNs, Greedy Forwarding, Data Forwarding, Low Duty Cycle, Sleep Latency.

1 Introduction

Wireless Sensor Networks have been used for many long-term applications such as military surveillance [1], infrastructure protection [2], and scientific exploration [3]. Energy consumption in each sensor should be minimized in order to increase the network lifetime. Among many operations of sensors, the energy consumption due to communication is much more critical than that of other computation operations.

* Corresponding author.

Therefore, a reliable and energy-efficient packet transmission is important. Unlike an idealistic model, in a real WSN, due to the effects of fading, attenuation, and interference, link reliability for the packet transmissions can be poor [4, 5]. In order to solve the link unreliability problem, several schemes have been proposed [7, 8, 9, 10]. Moreover, as previous reliable routing schemes do not consider link asymmetry, Expected Transmission Cost (ETC) is a metric that considers the unreliability and asymmetry of wireless link and the distance to the destination node [9].

However, many greedy forwarding schemes have high communication energy and it consumes most of time to be ready for potential incoming packets. This problem is commonly called as idle listening. In order to solve idle listening problem, we consider a low-duty-cycle environment for enhancing the network lifetime. Although many greedy forwarding schemes consider a low-duty-cycle environment, high end to end delay occurred due to the sleep latency. When a node has a packet that is ready to be sent and if all its neighboring nodes are in the sleep state then, the sender should wait until one of its neighbors is in the active state in order to forward its packet. The time spent on waiting for a neighbor to wake up at the sender is called as the sleep latency.

Our proposed scheme, Enhanced Greedy Forwarding based on low Duty Cycle (GFDC), forwards data and ACK packets by using a path with a high weight (w) to solve the sleep latency problem and transmit packets efficiently. The most important object is to select the best next-hop forwarder which has not only the high link quality but also small sleep latency so that the number of retransmissions and communication delay is decreased. Moreover, GFDC reduces end to end delay and maximizes energy efficiency in low-duty-cycle WSNs. Simulation results show that GFDC improves end to end delay by about 26% and packet delivery rate by about 4% and energy efficiency by about 6% compared to MAGF [9].

The rest of the paper is organized as follows. Related work is introduced in Section 2. Section 3 explains assumptions, link model, and energy model of GFDC. Section 4 describes GFDC in detail and Section 5 analyses the simulation results of GFDC and previous routing schemes. Finally, Section 6 concludes the paper.

2 Related Work

One of the popular geographic routing schemes is original greedy for-warding (OGF) [7]. In OGF, each node knows about its geographic in-formation and its neighbors and the source node knows the location of the destination node. OGF selects a neighbor that is closest to the destination as a next forwarding node. It has a benefit of not maintaining state information as it only uses the location information of its neighbors. However, the delivery rate decreases in a real lossy wireless environment as it only considers the distance. Besides, the end to end delay also increases as this forwarding scheme does not consider the sleep latency between the sender node and destination node in low-duty-cycle WSNs.

Some energy efficient routing schemes are proposed by considering the packet loss that depends on the distance in WSNs [8, 9]. In PRR \times Distance greedy forwarding, a node with the highest multiplication value of Packet Reception Rate (PRR) and Distance is selected as the next forwarding node. Distance is about how a forwarding

node gets closer to the destination node. By multiplying PRR and Distance, the scheme maintains the balance between PRR and the distance. Moreover, it solves the unreliable link problem. Nevertheless, $PRR \times \text{Distance}$ greedy forwarding does not take into account the sleep latency in low-duty-cycle environment so that not only high end to end delay goes up but also the network lifetime is decreased.

Multihop ACK-based greedy forwarding (MAGF) [9], forwards data and ACK packets by using a path with a minimum Expected Transmission Cost (ETC) to solve the link asymmetry problems and transmit packets efficiently [9]. MAGF minimizes energy consumption for packet transmission by selecting a neighbor that has the minimum multiplication value of the expected number of packet transmissions and expected hop count to the destination. Moreover, in contrast to previous routing schemes that use the same path for both data and ACK packets, MAGF uses different paths for transmission data and ACK packets which are expected to be the most efficient. With the help of such a path selection, MAGF reduces the packet retransmission cost and maximizes the energy efficiency in using the limited battery resource of sensor node. However, MAGF still does not consider sleep latency in low-duty-cycle environment so that the end to end delay from the sender node to destination node is high.

3 Preliminaries

This section defines the assumptions and network model related to GFDC. They are considered for reliable and energy efficient packet transmission in the low-duty-cycle environment.

3.1 Assumptions

We assume that sensor nodes are homogenous, static once deployed, and they are locally synchronized with their neighbors. Each sensor node knows its own geographic location as well as the location of its neighbors. This information can be obtained by using the GPS module at each sensor or by localization mechanisms. Sensor nodes work in the low duty-cycle mode. One duty-cycle period T of a sensor v is divided into T time slots with equal length τ . One time slot is long enough to be able to send or receive a DATA packet or an ACK packet. A sensor node, v randomly selects one time slot as its active time slot t_v and keeps its radio on to only receive data during that time slot. This is as shown in Fig. 1. In the other time slots, sensor node remains sleep unless it needs to send data. In other words, a sensor node can wake up to transmit a packet at any time but it can receive packets only when it is in its active time slots.

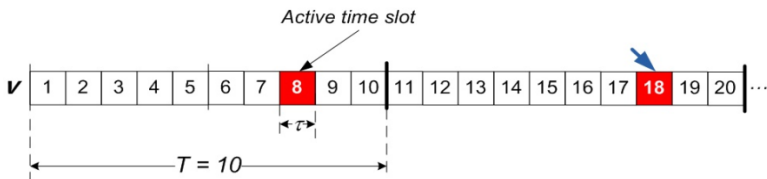


Fig. 1. Working schedule of a sensor node v

3.2 Link Quality Model

In this paper, the unreliability of the wireless links between sensor nodes is modeled by using a realistic link layer mode which is presented in [13]. In the log-normal path loss model, the received signal strength (P_r) at the distance d is calculated as below:

$$P_r = P_t - PL(d) = P_t - PL(d_0) + 10n \log_{10} \left(\frac{d}{d_0} \right) + X_\sigma \quad (1)$$

In Eq. (1) where, P_t is the power of the transmitter, n is the path loss exponent, X_σ is a zero-mean Gaussian (in dB) variable with standard deviation σ , d_0 is the reference distance, and $PL(d_0)$ is the power decay for the reference distance. Given the power of the transmitter P_t , the signal to noise ratio (SNR) (dB) at a distance d is also a random variable:

$$\gamma(d) = P_t - PL(d) - P_n \quad (2)$$

In Eq. (2), P_n : the noise floor is the measure of the signal that is created from the sum of all the noise sources and unwanted signals within a system. This parameter depends on both the radio and the environment. The temperature of the environment influences the thermal noise that is generated by the electronic components. Moreover, the environment can also further influence the noise floor due to the interfering signals.

When Manchester encoding and NCFSK modulation scheme are used, and in the presence of Additive White Gaussian Noise (AWGN), the probability of bit error P_e of the receiver is calculated as below:

$$P_e = \frac{1}{2} \exp \left(-\frac{\gamma(d)}{2} \right) \quad (3)$$

A frame is received successfully if all bits are received correctly. For a frame of length f , the probability of receiving a packet (Packet Reception Rate, PRR) successfully for a distance d between the transmitter and the receiver becomes a random variable. This random variable is given by:

$$PRR = p = (1 - P_e)^{8f} = \left(1 - \frac{1}{2} \exp \left(-\frac{\gamma}{2} \right) \right)^{8f} \quad (4)$$

3.3 Energy Model

When a node in the transmitting or receiving state, the energy consumption depends on the current and the supply voltage [14]:

$$e = P \times T = I \times V \times T \quad (5)$$

In Eq. (5), e is the energy consumption in one state of the sensor, P is the power consumption, V is the supply voltage, I is the current, and T is the time duration which is spent on that state.

$$T_{byte} = \frac{8}{19200} (s) \quad (6)$$

Table 1 shows the power model of Mica2 [15]. For Mica2 mote, the data rate is 19.2kbps. Therefore, the amount of time to transmit one byte, T_{byte} is calculated as below:

Assume that the transmission power is -5dBm, the current is 7.1mA, and the energy consumption to transmit one byte is defined as in Eq. (7):

$$e_t = I_t \times V \times T_{byte} = 7.1 \times 10^{-3} \times 3 \times \frac{8}{19200} = 8.87(\mu J/ \text{byte}) \tag{7}$$

Table 1. Power model of Mica2

Mode	Current	Mode	Current
CPU		Write	18.4 mA
Active	8.0 mA	Write Time	12.9 ms
Idle	3.2 mA	Radio	
ADC Noise Reduce	1.0 mA	Rx	7.0 mA
Power-down	103 μ A	Tx (-20 dBm)	3.7 mA
Power-save	110 μ A	Tx (-19 dBm)	5.2 mA
Standby	216 μ A	Tx (-15 dBm)	5.4 mA
Extended Standby	223 μ A	Tx (-8 dBm)	6.5 mA
Internal Oscillator	0.93 mA	Tx (-5 dBm)	7.1 mA
LEDs	2.2 mA	Tx (0 dBm)	8.5 mA
Sensor Board	0.7 mA	Tx (+4 dBm)	11.6 mA
EEPROM access		Tx (+6 dBm)	13.8 mA
Read	6.2 mA	Tx (+8 dBm)	17.4 mA
Read Time	565 μ s	Tx (+10 dBm)	21.5 mA

The energy consumption to receive one byte is calculated as shown below:

$$e_r = I_r \times V \times T_{byte} = 7.0 \times 10^{-3} \times 3 \times \frac{8}{19200} = 8.87(\mu J/ \text{byte}) \tag{8}$$

The data packet size is 100 bytes and the ACK packet size is 11 bytes. Therefore, the energy consumption to transmit and receive a data packet is defined as given in Eq. (9):

$$e_{data} = 100 \times (e_t + e_r) = 1762(\mu J) \tag{9}$$

The energy consumption to transmit and receive an ACK packet is calculated as below:

$$e_{ACK} = 11 \times (e_t + e_r) = 193.82(\mu J) \tag{10}$$

4 Proposed Scheme

4.1 Motivation

Sleep-wake scheduling (or duty-cycling) of radio transceivers is a well-known technique for achieving energy conservation. The duty-cycle scheduling aims to prolong the network lifetime by using the method of making the sensor nodes sleep for the most time and wake up in a small portion of the working period. For e.g. the working period is 100 units of time, the duty-cycle is 1%, the sensor is active (wake) in 1 unit of time and sleep in 99 remaining units of time in this working period.

However, the main disadvantage of the low-duty-cycle WSNs is the high sleep latency. When a node has a packet that is ready to be sent but all of its neighboring nodes are in the dormant state, the sender has to wait for one of its neighbors to wake up in order to forward its packet. The time spent on waiting for a neighbor to wake up at the sender is called as the sleep latency. The sleep latency is much longer compared to other delivery latencies such as processing delay, transmission delay, and propagation delay. Consequently, the sleep latency dominates the delay in low duty-cycle WSNs. Besides, many applications such as real time applications, military surveillance, disaster response applications, and so on require the data which has to be received in a time deadline. The authors argue that the link quality and the duty cycle of the sensor nodes can significantly impact on the data delivery [12]. Indeed, in the previous data forwarding protocols, the next hop forwarder is selected greedily based on its advance distance to the destination. However, in certain cases, the neighbor which has the best advance distance has the high sleep latency with the source node. This neighbor is selected as the next hop forwarder and it leads to the high end to end delay. If the end to end delay is greater than the time constraint for some specific applications then, the information that is contained in the data is useless although the data is received successfully at the destination. This problem becomes more and more severe in the wireless environment with lossy radio links. Due to the retransmissions, the latency increases.

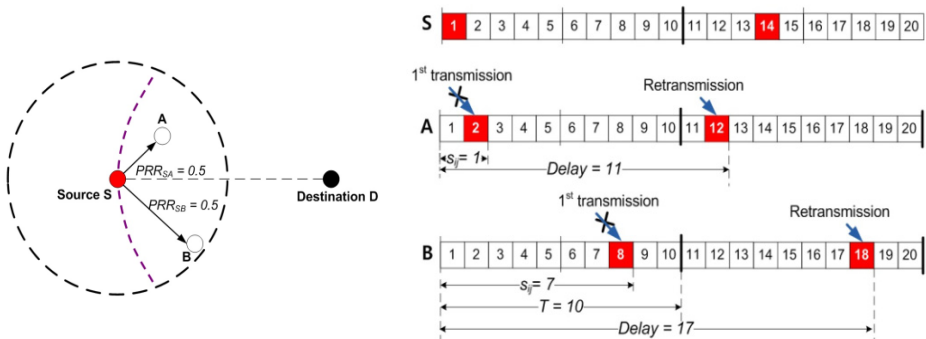


Fig. 2. Sleep latency in low-duty-cycle WSNs

For an instance, in Fig. 2, node B has better advance distance to destination D than node A but the sleep latency of node B with the source S is 7 time slots while that of node A is just 1 time slot. Assuming the link quality between source S and node A and the link quality between source S and node B are the same (0.5), the average number of transmissions from source S to node A and node B is 2. If node B is selected as the next hop forwarder of source S then, the average delay is 17 time slots. On the other hand, if node A is selected as the next hop forwarder of source S, the average delay is just 11 time slots. For some specific applications that require the delay is not greater than 15 time slots, if node B is selected for data forwarding, the data transmission is considered fail.

There are new challenges for the development of new data forwarding protocols which can reduce the effect of high sleep latency in low duty-cycle WSNs. Therefore, in this paper, we consider a forwarding scheme called as Enhanced Greedy Forwarding based Duty Cycle (GFDC) to select the next-hop forwarder based on the sensor’s duty cycle. Moreover, the link quality between the forwarding node and its neighbors, and the advance distance of the neighbor is compared with the forwarding node. The objective is to select the best next-hop forwarder that has both the high link quality and the small sleep latency so that the number of retransmissions and the communication delay are decreased. Consequently, the data delivery ratio and the energy efficiency are increased.

4.2 Calculating the Sleep Latency

The sleep latency is calculated by the following equation:

$$S_{i,j} = \begin{cases} (A_j - A_i) \text{ mod } T & \text{if } A_i \neq A_j \\ T & \text{if } A_i = A_j \end{cases} \quad (11)$$

In Eq. (11), $S_{i,j}$ is the sleep latency for data forwarding between node i and node j ; T is the working period; and A_i and A_j are the active time slots of node i and node j respectively. Fig. 3(a) shows the sleep latency when two nodes have the same active time slots. Fig. 3(b) presents the sleep latency between node i and node j when they have different active time slots.

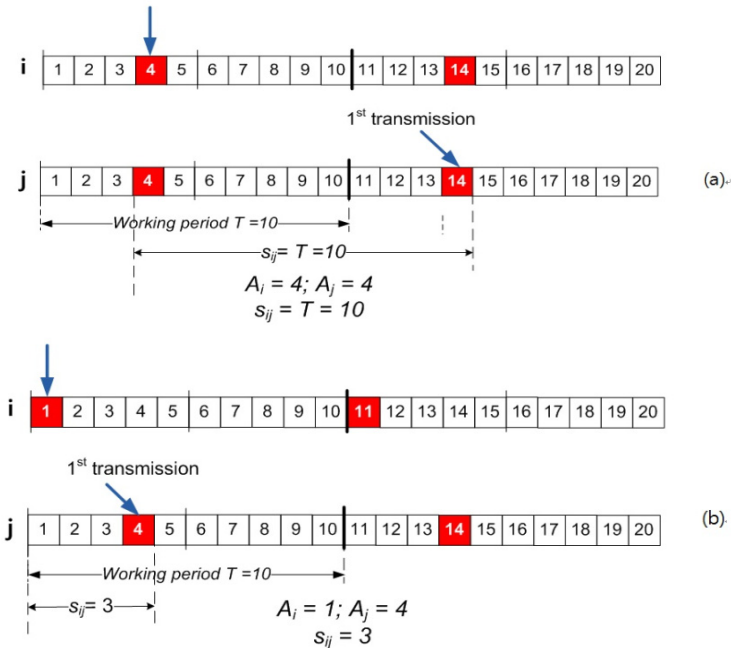


Fig. 3. Calculating sleep latency between two sensor nodes

4.3 Forwarding Metric

The sleep latency is calculated by using the following equation:

$$w = \left(\alpha \times \frac{1}{ED} \right) + (\beta \times PRR \times D) \tag{12}$$

The main idea is that we want select the neighbor which has the smallest sleep latency and the link quality as well as is the nearest node with the destination as the next-hop forwarder. In Eq. (12), PRR is the packet reception rate of a successful transmission (including a successful data forwarding and ACK transmission) and it is calculated as: $PRR = PRR_{forwarding} \times PRR_{ACK}$. The expected number of transmission is: $= \frac{1}{PRR}$. The expected delay of the transmission from the current forwarding node to its neighbor, ED is calculated as follows:

$$ED = \begin{cases} S_{i,j} + (ET - 1) \times T & \text{if } A_j \neq A_i \\ S_{i,j} + (ET \times 1) = (ET + 1) \times T & \text{if } A_j = A_i \end{cases} \tag{13}$$

In Eq. (13), the neighbor, which has the smaller sleep latency and higher link quality, is given high priority to be selected as the next-hop forwarder. Fig. 4 shows the expected delay for one hop transmission between the two sensor nodes when the link quality between them is $PRR_{i,j} = 0.5$.

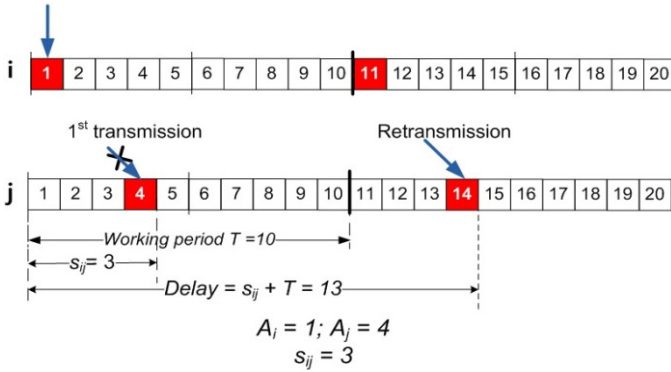


Fig. 4. The expected delay for one hop transmission

AD is the advance distance of the neighbor, $AD = 1 - \frac{d(\text{neighbor,destination})}{d(\text{forwarding node,destination})}$ with, $d(a,b)$ is the Euclidean distance between two nodes a and b . Fig. 5 illustrates how to calculate the advance distance of a neighbor e of the source node s .

α, β are two coefficients that denote the effect of the sleep latency, and link quality and the advance distance of a neighbor respectively. If the value of α is high, the neighbor with the high link quality and small sleep latency has higher priority compared to other neighbors that are nearer to the destination but have low link quality and high sleep latency. Otherwise, if the value of β is high, the neighbors that are nearer to the destination have higher opportunity to become the next-hop

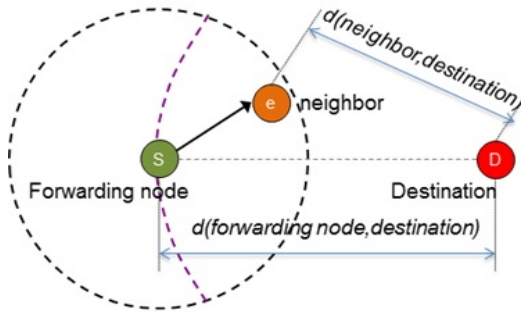


Fig. 5. Calculating the advance distance of a neighbor node

forwarder. Based on Eq. (12), the forwarding node selects among its neighbors the sensor node which has the highest value of w as the next hop-forwarder.

5 Performance Evaluation

In this section, we compare the performance of GFDC with that of OGF + DC (Duty Cycle), PRR×Distance greedy forwarding + DC (Duty Cycle) and MAGF+DC (Duty Cycle) while varying the node density and duty cycle. For each simulation run, 50 source and destination pairs are randomly chosen to send one data packet and the results are computed as the average of 1000 runs. Nodes are randomly deployed on the network topology and the node density is defined as the number of nodes that are within the transmission range. Automatic repeat request (ARQ) is 10 and a neighbor node is blacklisted when it has a link between itself and the current forwarding node with PRR less than 0.01. Parameters used in the simulation are shown in table 2 and three metrics are used to show the performance enhancement of GFDC:

- *Delivery Rate*: the ratio of packets sent from source node to packets received by destination node, ranging from 0 to 1
- *Energy Efficiency*: the amount of data (bits) delivered to destination node per unit energy (bits/mJ)
- *End to End Delay (time slots)*: number of time slots are calculated when the destination node receives the data packet from the source node

In this section, end to end delay, delivery rate and energy efficiency of OGF+DC, PRR×Distance greedy forwarding+DC, and MAGF+DC are compared at different node densities and duty cycles. In the subsection 5.1, 100 nodes are randomly deployed on the network and the three schemes’ performances are compared by varying the node density from 10 to 500. In subsection 5.2, we fix the node density as 100 and compare the four schemes’ performances by varying the duty cycles from 1 to 20%.

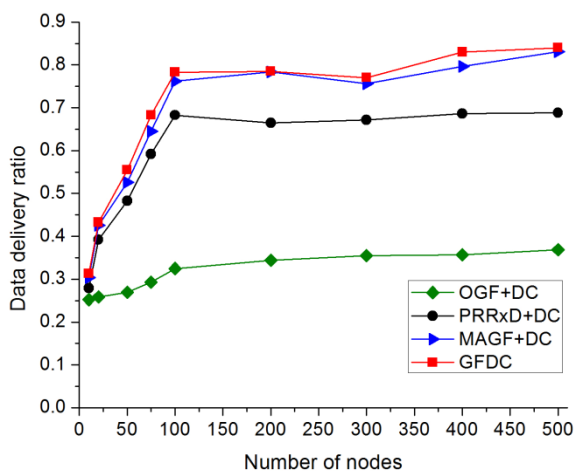
Table 2. Simulation parameters

Radio Parameters			
Modulation	NC-FSK	Encoding	Manchester
Output Power	-5 dBm	Path Loss Exp	3
Noise Floor	-105 dBm	Data Rate	1.92 kbps
Packet Size and Energy Consumption			
Data Size	100 bytes	e_{DATA}	1762 μ J
ACK Size	11 bytes	e_{ACK}	193.82 μ J

5.1 Comparison of Delivery Rate

Fig. 6 shows the delivery ratio with different number of sensor nodes for four algorithms: OGF+DC, PRR \times Distance+DC, MAGF+DC, and the proposed algorithm GFDC when the duty cycle is fixed to 5%. With all the number of sensor nodes in the network, the GFDC outperforms other three algorithms in terms of delivery ratio. When the number of sensor nodes is small, the number of neighbors for each node is small so that the next forwarding node selection of three algorithms is similar in many cases. Therefore, the difference of data delivery ratio of three algorithms is small.

However, when the number of sensor nodes increases, the node that uses GFDC can select a better next forwarding node among its neighbors. The end to end delay

**Fig. 6.** Comparison of Delivery Rate with Different Node Densities

transmission decreases so that the number of data transmissions which miss the application delay time decreases. This leads to a high data delivery ratio compared to other two algorithms. The data delivery ratio of PRR×Distance+DC is also higher than that of OGF+DC in duty-cycle environment. This is due to the PRR×Distance+DC algorithm considers the link quality between the source node and its neighbor when selecting the next forwarding node. It selects the neighbor with high link quality as the next forwarding node. Hence, the number of lost packets drops and the data delivery ratio rises.

Fig. 7 illustrates the data delivery ratio of three algorithms with a variation in the duty cycles. With all values of the duty-cycle, the data delivery ratio of GFDC is the highest one. However, the difference between GFDC and MAGF+DC is smaller when the duty-cycle is high. Because when the duty-cycle is high, the effect of the sleep latency is not too much.

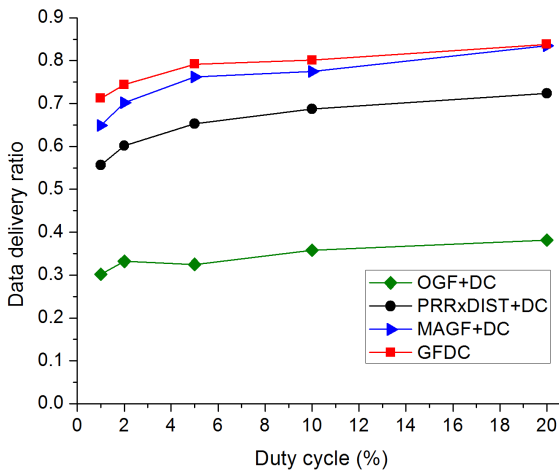


Fig. 7. Comparison of Delivery Rate with Different Duty Cycles

5.2 Comparison of Energy Efficiency

Figure 8 shows the energy efficiency (bits/mJ) of four algorithms when the number of sensors nodes varies from 10 to 500 nodes. When the number of nodes is small, the delivery ratio is low so that the energy efficiency of the three algorithms is small. When the number of nodes increases, GFDC and MAGF+DC have more opportunities to select the best neighbor as the next forwarding nodes. Hence, the data delivery ratio as well as the energy efficiency increases.

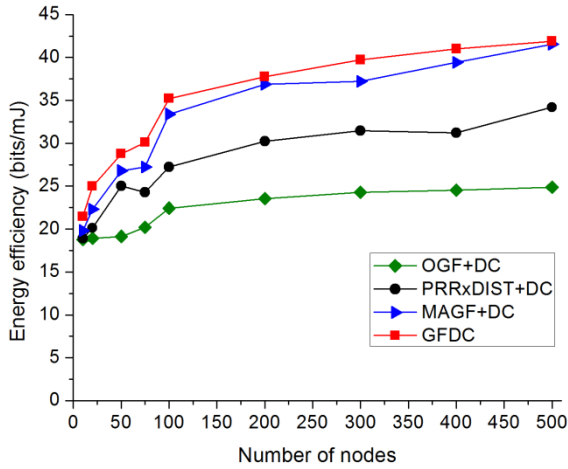


Fig. 8. Comparison of Energy Efficiency with Different Node Densities

Fig. 9 illustrates the energy efficiency (bits/mJ) of three algorithms with the variation of the duty cycles. With all values of the duty-cycle, the energy efficiency of GFDC is the highest one. However, the difference between GFDC and the other schemes is bigger when the duty-cycle is low such as duty rate is 1%. Because when the duty-cycle is low, the effect of the sleep latency is increased.

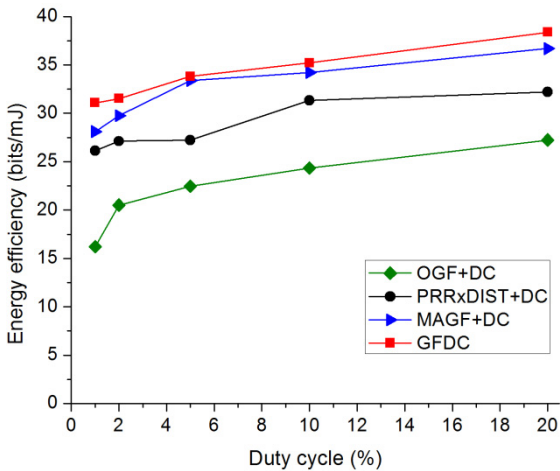


Fig. 9. Comparison of Energy Efficiency with Different Duty Cycles

5.3 Comparison of End to End delay

Fig. 10 presents the comparison of end to end delay of three algorithms when the duty cycle is 5% and the number of sensor nodes varies from 10 to 500 nodes. The end to end delay of OGF+DC, PRR×Distance+DC, and MAGF+DC does not change significantly when the number of sensor nodes increases while that of GFDC decreases steeply when the number of sensor nodes increases from 10 to 100 nodes. The reason for this is, when the number of nodes increases, a source node that uses GFDC can select a neighbor which balances the sleep latency, link quality, and advance distance as the next forwarding so that the end to end transmission delay will decrease. When the number of nodes continues to increase, because the duty cycle is fixed, the effect of number of the nodes is not too much and so, the end to end delays of GFDC are quite similar while the number of nodes changes from 100 to 500 nodes.

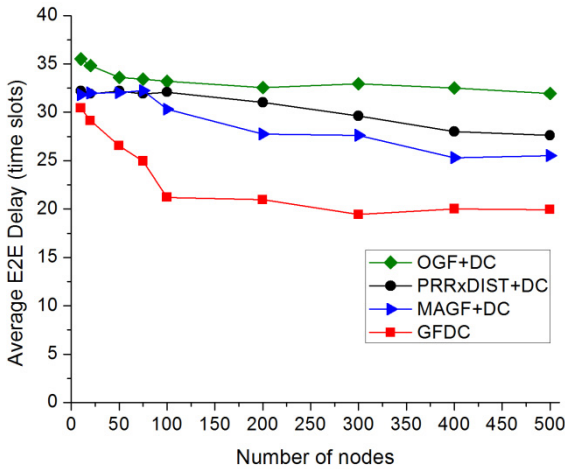


Fig. 10. Comparison of E2E Delay with Different Node Densities

In Fig. 11, by taking into account the sleep latency in the forwarding metric, GFDC can reduce the average end to end delay especially when the duty cycle is small (the length of the working period is high). For example, with a duty cycle of 2%, the average end to end delay for OGF+DC, PRR×Distance+DC, MAGF+DC, and GFDC are 168, 152, 171, and 121, respectively. When the duty cycle increases, the effect of sleep latency for the data transmission in each hop decreases. Therefore, the average end to end delay decreases for the four algorithms.

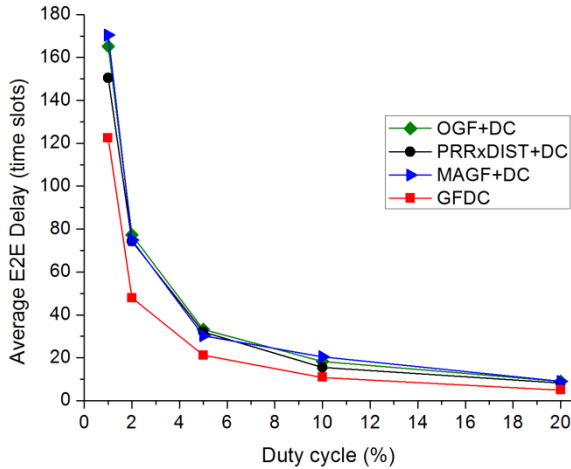


Fig. 11. Comparison of E2E Delay with Different Duty Cycles

6 Conclusion

We proposed a reliable and energy efficient forwarding scheme called as Enhanced Greedy Forwarding based on low Duty Cycle (GFDC) in low-duty-cycle WSNs. GFDC selects the forwarding node by considering the neighbor node information along with the transmission range. In order to select forwarding node, each node calculates weight (w) based on not only the sleep latency between the sender and receiver node but also on the link unreliability and asymmetry and the advance distance to destination node. Then, it selects a neighbor node with high weight (w) as the next forwarding node. The objective of GFDC is to solve the problems that the previous schemes have in low-duty-cycle WSNs. As a result, fast dissemination and energy consumption that are related to the packet transmission is guaranteed. In the future, we shall extend our work by applying the flexible backoff scheduling in order to reduce collision. Therefore, further research can provide better performance.

Acknowledgement. This research was supported in part by MKE and MEST, Korean government, under ITRC NIPA-2012-(H0301-12-3001), WCU NRF (No. R31-2010-000-10062-0) and PRCP(2011-0018397) through NRF, respectively.

References

1. Szewczyk, R., Mainwaring, A., Anderson, J., Culler, D.: An Analysis of a Large Scale Habit Monitoring Application. In: *SenSys 2004* (2004)
2. Xu, N., Rangwala, S., Chintalapudi, K.K., Ganesan, D., Broad, A., Govindan, R., Estrin, D.: A Wireless Sensor Network for Structural Monitoring. In: *SenSys 2004* (2004)

3. Tolle, G., Polastre, J., Szewczyk, R., Turner, N., Tu, K., Burgess, S., Gay, D., Buonadonna, P., Hong, W., Dawson, T., Culler, D.: A Macroscopic in the Redwoods. In: *SenSys 2005* (2005)
4. Zhao, J., Govindan, R.: Understanding Packet Delivery Performance in Dense Wireless Sensor Networks. In: *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, pp. 1–13 (2003)
5. Ganesan, D., Krishnamachari, B., Woo, A., Culler, D., Estrin, D., Wicker, S.: Complex Behavior at Scale: An Experimental Study of Low-Power Wireless Sensor Networks. Technical Report UCLA/CSD-TR 02-0013 (2002)
6. Lee, S., Bhattacharjee, B., Banerjee, S.: Efficient Geographic Routing in Multihop Wireless Networks. In: *Proceedings of the 6th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 230–241 (2005)
7. Zamalloa, M.Z., Seada, K., Krishnamachari, B., Helmy, A.: Efficient Geographic Routing over Lossy Links in Wireless Sensor Networks. *ACM Transactions on Sensor Networks* 4(3), article No. 12 (2008)
8. Yağan, O., Makowski, A.M.: Designing securely connected wireless sensor networks in the presence of unreliable links. In: *Proceedings of the IEEE International Conference on Communications (ICC 2011)*, Kyoto, Japan (June 2011)
9. Bae, D., Choi, W., Choo, H.: Multihop ACK-based greedy forwarding using expected transmission cost in wireless sensor networks. In: *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, ICUIMC 2011*, vol. (119) (2011)
10. Yousefi, H., Yeganeh, M.H., Movaghar, A.: Long lifetime routing in unreliable wireless sensor networks. In: *Proceedings of the 8th IEEE International Conference on Networking, Sensing and Control (ICNSC 2011)*, pp. 457–462 (2011)
11. Zamalloa, M.Z., Seada, K., Krishnamachari, B., Helmy, A.: Efficient Geographic Routing over Lossy Links in Wireless Sensor Networks. *ACM Transactions on Sensor Networks* 4(3), article No. 12 (2008)
12. Gu, Y., He, T.: Data Forwarding in Extremely Low-Duty-Cycle Sensor Networks with Unreliable Communication Links. In: *The 5th ACM Conference on Embedded Networked Sensor Systems, Sensys 2007* (2007)
13. Zuniga, M., Krishnamachari, B.: Link Layer Models for Wireless Sensor Networks. Tutorial, <http://anrg.usc.edu/www/downloads/LinkModellingTutorial.pdf>
14. Nguyen, D.T., Choi, W., Choo, H.: PDF: A Novel Probability-based Data Forwarding Scheme in Lossy Wireless Sensor Networks. In: *Reliable and Autonomous Computational Science* (2010)
15. Shnayder, V., Hempstead, M., Chen, B., Allen, G.W., Welsh, M.: Simulating the Power Consumption of Large-Scale Sensor Network Applications. In: *Proc. Sensys 2004* (2004)

Neural Network and Physiological Parameters Based Control of Artificial Pancreas for Improved Patient Safety

Saad Bin Qaisar, Salman H. Khan*, and Sahar Imtiaz

Department of Electrical Engineering,
National University of Sciences and Technology,
Islamabad, Pakistan
{saad.qaisar,salman.khan,1mseeimtiazh}@seecs.edu.pk

Abstract. Acyber-physical system (CPS) establishes a close interaction between system's computational core and the control of physical process. In case of Diabetes, failure in endogenous insulin production requires exogenous infusion of required drug amount. We have proposed an architecture for artificial pancreas and checked its validity in simulations. The aim is to control blood glucose level (BGL) of a patient suffering from diabetes and to prevent the harmful state of Hypoglycemia. For this, vital signs monitoring is introduced through which hypoglycemic condition can be efficiently detected and avoided. Electrocardiogram, Heart beat rate, Electroencephalography and skin resistance are known to depict an irregularity in blood glucose. Upon detection, a specified amount of Glucagon is infused into patient's body. The system consists of an insulin infusion and glucagon pump, through which insulin/glucagon is entered into the patient's body subcutaneously, based on the current BGL. A neural network predictive controller is designed to keep the glucose level inside the desired 'safe range'. The simulations have shown that patient safety can be improved through this strategy.

Keywords: diabetes, insulin infusion, exogenous, endogenous, feedback control, simulation, neural network.

1 Introduction

High confidence Medical cyber physical systems (MCPS) are the future of Medical device Industry. *Lee and Sokolsky* have identified physiological closed loop systems, continuous monitoring and care systems with real time diagnostic and treatment capability as a new trend in Medical devices [1]. Model based development, Safety and Patient modeling and simulation are identified as key challenges in MCPS [1,2].

Diabetes Mellitus is an incurable metabolic disorder plaguing lives of millions across the world. According to WHO report, today more than 177 million people

* Corresponding author.

are suffering with diabetes, and this number is expected to rise to 300 million by 2025 [15]. It is a result of partial or complete failure of the human pancreas to produce insulin to counter elevated blood glucose levels (BGL). Extremely crucial and chronic complications of diabetes which can cause permanent pathological changes such as blindness, amputations, kidney problems, strokes, cardiovascular diseases and non-traumatic limb amputation are common. Elevated blood glucose level for long time is the basic reason of all these adverse effects [5].

The vast majority of diabetes cases fall into three broad three aetiopathogenetic categories: Type 1, Type 2 and gestational diabetes. In Type 1 Diabetes Mellitus (T1DM) insulin producing pancreatic beta cells are destroyed because of an autoimmune pathologic process occurring in the pancreatic islets. Type 1 diabetes eventually leads the patient to acute insulin deficiency and build up ketones, thus increasing a risk for diabetic ketoacidosis [5]. The patient is totally dependent on externally infused insulin at an appropriate rate to maintain blood glucose level. The blood glucose level must be contained within the range of 60-120mg/dl and the ideal value is 81mg/dl [4]. Hyperglycemia state occurs when the plasma glucose level rise above 120mg/dl while in an opposite situation plasma glucose levels fall below 60mg/dl which is known as hypoglycemia state. Both these states are harmful, but hypoglycemia is a more serious threat and can cause coma or may be even fatal. Hyperglycemia state causes problems in long run.

A fully functional Artificial Pancreas is close to become a reality [20,21]. Such an ambulatory closed-loop system capable of maintaining normoglycemia for long periods can dramatically enhance the quality of life of T1DM patients [17]. Continuous glucose monitoring and portable insulin infusion solutions with small form factor are available today [22]. Subcutaneous route (SC) is the least invasive and most secure solution for insulin delivery and BGL measurement that gives it an edge over intravenous route, which is best suited for control [13]. System setups using subcutaneous sensing and infusion have been successfully studied and implemented in past [18,19].

Although the concept of closed loop insulin control for T1DM has got older yet it is not fully optimized. Traditional automated insulin delivery system compose of: (a) A continuous glucose sensing device (systems afferent part), (b) A drug infusion pump (systems efferent part), and (c) A controller/algorithm linking the measured blood glucose concentration and insulin delivery [24]. Fig 1 shows block diagram of such a system.

One of the major problems with this architecture is that many T1DM patients have irregular and insufficient release of the pancreatic counter-regulatory endocrine hormone glucagon, which causes extended hypoglycemia episodes which can be fatal. Moreover due to time delay in insulin action via subcutaneous route, over dose of insulin is possible because of elevated glucose levels after meals, which can lead patient to the hypoglycemia. Though closed loop treatments in clinical environments have greatly enhanced diabetes control and care over the last few decades [5], hypoglycemia remains a major unsolved problem to achieve strict glycaemic control [5,7].

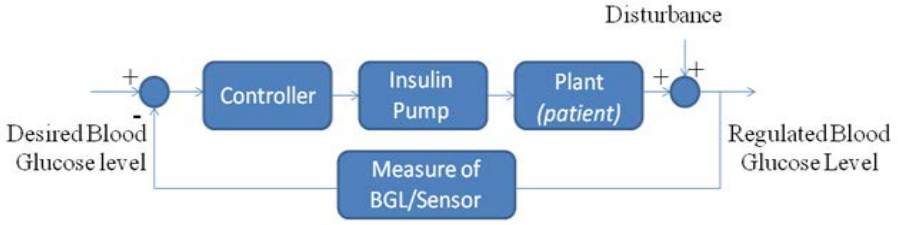


Fig. 1. Block Diagram of Conventional Artificial Pancreas

Continuous vital sign monitoring (VSM) can be used to detect hypoglycemia, which can eventually cause coma, seizure or death. Electroencephalography (EEG), Electrocardiography (ECG), heart rate and skin resistance are important in detecting hypoglycemia condition. For an ambulatory system ECG, heart rate and skin resistance are more important and practical parameters to detect hypoglycemia. In this paper, we have proposed an architecture in which any irregularity above a specified threshold in all the three vital signs will trigger a signal from controller which in turn will infuse glucagon in to patients body. Fig. 2 shows block diagram of system.

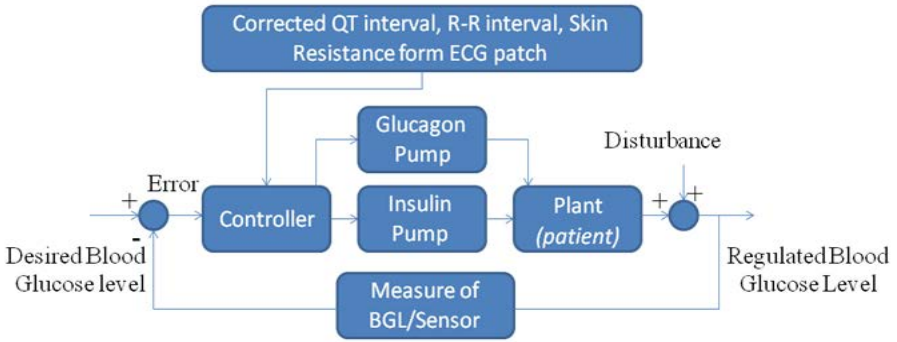


Fig. 2. Block Diagram of Modified System

2 Materials and Methods

2.1 Mathematical Modeling

Mathematical modeling of glucose-insulin interaction is a useful tool in testing the functionality of control algorithms [25]. Various methods have been developed for mathematical modeling of glucose insulin interactions [16] but Bergman minimal model is widely used in physiological research to approximate the blood

glucose and insulin interaction dynamics. It defines an important constant parameter which is product of amount of insulin released and sensed, called disposition index [11]. It assumes closed-loop relationship between a minimum number of three compartments which are described by following equations [3]:

$$\frac{dG(t)}{dt} = -p_1[G(t) - G_b] - X(t)G(t) + [D(t) + C(t)] \tag{1}$$

$$\frac{dX(t)}{dt} = -p_2X(t) + p_3[I(t) - I_b] \tag{2}$$

$$\frac{dI(t)}{dt} = -n[I(t) - I_b] + \gamma[G(t) - h]^+t + r(t) \tag{3}$$

where $G(t)$ is the instantaneous glucose concentration in blood (mg/dl), $X(t)$ is the effective amount of insulin used in disappearing plasma glucose(min^{-1}) and $I(t)$ is the instantaneous insulin concentration in plasma ($\mu\text{U/ml}$). G_b and I_b are the concentration of glucose and insulin in blood before any exogenous infusion of either of these two. This lower concentration is called as basal level. p_1 , p_2 and p_3 are the model parameters and n is the rate at which insulin is being used up in plasma(min^{-1}), h is the lowest value of blood glucose above which endogenous insulin is secreted (mg/dl), γ is the rate of endogenous release of insulin when glucose is infused exogenously and concentration of blood glucose is above h threshold [$(\mu\text{U/ml min}^{-2} (\text{mg/dl})^{-1})$]. The values of all above quantities for an average person are listed in table 1 [29].

Table 1. Modified Bgregman Model parameters

Parameters	Values
p_1	0.0337 min^{-1}
p_2	0.0209 min^{-1}
p_3	$7.510^{-6} \text{min}^{-2} (\mu\text{U/ml})^{-1}$
X	0.0054 $\mu\text{U/ml}$
$\overline{G} \overline{G} \overline{G} \overline{G}$	0.81 mg/ml
G_b	0.811 mg/ml
n	0.214 min^{-1}
T	5 min

$D(t)$ (mg/dl/min) is the disturbance to the patient and it accounts for the rate at which glucose is absorbed from the intestine to blood after food intake. $C(t)$ (mg/dl/min) is the rate of exogenous glucose (glucagon) infusion which is desired form of disturbance to avoid hypoglycemia. $r(t)$ ($\mu\text{U/ml/min}$) is the controller’s output and it compensates for any disturbances and acts so as to maintain BGL equal to basal level.

Plasma insulin compartment is represented by eq.(3), Dynamic insulin compartment by eq.(2) and Plasma glucose compartment by eq.(1). All interactions are shown in block diagram of Fig. 3.

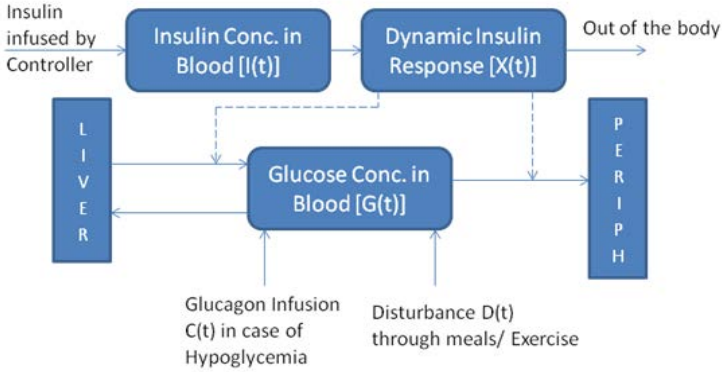


Fig. 3. Block Diagram of Bergman Minimal Model

A modified form of eq. (3) in which endogenous insulin secretion in the original nonlinear minimal model [3] is replaced by the term $\tau.r(t)$. The conversion factor τ converts the units of $r(t)$ to U/h, which are consistent with clinical convention of insulin delivery rate prescription [28]:

$$\frac{dI(t)}{dt} = -n[I(t) - I_b] + \tau.r(t) \tag{4}$$

In order to further linearize the relationships, linear form of Bergman minimal model is proposed in [28]:

$$\frac{dG(t)}{dt} = -p_1[G(t) - G_b] - \bar{X}.G(t) - \overline{GGGGG}.X(t) + \overline{GGGG}.\bar{X} + [D(t) + C(t)] \tag{5}$$

$$\frac{dX(t)}{dt} = -p_2X(t) + p_3[I(t) - I_b] \tag{6}$$

$$\frac{dI(t)}{dt} = -n[I(t) - I_b] + \tau.r(t) \tag{7}$$

Where \overline{GGG} and \bar{X} are the average values of $G(t)$ and $X(t)$. By putting $G(t)$ for \overline{GGGG} and $X(t)$ for \bar{X} we can easily obtain the original Bergman model from eqs.(5), (6) and (7).

2.2 Closed Loop System

Continuous glucose monitoring and automatic insulin delivery are integrated into one system to establish a closed loop control. Closed loop glycemic treatment leads to improved diabetes control [8,9].

Various authors have proposed and practically implemented the closed loop insulin delivery system [14,23].

2.3 Neural Network Predictive Controller

A neural network predictive controller is a type of model predictive control (MPC). It is based on receding horizon technique and is useful in estimating the future response of a dynamic nonlinear system and thus controlling its output in the desired manner. This strategy is useful because it takes in to account the nonlinearities of the plant and effectively models the system with multilayer perceptrons. We have used Neural Network Toolbox in Matlab for designing the predictive controller [31].

The controller generates a neural network model of plant and then predicts the future behavior of system. Based on this prediction a control output is calculated that optimizes the plant performance over the future horizon. The first step of calculating the neural network plant model is termed as System Identification. It involves training a neural network to see the future behavior of plant. The prediction error between the actual plant output and the neural network predicted output is used as the training signal for neural network. We have used two layer neural network for predicting the plant output. The network uses past inputs and outputs to estimate future profile of the plant output. Once future output is estimated, the following equation is used to produce such control signal u that minimizes the cost function J over a specified horizon

$$J = \sum_{j=N_1}^{N_2} (y_r(t+j) - y_m(t+j))^2 + \rho \sum_{j=1}^{N_u} (u'(t+j-1) - u'(t+j-2))^2 \quad (8)$$

The network can be trained either online or off line but we have preferred to train our neural network in offline batch mode. [26,30]

2.4 Subcutaneous Route

Intravenous (IV) route is best suited from control perspective and is investigated in small studies [27] but it is not used because it accompanies with severe risks such as cellulitis, thrombosis, edema, phlebitis, sepsis, embolization, and intravascular infection [10]. In contrast, subcutaneous (SC) route for insulin delivery is much safer and easy to manage [13].

However, SC is not ideal and obtaining a satisfactory glycaemic control is tricky because of the time delay in absorption by this route. The recent introduction of the rapid acting insulin analogs, like Lispro insulin with 2-3 times faster absorption, has significantly improved the quality of control obtainable through this route [13]. Even then the SC infusion confers risks when used for closed-loop BGL management because of efferent delays. Meals intake and higher glucose levels can cause overdose of drug due to these delays, which in turn can cause hypoglycemia. Weinzimer et.al have proposed a hybrid approach to tackle this problem and have shown tighter glycaemic control than fully closed loop treatment [11]. But this strategy requires detailed meal timings/ food intake schedule of each patient which can be risky in case of any deviation in plan. Moreover, specification of food intake is a burden for patient in normal daily life.

2.5 ECG Patch

We propose that a continuous monitoring ECG patch must be incorporated in the closed loop system so that Hypoglycemia can be detected immediately and proper amount of glucose can be entered to keep the BGL in safe range. The ECG patch will be continuously monitoring QT interval, R-R peak distance and skin resistance through any one of its contacts on human body.

3 Vital Signs for Hypoglycemia Detection

Hypoglycemia is a major limiting factor in achieving high quality diabetes control. Hypoglycemia can cause coma or even death. It mainly affects the central nervous system and sweating and arrhythmia in cardiac response are associated with hypoglycemia [34, 35]. EEG, ECG and skin resistance have found to be affected by low BGL (<60 mg/dl) [36]. Since EEG cannot be monitored continuously so QT and RR interval in ECG along with skin resistance are more important in hypoglycemia detection. It is found that during hypoglycemia episodes heart rate of patients increases (correlation of 1.02 ± 0.26 vs. 1.07 ± 0.31 , $P < 0.053$), corrected QT interval increases (correlation of 1.03 ± 0.08 vs. 1.05 ± 0.08 , $P < 0.002$) and skin impedance decreased 15 ± 6 min after hypoglycemia [33, 39]. In another study heart rate was noted to deviate from 72 ± 9 to 80 ± 11 bpm, with $P < 0.01$ [37].

From the ECG signal R-R peaks distance will give the heart rate and corrected QT interval is found by calculating the distance between the start of Q to the end of T interval. We used Bazett's formula for correction given by,

$$QT_c = \frac{QT}{\sqrt{RR}} \quad (9)$$

The infusion of glucagon in case of hypoglycemia episode significantly reduces the time spent in hypoglycemic range (15 ± 6 vs. 40 ± 10 min/day, $P = 0.04$) [38].

4 Modeling and Simulation

We modeled our proposed system in Matlab Simulink. The patient model is shown in Fig. 4 which is based on laplace transform of eqs. (5),(6) and (7) given by,

$$G(s) = T(s)_1 R(s) + T(s)_2 [D(s) + C(s)] + T(s)_3 I(s) \quad (10)$$

Where $I(s)$ is Laplace transform of impulse and $T(s)$ are the transfer functions.

ECG signal was generated in Matlab artificially. Disturbances due to meals or due to exercise are modeled by adding the various terms of the form:

$$H = H_0 \times e^{-a[\log(bt)-c]^2} \quad (11)$$

Where H_0 is the maximum value and constants a , b and c set the shape of the curve. Hypoglycemia is detected once all three vital signs are showing abnormality.

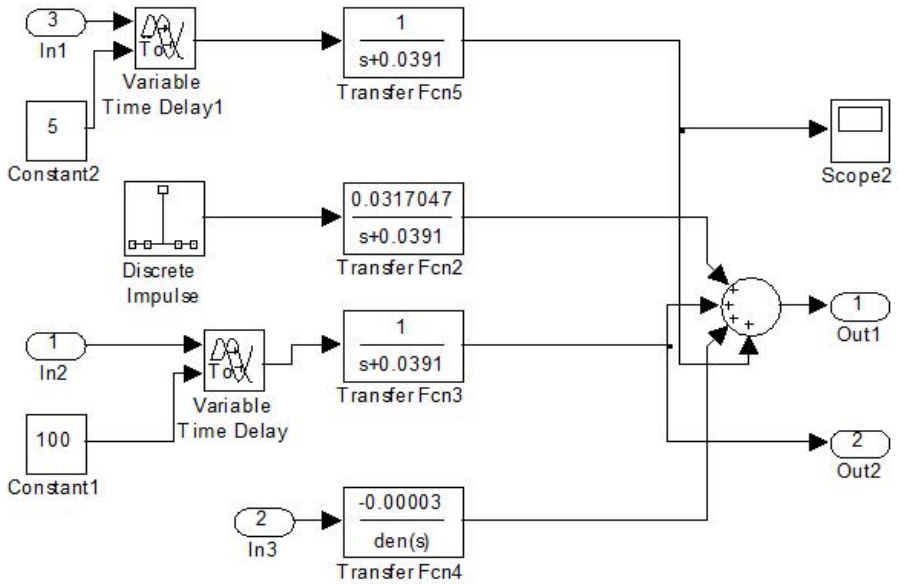


Fig. 4. Patient Model Based on eqs. (5), (6) and (7)

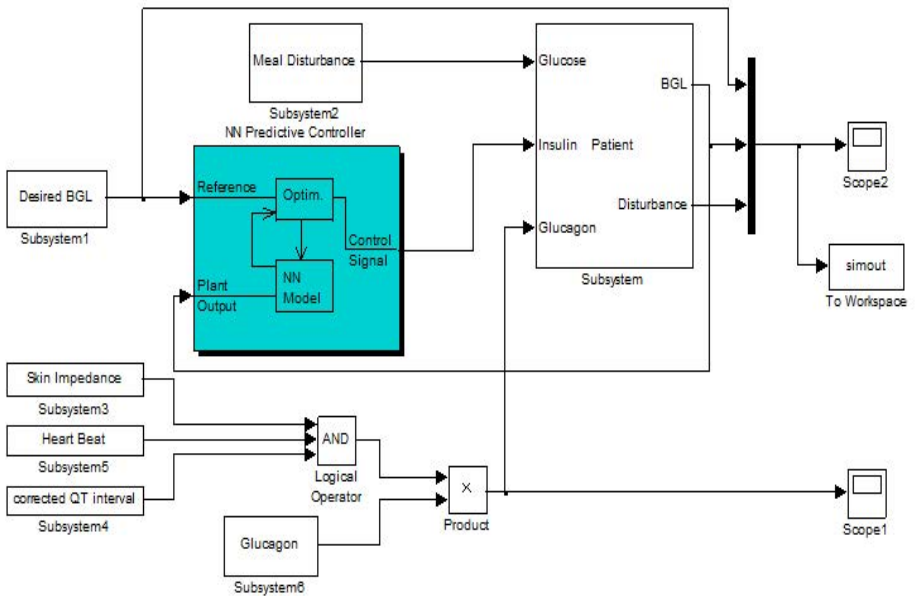


Fig. 5. Model of the Complete Insulin-Glucagon infusion system

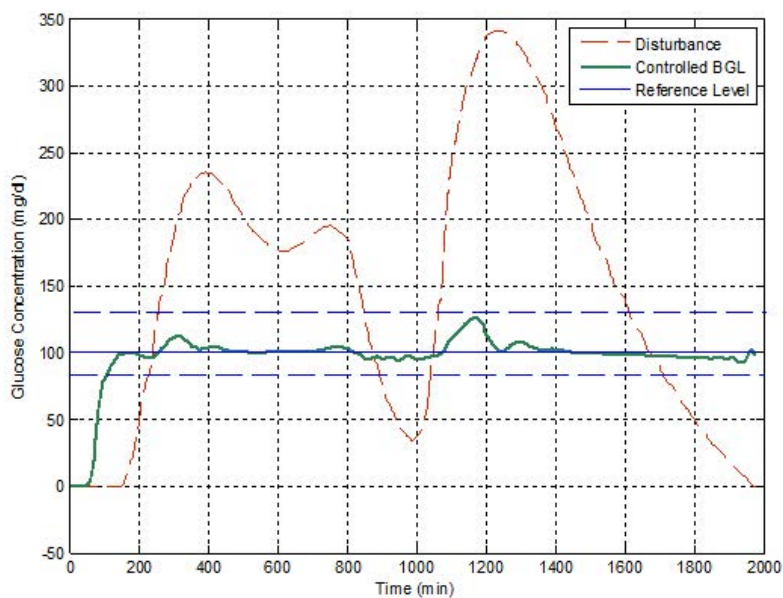


Fig. 6. System response to a huge disturbance

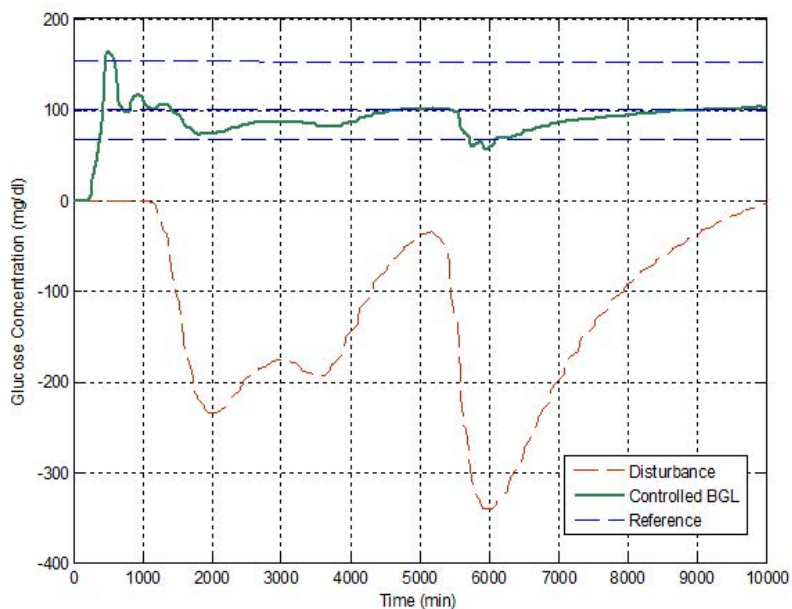


Fig. 7. A huge disturbance triggering hypoglycemia without glucagon infusion

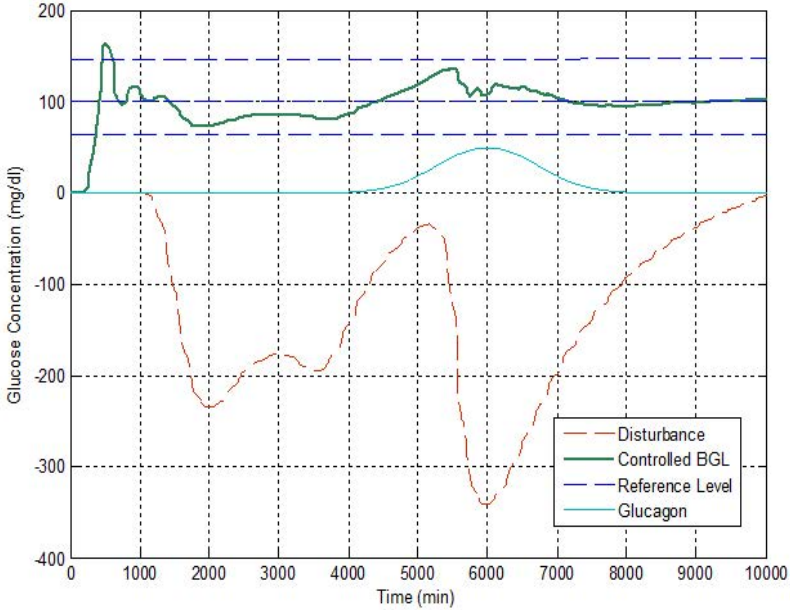


Fig. 8. With glucagon infusion hypoglycemia duration is avoided

In figs. [6](#), [7](#) and [8](#), it can be seen that the controller is significantly controlling the BGL. For the extreme disturbance value of nearly 350mg/dl the controller is maintain the BGL around 120mg/dl.

Glucagon can be administered intravenously or subcutaneously at 50 unit. An amount in this range is infused as soon as the hypoglycemia condition is detected. The effect of infusing glucagon is shown in fig. [8](#). The duration of hypoglycemia episode has been successfully avoided.

5 Conclusion

In view of the fact that not a single approved artificial pancreatic system is available, methods focusing patient safety need to be devised [\[32\]](#). We have proposed architecture based on vital signs monitoring to prevent the situation of hypoglycemia. The system is modeled in Matlab Simulink and results are encouraging. The study has shown that for an ambulatory system EEG monitoring is not practical. From the ECG waveform, RR and QT intervals along with skin impedance can depict the hypoglycemia condition. Automatic tuning of PID controller was found more suitable. Bergman model was suitably modified to incorporate glucagon infusion. In future, proposed architecture can also be tested using advanced model based or hybrid controller design techniques. Moreover, model checking tools like UPPAAL can be used in future to confirm the validity of our model, under strict timing constraints.

Acknowledgement. This research was supported in part by Higher Education Commission Pakistan grants National Research Program for Universities:1667 and 1668, King Abdul Aziz City for Science and Technology (KACST) grants: NPST-11-INF1688-10 and NPST-10-ELE1238-10 and National ICTRDF Pakistan grant SAHSE-11.

References

1. Lee, I., Sokolsky, O.: Medical Cyber Physical Systems. In: Proc. of DAC, Anaheim, California, USA (2010)
2. High Confidence Software and Systems Coordinating Group. High-confidence medical devices: Cyber-physical systems for 21st century health care. A Research and Development Needs Report, NCO/NITRD (February 2009)
3. Bergman, R.N., Phillips, L.S., Cobelli, C.: Physiologic evaluation of factors controlling glucose tolerance in man. Measurement of insulin sensitivity and β -cell glucose sensitivity from the response to intravenous glucose. *Journal of Clinical Investigation* 68(6), 1456–1467 (1981)
4. Dua, P., Doyle, F.J., Pistikopoulos: Model-Based Blood Glucose Control for Type 1 Diabetes via Parametric Programming. *IEEE Transactions on Biomedical Engineering* 53, 1478–1491 (2006)
5. Diabetes Control and Complications Research Group, The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N. Engl. J. Med.* 329(14), 977–986 (1993)
6. Banting, F.G., Best, C.H., Collip, J.B., Campbell, W.R., Fletcher, A.A.: Pancreatic extracts in the treatment of diabetes mellitus: preliminary report. *Cmaj* 145(10), 1281–1286 (1922) (reprinted 1991)
7. Cryer, P.E.: Hypoglycaemia: the limiting factor in the glycaemic management of Type I and Type II diabetes. *Diabetologia* 45(7), 937–948 (2002)
8. Hovorka, R., Chassin, L.J., Wilinska, M.E., Canonico, V., Akwi, J.A., Federici, M.O., Massi-Benedetti, M., Hutzli, I., Zaugg, C., Kaufmann, H., Both, M., Vering, T., Schaller, H.C., Schaupp, L., Bodenlenz, M., Pieber, T.R.: Closing the loop: the adicol experience. *Diabetes Technol. Ther.* 6(3), 307–318 (2004)
9. Farmer Jr., T.G., Edgar, T.F., Peppas, N.A.: The future of open- and closed-loop insulin delivery systems. *J. Pharm. Pharmacol.* 60(1), 1–13 (2008)
10. El Youssef, J., Castle, J., Kenneth Ward, W.: A Review of Closed-Loop Algorithms for Glycemic Control in the Treatment of Type 1 Diabetes. *Algorithms* 2, 518–532 (2009)
11. Weinzimer, S.A., Steil, G.M., Swan, K.L., Dziura, J., Kurtz, N., Tamborlane, W.V.: Fully automated closed-loop insulin delivery versus semiautomated hybrid control in pediatric patients with type 1 diabetes using an artificial pancreas. *Diabetes Care* 31(5), 934–939 (2008)
12. Elbein, S.C., Wegner, K., Kahn, S.E.: Reduced beta-cell compensation to the insulin resistance associated with obesity in members of caucasian familial type 2 diabetic kindreds. *Diabetes Care* 23(2), 221–227 (2000)
13. Bellazzi, R., Nucci, G., Cobelli, C.: The subcutaneous route to insulin-independent diabetes therapy. *IEEE Engineering in Medicine and Biology* 20, 54–64 (2001)
14. Gómez, E.J., Pérez, M.E.H., Vering, T.: The INCA System: A Further Step Towards a Telemedical Artificial Pancreas. *IEEE Transactions on Information Technology In Biomedicine* 12(4) (2008)

15. World Health Organization, Fact Sheet No. 138 (April 20, 2011) (Online), <http://www.who.int/mediacentre/factsheets/fs138/en/>
16. Carson, E.R., Cobelli, C. (eds.): *Modelling Methodology for Physiology and Medicine*. Academic Press, San Diego (2001)
17. Parker, R.S., Doyle III, F.J., Peppas, N.A.: The intravenous route to blood glucose control. *IEEE Eng. Med. Biol. Mag.* 20(1), 65–73 (2001)
18. Bellazzi, R., Nucci, G., Cobelli, C.: The subcutaneous route to insulin-dependent diabetes therapy. *IEEE Eng. Med. Biol. Mag.* 20(1), 54–64 (2001)
19. Hovorka, R.: Continuous glucose monitoring and closed-loop systems. *Diabet. Med.* 23, 1–12 (2005)
20. Bequette, B.W.: A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas. *Diabetes Technol. Ther.* 7(1), 28–47 (2005)
21. Owens, C., Zisser, H., Jovanovic, L., Srinivasan, B., Bonvin, D., Doyle III, J.: Run-to-run control of blood glucose concentrations for people with type 1 diabetes mellitus. *IEEE Trans. Biomed. Eng.* 53(6), 996–1005 (2006)
22. Dudde, R., Vering, T., Piechotta, G., Hintsche, R.: Computer-aided continuous drug infusion: setup and test of a mobile closed-loop system for the continuous automated infusion of insulin. *IEEE Trans. Inf. Technol. Biomed.* 10(2), 395–402 (2006)
23. Phee, H.K., Tung, W.L., Quek, C.: A personalised approach to insulin regulation using brain inspired neural semantic memory in biabetic glucose control. In: *IEEE CEC* (2007)
24. Hovorka, R.: The future of continuous glucose monitoring: closed-loop. *Current Diabetes Reviews* 4(3), 269–279 (2008)
25. Chee, E., Fernando, T., van Heerden, P.V.: Simulation study on automatic blood glucose control. In: *7th Australian and New Zealand Intelligent Information Systems Conference* (2001)
26. Demuth, H., Beale, M.: *Neural Network Toolbox for use with MATLAB*, Mathworks
27. Renard, E., Costalat, G., Chevassus, H., Bringer, J.: Artificial beta-cell: clinical experience toward an implantable closed-loop insulin delivery system. *Diabetes Metab.* 32(5 Pt 2), 497–502 (2006)
28. Chee, F., Savkin, A.V., Fernando, T.L., Nahavandi, S.: Optimal H_∞ insulin injection control for blood glucose regulation in diabetic patients. *IEEE Transactions on Biomedical Engineering* 52(10), 1625–1631 (2005)
29. Chen, J., Cao, K., Sun, Y., Xiao, Y., Su (Kevin), X.: Continuous Drug Infusion for Diabetes Therapy: A Closed-Loop Control System Design. *EURASIP Journal on Wireless Communications and Networking*, Article ID 495185 (2008)
30. El Jabali, A.K.: Neural network modeling and control of type 1 diabetes mellitus. *Bioprocess Biosyst. Eng.* 27, 75–79 (2005)
31. Matlab Simulink R2011a Documentation (Online), <http://www.mathworks.com/help/toolbox/slcontrol/ug/br684zf.html>
32. Klonoff, D.C.: The Artificial Pancreas: How Sweet Engineering Will Solve Bitter Problems. *Journal of Diabetes Science and Technology* 1(1) (2007)
33. Nguyen, H.T., Ghevondian, N., Jones, T.W.: Real-time Detection of Nocturnal Hypoglycemic Episodes using a Novel Non-invasive Hypoglycemia Monitor. In: *31st Annual International Conference of the IEEE EMBS* (2009)
34. Heller, S.R., Macdonald, I.A.: Physiological disturbances in hypoglycemia: effect on subjective awareness. *Clin. Sci.* 81, 1–9 (1991)

35. Gale, E.A.M., Bennett, T., MacDonald, I.A., Holst, J.J., Matthews, J.A.: The physiological effects of insulin-induced hypoglycemia in man: responses at differing levels of blood glucose. *Clin. Sciences* 65, 263–271 (1983)
36. Marques, J.L., et al.: Altered ventricular repolarisation during hypoglycaemic in patient with diabetes. *Diabetic Med.* 8, 648–654 (1997)
37. Koivikko, M.L., Salmela, P.I., et al.: Effects of Sustained Insulin-Induced Hypoglycemia on Cardiovascular Autonomic Regulation in Type 1 Diabetes. *Diabetes* 54, 745–750 (2005)
38. Castle, J.R., Engle, J.M., Youseff, J.E., et al.: Novel Use of Glucagon in a Closed-Loop System for Prevention of Hypoglycemia in Type 1 Diabetes. *Diabetes Care* 33(6), 1282–1287 (2010)
39. Heger, G., Howorka, K., Thoma, H., Tribl, G., Zeitlhofer, J.: Monitoring setup for selection of parameters for detection of hypoglycemia in diabetic patients. *Medical & Biological Engineering & Computing* (1996)

A Genetic Algorithm Assisted Resource Management Scheme for Reliable Multimedia Delivery over Cognitive Networks*

Salman Ali, Ali Munir, Saad Bin Qaisar, and Junaid Qadir

School of Electrical Engineering & Computer Science
National University of Science & Technology
Islamabad, Pakistan
salmanali@ieee.org,
{ali.munir, saad.qaisar, junaid.qadir}@seecs.edu.pk

Abstract. The growth of wireless multimedia applications has increased demand for efficient utilization of scarce spectrum resources which is being realized through technologies such as Dynamic Spectrum Access, source and channel coding, distributed streaming and multicast. Using a mix of DSA and channel coding, we propose an efficient power and channel allocation framework for cognitive radio network to place multimedia data of opportunistic Secondary Users over the unused parts of radio spectrum without interfering with licensed Primary Users. We model our method as an optimization problem which determines achievable physical transmission parameters and distributes available spectrum resources among competing secondary devices. We also consider noise contributions and channel capacity as design factors. We use Luby Transform codes for encoding multimedia traffic in order to reduce dependencies involved in distributing data over multiple channels, mitigate Primary User interference and compensate channel noise and distortion caused by sudden arrival of Primary devices. Tradeoffs between number of competing users, coding overhead, available spectrum resources and fairness in channel allocation have also been studied. We also analyze the effect of number of available channels and coding overhead on quality of media content. Simulation results of the proposed framework show improved gain in-terms of PSNR of multimedia content; hence better media quality achieved strengthens the efficacy of proposed model.

Keywords: Fountain Codes, Genetic Algorithm, Distributed Streaming, Secondary Users.

1 Introduction

Scarce communication spectrum poses significant challenges for reliable transmission of emerging bandwidth hungry multimedia applications. However, actual measurements

* This research was supported in part by Higher Education Commission Pakistan grants National Research Program for Universities:1667 and 1668, King Abdul Aziz City for Science and Technology (KACST) grants: NPST-11-INF1688-10 & NPST-10-ELE1238-10 and National ICTRDF Pakistan grant SAHSE-11.

taken in Berkley [9] show underutilization of wireless spectrum resources, especially over the TV band. To improve upon usage efficiency, Cognitive Radio (CR) framework was proposed as a result of Federal Communication Commission's (FCC) proposal of secondary spectrum utilization. A CR may utilize non-contiguous set of sub-channels upon availability, requiring multimedia content to be disseminated over a multiband spectrum. This diversity can cause significant loss in quality of multimedia content compared to good quality equal contiguous spectrum if it does not properly exploit the diverse characteristics of different bands. To mitigate such losses, a class of rate-less codes called Digital Fountain Codes (DFC), can be used allowing scalable media to be distributed in secondary channels. In codes, the encoder produces variable number of message packets and only a small portion of these is required for decoding at the receiver. The order of reception of these packets does not matter as long as receiver gets a desired number of packets that can correctly decode. This property of fountain codes can be utilized to combat loss due to PU interference [10] and other channel conditions like path loss, noise and fading etc. In literature rate-less codes have been used in cognitive network to assist message transmissions and to exploit quality links between source-destination. In this work, a more specific realization of the fountain codes, Luby Transform (LT) codes [1], has been utilized.

There are many trade-offs involved in selection of suitable channels and corresponding associated overhead for transmissions with rate-less codes. The addition of rate-less property further boosts the spectrum opportunities for cognitive networks. Multiple channels can be used to enhance signal-to-interference-plus noise ratios, balance traffic requests and to avail spectrum opportunities via cognitive space-time-frequency coding techniques. In this paper we address the task of power and channel allocation for cognitive radio nodes. Two types of nodes are considered namely source and destination nodes. The channel between the source and destination is a direct channel and no relay or similar terminal in between is available. In order to improve end-to-end throughput and optimize system performance we need to exploit available channels jointly while including power allocation in an optimization framework.

To elaborate our work, rest of the paper is arranged as follows. In the related work, we discuss the research work in similar domain. In the system modeling section, we highlight problem under discussion and based on our problem, appropriate system model is discussed. Resource allocation strategy is discussed in after the modeling section followed by results and discussion. Finally the paper is concluded.

2 Related Work

Various research efforts have focused quite recently on analyzing benefits of erasure codes in Cognitive environment. Primary work in this regard has been done by Harikeshwar Kushwaha et al. [5]. They discuss the impact of erasure coding for multimedia applications. A channel access mechanism has been introduced to distribute data in unused parts of the spectrum. In particular, authors focus on determining the required number of channels for successful delivery of a given number of coded packets.. In the work by T. Weiss et al [3], a method of channel availability and corresponding utilization has been proposed. The goal of the proposed

spectrum pooling mechanism is to improve spectral efficiency by overlaying a new terminal device system on an existing one in the relevant channel. The work by D. Cabric et al [4][9] looks into opportunistic utilization of spectrum by using techniques that are used to adapt to local spectrum availability. The approach termed as CORVUS performs channel allocation in a coordinated manner. However no attempt has been made to quantify or measure impacts of interference from PUs. Work done by J. Wagner et al [6] is an attempt towards quantifying the benefits of rate-less codes in streaming media applications. Similarly there are numerous other research findings that may have partial overlaps with this work. In contrast to the related work, we incorporate an integrated concept of quantifying interference and benefiting from erasure based rate-less codes while extending it to multiple nodes in a resource allocation problem. Recently, there has been some work on coding based transmission and erasure correcting capabilities in cognitive radios [16][17]. Finally there have been some significant contributions in coding based transmission with relay assistance for cognitive networks [18].

3 System Model

A. Assumptions and Goals

This work proposes a power and channel allocation framework for cognitive nodes in DSA environment. For this purpose we assume an overlay approach of DSA while LT codes have been utilized to mitigate effects of major channel losses. A spectrum pooling concept similar to the work by Weiss *et al* [3], is used to introduce secondary sub-channels while exploiting traffic modeling and relationships. We extend it to our multi-user environment of spectrum overlay for a variable number of fountain coded packets to achieve reliable content delivery. The channels between the source-destination pair are assumed to be block fading Rayleigh channels with flat frequency response. The block length of the channel is equal to the transmission time slots, hence the gain of the channel remains constant throughout the time slot and only changes from one slot to another. Simulations at the end provide a valuable insight into the trade-offs involved between varying numbers of users, coded packets and sub-channels.

B. Analytical Framework

We model our system as follows. Let the number of source nodes competing for channels be represented by $M = \{1,2,3, \dots\}$ where m would represent a node from the set M . The total number of available spectrum bands is $|\Psi|$, where channels in a band x from $|\Psi|$, are represented by ξ_x^c : $c = \{1,2,3, \dots\}$. where c is the set of sub-channels. There is an associated effective capacity with each channel denoted by C_x^c and a channel bandwidth W_x^c . Each channel has a specific quality level that is estimated over time and given by μ . Intuitively, this gives a percentage of interference from PU as seen by a SU. We call this interference as Primary User Susceptibility (PUS) of a channel.

Secondary nodes transmit DFC packets that are modeled as follows. Let there be an associated number of fountain coded packets with each node denoted by K^i where $i = \{1,2,3, \dots\}$. Each packet being generated has a size of s bits. Since each node can send different types of media content therefore it has a specific quality requirement or channel availability constraint denoted by δ_m where: $\delta_m : 0 \leq \delta_m \leq 1$.

C. Probability of Primary User

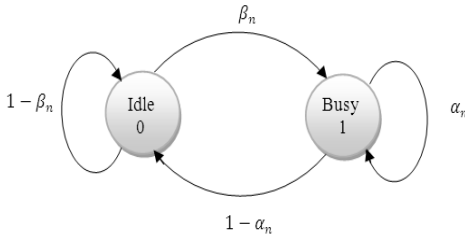


Fig. 1. Markov model for channel availability

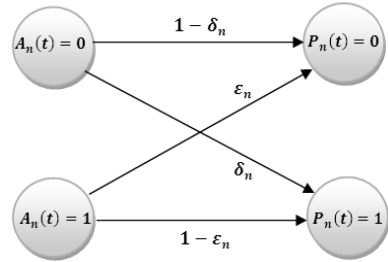


Fig. 2. Asymmetric channel for actual channel status and sensing results

Channel availability mechanism is modeled on the principles in [14]. We assume that the base station is capable of sensing the whole channel set of Primary networks, while the individual nodes are only capable of accessing and sensing a subset β of the whole spectrum, i.e. $\beta \subseteq |\Psi|$. Let the channel availability for the SU be modeled by a Markov process given in Fig 1. To cater for the errors in the sensing process, we can formulate two vectors, one representing the actual condition of the channel and the other representing the actual results of the sensing process. Assuming $A_n(t)$ denotes the spectrum sensing results achieved by a CR node at time t for channel n (taken from the set c defined before) and $P_n(t)$ be the actual channel state at time t . Since the channel can be in either busy or idle state, errors induced in the sensing process and resulting false values of $A_n(t)$ can be represented by asymmetrical channel as shown in Fig 2.

The channel sensing vector is given by $\vec{A}_n(t) = [A_0(t), A_1(t), \dots, A_{n-1}(t)]$ and the actual channel condition vector is given by $\vec{P}_n(t) = [P_0(t), P_1(t), \dots, P_{n-1}(t)]$. As a result, the Primary User susceptibility vector would be given by $\vec{\mu}_n(t) = [\mu_0(t), \mu_1(t), \dots, \mu_{n-1}(t)]$. Using Bayes law and conditional probability we can compute $\mu_n(t)$, the probability of channel being in busy state. Given some previous history of the channel occupation, ω_i , we write initial probability as

$$\mu_n(t) = \alpha_n \mu_n(t - 1) + \beta_n (1 - \mu_n(t - 1)) = \pi_n(t) \tag{1}$$

The history ω_i can be manipulated as:

1. If an acknowledgement is received in time t , then in interval $t - 1$, the channel can be assumed busy in transmission.

2. If there is a successful transmission negotiation in the current slot then in the previous slot, the channel will be busy in carrying requests.
3. If data is received in time interval t then in interval $t - 1$, the channel was busy in transmission.

Incorporating and conditioning on channel sensing result A_n , we can write probability for correctly deciding upon the busy channel statistics as:

$$\mu_n(t) = \frac{\pi_n(t)(1 - \epsilon_n)}{\pi_n(t)(1 - \epsilon_n) + [1 - \pi_n(t)]\delta_n} \tag{2}$$

where $\pi_n(t) = \alpha_n\mu_n(t - 1) + \beta_n(1 - \mu_n(t - 1))$ is based on stationary memory-less Markov model assumption. Let the transmission probability at time instant t be given by $p_n^{tr}(t)$. Then for unsuccessful transmission on channel n , we can write the probability of transmission failure or error as

$$p_{fail} = \mu_n(t) \times p_n^{tr}(t) \tag{3}$$

The probability of transmission can be defined in terms of collision range under which we allow a node m to transmit. Let the range be termed as θ_m , where $0 < \theta_m < 1$ ranges between 0 and 1. As we want the transmission failure probability range to be as limited as it can be, we set $\mu_n(t) \times p_n^{tr}(t) \leq \theta_m$. Hence, the node transmits deterministically or opportunistically as

$$p_n^{tr}(t) = \min\left(1, \frac{\theta_m}{\mu_n(t)}\right) \tag{4}$$

The channel gains are exponentially distributed and independent of each other. Let g_i be the channel gain from a source to destination on the i -th channel for direct communication between source and destination. The channel gain mean is represented by \bar{g} . An SINR threshold is assumed at each receiver represented by δ . A coded packet will be dropped as an erasure in the current time slot if the receiver gets the coded packet with less than the threshold value and a successful reception will occur otherwise. The erasure probability is therefore represented for source-destination communication as

$$p_{ei}^{sd} = \Pr\left(\frac{P_s g_i}{n_r^i} < \delta\right) = 1 - \exp\left(-\frac{n_r^i \delta}{P_s \bar{g}_i}\right) \tag{5}$$

where n_r^i is the noise power at the corresponding receiver for the source that transmit with power P_s . A message is completely recovered when the receiver obtains any of the $H = (1 + \epsilon)K^i$ coded packets at least where ϵ is the overhead for decoding.

4 Resource Model

A. Transmit Power Constraint

Let P_s^j be the transmit power of the source on the j -th spectrum band Then the power allocation vector can be represented by $P_s = [P_s^1, P_s^2, \dots, P_s^j]$. In particular, a zero value for an entity in the allocation vector implies that the band cannot be used at the source.

To avoid harmful interference to the PU, we introduce a power limitation for the source on each band denoted by P_{max} which can be regarded as the per band power constraint. Hence we have the following power constraint $P_s^j \leq P_{max}$, $j = 1,2,3, \dots$. The exact value of the P_{max} limit will depend upon the CSI mechanism involved and the physical spectrum sensing mechanism involved. However for simplicity, the sum power constraints at the source derived from the per power constraints is expressed as

$$\sum_{\forall s} P_s^j \leq P_{max}^s, \quad j = 1,2,3, \dots$$

where P_{max}^s is the aggregated maximum power that the source is able to transmit.

B. End-to-End Throughput

Since we can use multiple channels for end-to-end parallel transmission in direct communication mode, the capacity of the channel would vary according to the combination used. If we represent capacity of the channel by Shannon theory, we can write

$$C(j) = B \log(1 + j) \tag{6}$$

Where $C(j)$ is the Shannon capacity for channel j with Bandwidth B . But the actual throughput would be based upon the power allocations and the channel gains respectively. Hence we can write for the channel combination, the end-to-end throughput equations as:

$$T_{direct} = C(P_s^j d_i) \tag{7}$$

where T_{direct} is the end-to-end throughput for the channel utilization scheme with focus on half-duplex scenario. The average flow requirement of data is determined by the nodes themselves and communicated to the central entity before channel assignment phase. Here, we develop a framework in the form of an optimization problem to determine the node’s physical transmission capability parameters that would be utilized for data transmission. For simplicity we focus on three parameters only, the modulation scheme, transmission power and bit error rate. The main idea here is that the basic infrastructure upon which CRs are built is the software defined radio where majority of the communication system blocks like modulation, encoding, encryption etc. are implemented in software for ease of adaptation to underlying channel conditions. The generalized fitness function for cognitive radio nodes can be defined as

$$\text{maximize } f(x) = \sum_{x = \text{Power, Throughput}} w \times f(x)$$

$$\text{subject to } g(x) \leq \text{Constraint}$$

where we need to maximize Power and Throughput fitness functions under some constraint function $g(x)$ using weights w . Let $Mod = \{1,2, \dots, m_{mod}\}$ be the modulation schemes available at a particular node in increasing order of constellation size. For example, scheme 1 can represent BPSK and the highest available be 128-QAM (Table 1). Also let the power with which the node can transmit, be in constant intervals of Δp , and represented by the set $P = \{1,2,3, \dots, p\}$.

Table 1. Physical Parameters considered in the work

Parameter	Schemes Used
Modulation	BPSK, 8-PSK, 16,64 and 128-QAM
Power Level	0mW-50mW (0-27dBm)

For example maximum transmission power that a node can support is 50mW and $\Delta p = 5mW$, then we would have 10 power levels from 0W to 5mW. With the increase in power levels and modulation scheme to choose from, the search space for an appropriate combination to use also increases.

In CRs it is desirable to communicate with low power transmissions when occupying PU channels, thus we need to incorporate tradeoffs involved in the selection of power levels and modulation schemes. A larger power level can not only increase the signal-to-noise ratio but also cause interference with other transmissions. A higher modulation scheme results in a higher bit rate but we are forced to switch to lower schemes when a higher constellation scheme tends to produce higher bit-error rates due to degrading channel conditions. Hence on one side we want to keep the power levels as low as possible and at the other end we want to use the highest modulation scheme to achieve high throughputs.

C. Fitness Functions

In this section, we define a power minimization function and a throughput maximization function. These functions are then combined in a filtering equation according to the environmental characteristics that includes bit error rate achieved on the channel. The power minimization problem can be represented as:

$$f_{\min_pow} = 1 - \frac{\sum_{i=1}^{N-1} P_i}{N \times P_{max}} \tag{8}$$

where N represents the maximum number of sub-carriers over which transmission can take place. The fitness function for maximizing the throughput in terms of the modulation scheme can be represented as;

$$f_{\max_thr} = \begin{cases} 0 & \text{if } \sum_{i=1}^{N-1} Mod_i < thr_{lower} \\ \frac{thr_{upper} - \frac{\log_2(Mod_i)}{\log_2(Mod_{max})}}{thr_{upper} - thr_{lower}} & \text{if } thr_{upper} < \sum_{i=1}^{N-1} Mod_i < thr_{lower} \\ 1 & \text{if } \sum_{i=1}^{N-1} Mod_i > thr_{upper} \end{cases} \tag{9}$$

Where $[thr_{upper}, thr_{lower}]$ represent the interval limits which effect user utility the most. Both f_{\min_pow} and f_{\max_thr} result in values ranging from 0 to 1. P_i and Mod_i are current candidate power and modulation parameters and P_{max} and M_{max} are the maximum available parameters respectively. To maximize the throughput we overall need to assign power with sum and per channel power constraint.

$$F_{SD} = \max_{PS} \left\{ \sum_{i \in S} C(p_i^s g_i) \right\} \tag{10}$$

Subject to

$$\sum_i p_i^s \leq p_{max}^s$$

$$p_i^s \geq 0$$

D. Multiple Objective Fitness Function

To select the most appropriate combination to be utilized would then be determined by a multi-objective fitness function given as;

$$f_{multi_obj} = w_1 \times (f_{min_pow}) + w_2 \times (f_{max_thr}) \tag{11}$$

where the weights of w_1 and w_2 would determine the direction of search for better combination. The weights are indirectly derived from the channel characteristics depicted by the probability of bit error rate achieved by a specific coding scheme, i.e. $w_2 \propto (1 - P_{ber})$. Hence for higher P_{ber} we need to shift at a lower modulation scheme. Since we stick to the use of normalized weights, weightage for the single objective function to achieve minimization in power would be calculated as; $w_1 = w_2 - 1$. For a given combination of power and modulation scheme, the probability of bit-error-rates are defined for specific schemes, e.g. for BPSK, M-ary PSK and M-ary QAM, the equations are;

$$P_{ber_BPSK} = Q\left(\sqrt{\frac{P}{N}}\right)$$

$$P_{ber_MPSK} = \frac{2}{\log_2(M)} Q\left(\sqrt{2 \times \log_2(M) \times \frac{p}{N} \times \sin\left(\frac{\pi}{M}\right)}\right)$$

$$P_{ber_MQAM} = \frac{4}{\log_2 M} \left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\frac{3 \times \log_2 M}{M - 1} \times \frac{P}{N}\right)$$

Finally we can determine the average flow demand over some time period $[0 - T]$ as a function of the modulation index that achieves a specific throughput.

Table 2. The Dynamic Fixing Algorithm

1. Initialize w_2 with an initial random value and $w_1 = 1 - w_2$
2. Set initial P_{ber_curr} to the value achieved by the current MOD scheme
3. Initialize timer $t = 0$ and Δ interval
4. Set an initial allowable threshold change of P_{ber} as σ
5. After $t = t + \zeta$, check P_{ber}
6. If $ P_{ber_curr} - P_{ber} \geq \sigma$, decrease ζ
7. Else if $ P_{ber_curr} - P_{ber} < 0.1\sigma$ increase ζ
8. End if
9. Set w_2 as $(1 - P_{ber})$ and $w_1 = w_2 - 1$

E. Weights Adaptation

The weights defined for the filtering equations need to be modified according to the varying channel conditions. Thus, we formulate a simple algorithm for weights adaptability, using dynamic programming. The algorithm starts with an initial randomized value of w_2 . $P_{ber,curr}$ is set to the value achieved by the current scheme on the channel. Channel is checked after every Δ interval for the bit errors achieved. A bit error value lower than the threshold would initiate an increase in Δ interval, and decrease otherwise. For reduced oscillations, the interval should be kept large and initiated in emergency only if the communication statistics reach a poor level. The average flow demands are conveyed to the receiver only once at the start and later the weights adaptation leading to different throughputs are done on the same channel.

The values of the time steps and weights depend on the terminal systems processing capability. A high performance system capable of sensing the environment fast for achieved BER can set the time step low without degrading other transmission activities and vice versa. To determine the exact values, dynamic programming can be used with ease. The program begins with an initial very high value of $w_2^{(0)}$ and a very small value of time step ζ . From that point, next allowable time steps that the system can handle are calculated with a decrease in $w_2^{(0)}$ and an increase in time step. The minimum ratio $\lambda_i^{(j)}$ is calculated for the step j for all the points given as

$$\min_i \lambda_i^{(j)} = \left| \frac{\Delta \zeta_i^{(j)}}{\Delta w_2^{(j)}} \right| \tag{12}$$

Where

$$w_2^{(j)} = w_2 - |\Delta w_2^{(j-1)}| \quad \text{and}$$

$$\zeta_i^{(j)} = \zeta_i + |\Delta \zeta_i^{(j-1)}|$$

The optimal operation points are reached when $w_2^{(j)} \leq P_{BER}$. The sub optimal algorithm for determining weights is given in Table 2.

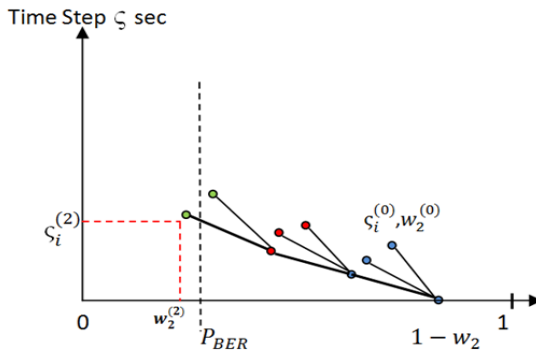


Fig. 3. Dynamic Programming for determining weights and time steps

5 Resource Allocation

Since we have multiple channels with which there are different associated capacities, we would have variable end-to-end throughput for different channel allocation strategies. The channel allocation method corresponds to the selection of the best mix that will maximize the overall throughput. The throughput would also depend upon the power allocation. To achieve the overall maximum throughput, the power allocation also needs to be maximized. The maximum throughput channel selection is to choose the best transmission modes from the channel mix. The parameters that we need to take decision on are the Secondary sub-channels and the nodes to which these channels would be assigned. Thus, we have to take a decision on the parameters N and Ψ in our genetic algorithm formulation. An assignment matrix $A(n \times k)$ would be utilized for the channels allocation at the end.

A. Fitness Objectives

Since it is possible that a channel with a larger capacity can have susceptibility to Primary User or increased interference levels, therefore it presents a tradeoff between channel bandwidth utilization and the channel availability demand by the transmitting node. Assigning a channel with a larger capacity to a node with few coded packets to send will be an underutilization of resources, where as assignment of low capacity channels to nodes with high data rates requirement would serve no purpose. Also assignment of a sub-channel with high interference levels to nodes with higher channel availability or quality requirements would kill the purpose. Hence, we develop a scheme that gives appropriate importance to each requirement in the form of weightings. Our algorithm serves to assign channels catering for the amount of bandwidth being utilized and the channel quality being offered to a node with a specific channel availability requirement. We define three objectives that the algorithm would focus on: 1) Achieve maximum bandwidth utilization in the form of throughput. 2) Provide sub-channels with better quality to nodes with higher requirement of quality or channel availability. 3) Assign channels that result in minimization of interference.

B. Fitness Functions

According to the objectives defined above we develop fitness functions that are used to compute the performance of the algorithm. Given the requirement of maximum bandwidth utilization and throughput, for a node i

$$f_{\max_BW} = \frac{\log_{10}(C_x^c)}{\log_{10}(K^i \times w_k)} \quad (13)$$

Also to cater for the quality requirements and service layer availability, we can formulate fitness function that minimizes the interference

$$f_{\min_inter} = \begin{cases} 1 & \text{if } \sum_{i=1}^{N-1} I_i < \text{inter}_{\text{lower}} \\ \frac{\text{inter}_{\text{upper}} - \log_{10} \left(\frac{(\mu_x^c)_{\text{max}}}{\mu_i^c} \right)}{\text{inter}_{\text{upper}} - \text{inter}_{\text{lower}}} & \text{if } \text{inter}_{\text{upper}} < \sum_{i=1}^{N-1} I_i < \text{inter}_{\text{lower}} \\ 0 & \text{if } \sum_{i=1}^{N-1} I_i > \text{inter}_{\text{upper}} \end{cases} \quad (14)$$

where $I = \sum_{i=1, i \neq j}^m \frac{I_j}{P_i}$ represents the interference level caused by radio j to all other radios from 1 to m excluding j . f_{\max_BW} and f_{\min_inter} result in values in the range $[0,1]$ where a higher value determines the suitability of channel allocation.

C. Multiple Objective Fitness Function

To be able to achieve the appropriate combination of sub-channels according to the nodes requirement we formulate a multi-objective equation that incorporates the single objective fitness functions in equation (13) and (14). This would give us a score between 0 and 1 which is the overall fitness score of the algorithm. The minimization and maximization equations are multiplied with a weighting factor;

$$f_{\text{multi_obj}} = w_1 \times (f_{\min_inter}) + w_2 \times (f_{\max_BW}) \quad (15)$$

Where w_1 is the percentage of channel availability requirements conveyed by a specific node and varies as $0 \leq w_1 \leq 1$. We stick to the use of normalized weights, such that: $1 = w_1 + w_2$. Hence the above multi objective formulation will work as a filtering equation and give results according to the weights giving importance of each objective in both the equations. Again the objective here is to rank channels according to the values of $f_{\text{multi_obj}}$ and then do channel assignment. A channel with a higher value of $f_{\text{multi_obj}}$ is more suitable for communication.

D. Fairness in Channel Assignment

For maintaining proportionality fairness while the channel assignment is done we constantly evaluate the ratio of total percentage of demand fulfilled and the average susceptibility of the assigned channels.

$$R_{\text{fair}} = \frac{\sum_m C_{x(m)}^c}{K^i \times w_k} / \text{avg}(\mu_i^c) \quad (16)$$

where R_{fair} value for the candidate nodes will be used as a measure to determine which nodes to test for channel assignment. A node with a comparative smaller value would depict the node likely to be tested for channel assignment. The reason we use proportionality fairness criterion is to prevent certain nodes with higher demand from starving other nodes with less requirements while manifesting itself as a simplistic approach towards fairness.

Table 4. Resource Allocation Algorithm

1.	Input: $N = (K^i, w_k, (\mu_x^c, C_x^c))$
2.	For $M = 1: m$
3.	Select node = $arg_{max}(\delta_m)$
4.	For $P_{pop} = 1: 1: pop$
5.	Select node with minimum of R_{fair}
6.	Generate population of sub-channels
7.	Evaluate $f_{max_BW}, f_{min_inter}$ and f_{multi_obj}
8.	Arrange in descending order the results from equation (15)
9.	Perform crossover and mutation
10.	End For
11.	Evaluate the first chromosome
12.	Evaluate $K^i \times w_k - C_x^c$
13.	Update R_{fair}
14.	Assign current channel to node in matrix $A(n \times k)$ and move to step (5)
15.	If criteria fulfilled, remove current node from list N and move to (3)
16.	End For
17.	Output: $A(m \times n)$ The channel assignment matrix

E. Algorithm

The channel assignment algorithm, discussed in Table 4, takes channel parameters as input and outputs a channel assignment matrix A . The multi objective formulation is constantly evaluated during the channel assignment phase, while the proportionality fairness ratio prevents any individual node from capturing the resources. After the end of the optimization algorithm, the results are arranged with decreasing order for fitness values representing the suitability of assignment to radio terminals. The highest of these is allocated and the resource vector is updated. To avoid resource scarcity not all the resources are allocated if the demand is low. This is important because channel requirements from transmitting terminals may turn up at any time.

6 Simulation Results

A. Genetic Optimization

We evaluate the convergence of throughput demand calculation from multi-objective fitness function. For a specific case, the channel assignment algorithm was evaluated for 35 channels and 3 source-destination pair nodes. The channel availability was kept random for the three nodes while the channels were also randomly selected with different capacities. The multi-objective function converges to 0.9 in about 10 iterations from an initial score of around 0.5. For normally around 50 channels, a fitness score of 0.9 can be obtained in an approximate of 10 iterations.

We compared our method to Equal Allocation to visualize the efficiency of the fairness criterion. Figure 6 depicts that when the demand vector is same for all resource competing terminals, our approach works just like an equal channel

allocation methodology to distribute resources. However when the demand vector is randomized with different values, the fairness ratio plays a more visible role and raises the average demand fulfillment to more than double the equal allocation method as depicted in Figure 5. The variance ranges specified with ‘I’. The model is evaluated for 15 to 50 channels availability.

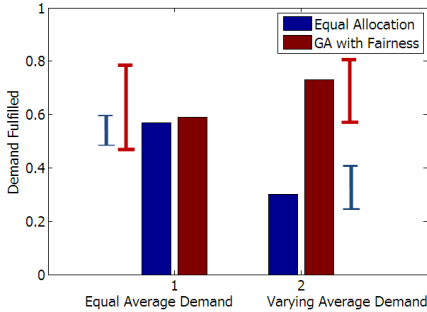


Fig. 4. GA with Fairness Compared to Equal Allocation

Table 5. Variance in convergence score relative to number of channels

Parameter	Value
Channels	15-50
Iterations	5-15
Convergence Score	>0.6

B. Fountain Codes Efficiency

Multimedia transmission over such opportunistic networks requires some kind of protection against losses for efficient streaming. We make use of fountain codes for efficient and reliable delivery of data across multiple channels. LT codes, due to their inherent property, help in eliminating the need for coordination among data packets of a single flow. This is due to the reason that in LT codes, the order in which packets arrive is not important, all we need is sufficient amount of packets to successfully decode data. A decoder with decoding capability x require, $k(1 + x)$ packets to decode a stream of k packets. For any such channel with erasure rate $q, k(1 + x)/q$, packets are required for successful delivery of data [1].

We measure performance of the proposed framework using MATLAB and ns-2 simulations. To capture effect of Sudden arrival of Primary Users, Figure 5 shows results for the packet loss probability depending on the channel losses introduced by Primary Users. Here the number of packets is assumed to be 1000 bytes and overhead ranges from 10 to 50 % in steps of 10. We can see that as the overhead increases the fraction of lost packets drop significantly and we have better video quality. After an overhead of 50% our analysis show that this performance starts degrading due to too much redundant information. We also see a trend that as the average erasure rate of the sub-channel increases, packet loss rate also increases and LT codes are not able to recover data completely.

For Cognitive networks scenario tradeoff between packet loss rate, overhead and number of Secondary channels is of significant importance. Figure 7 depicts this case and shows the effect of overhead on packet loss rate with respect to the amount of overhead and number of sub-channels. We can observe that as the number of sub-channels increase, the fraction of lost packet decreases and we can see that after 30% overhead we can decode file completely at the receiver, provided we have more than five sub-channels

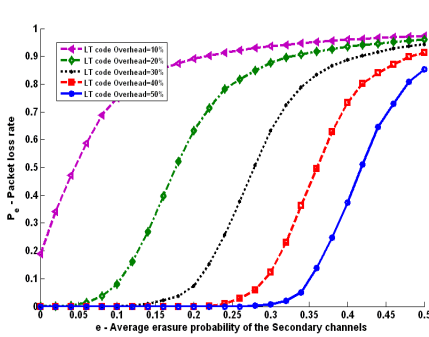


Fig. 5. Packet loss rate vs. channel erasure rate

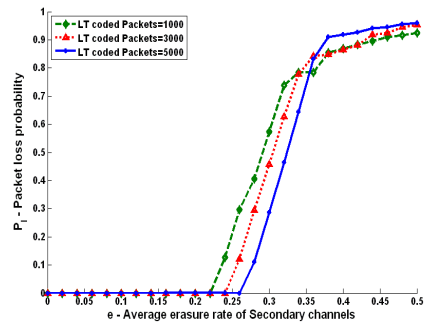


Fig. 6. Effect of Packet size on Packet loss (overhead 30% for fair comparison)

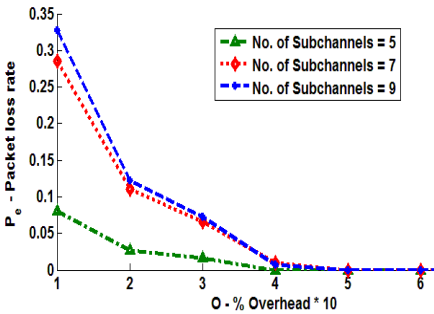


Fig. 7. Effect of Overhead on Packet erasure

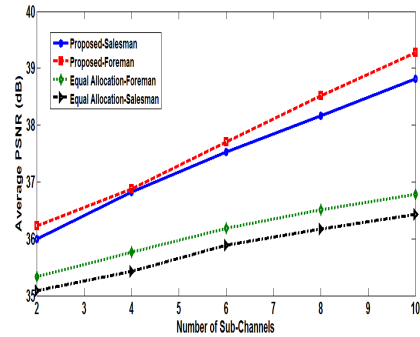


Fig. 8. Average PSNR of all users versus channels N

available. The erasure probability/ Primary User susceptibility of sub-channels is assumed as: $E = [0.03 \ 0.04 \ 0.02 \ 0.01 \ 0.025 \ 0.06 \ 0.01 \ 0.03 \ 0.015]$

The performance of network is dependent on the erasure rate or primary user arrival pattern thus the behavior may vary depending on the primary user arrival pattern.

In Figure 6 the packet loss with respect to variation in file size is shown. We can see that for large files the packet loss rate increases after a certain threshold of erasure probability. For example a video with 5000 packets have less PER before 35 % overhead, then its PER increases as the total number of lost packet increases due to too much redundant data.

Effect of number of channels on video quality can be observed from PSNR metric. We use foreman and salesman sequences to analyze performance of our proposed framework. Thus, Figure 8 shows the impact of number of channels N on the quality of video. We increase N from 2 to 10, in step size of 2. The average PSNR values of all users show that as the number of channels increases, the spectrum usage increases in CR networks. From figure we can observe that foreman sequence shows better performance as compared to salesman overall. However the average PSNR thresholds achieved are a depiction of the framework efficacy. We compare our proposed methodology to equal

allocation method. We can see that proposed framework achieves significant gains over equal allocation method and exploits available resources in a better way. The results have been shown with 95% confidence intervals.

The performance can be best understood by per-frame PSNR values. Figure 9 represents the variation in PSNR values on per frame basis. We have used salesman and foreman sequences with 300 frames. We observe that variation in salesman is more than that of foreman. The reason may be that the variation in frame sequence of salesman is more than that of foreman sequence. The results were produced for a network with three sub-channels per user and Primary User susceptibility 0.25.

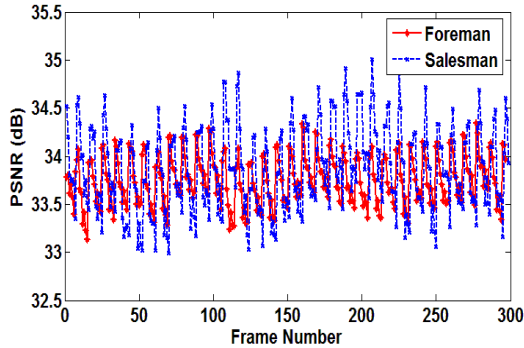


Fig. 9. Per-frame PSNR

7 Conclusion

In this paper we propose a resource allocation and efficient utilization scheme to place the distributed multimedia data of Secondary Users over the unused parts of the radio spectrum without interfering with licensed Primary Users of a Cognitive network. Genetic Algorithm has been used as an optimization scheme to distribute the available spectrum resources among the competing Secondary Users over Cognitive radio networks with central decision making entity. When compared with Equal Channel Allocation method, our framework performs with close results when given equal demand vector. The framework shows enhanced performance when the demand vector is random as in real network scenarios. This is mainly due to the fact that fairness criterion adopted into the resource allocation algorithm is critical for very theme of the framework. Fountain codes were proposed and used for encoding the Secondary User's data in order to compensate for the loss due to Primary User interference and channel noise. A method to determine the average flow demand has also been discussed. Tradeoffs have been studied between the number of competing users, the coding overhead, number of available spectrum resources and their bandwidth. The simulation results prove the effectiveness of the algorithm and use of fountain codes promises an intelligent routine for the Cognitive radio environment.

Acknowledgments. This work is supported by Myongji College with Prof. Dong-Young Lee as corresponding author.

References

- [1] Luby, M.: LT Codes. In: The 43rd Annual IEEE Symposium on Foundations of Computer Science (2002)
- [2] Goldberg, D.E.: Genetic Algorithms in Search Optimization and Machine Learning, p. 41. Addison Wesley (1989)
- [3] Weiss, T., Jondral, F.: Spectrum pooling: An innovative Strategy for the enhancement of spectrum efficiency. *IEEE Communication Magazine* 42, S8–S14 (2004)
- [4] Cabric, D., Mishra, S.M., Willkomm, D., Broderson, R.W., Wolisz, A.: A Cognitive Radio Approach for Usage of Virtual Unlicensed Spectrum. In: Proc. 14th 1st Mobile Wireless Communication Summit, Dresden, Germany (June 2005)
- [5] Kushwaha, H., Chandramouli, R.: Secondary Spectrum Access with LT Codes for Delay-Constrained Applications. In: 4th IEEE Consumer Communications and Networking Conference, CCNC 2007, Las Vegas, NV, USA, pp. 1017–1021 (2007)
- [6] Wagner, J., Chakareski, J., Frossard, P.: Streaming of scalable video from multiple servers using rateless codes. In: Proc. IEEE International Conference on Multimedia Expo. 2006 (July 2006)
- [7] Bck, T., Schwefel, H.: Evolutionary computation: An overview. In: International Conference on Evolutionary Computation (1996)
- [8] Rondeau, T., Le, B., Rieser, C., Bostian, C.: Cognitive radios with genetic algorithms: Intelligent control of software defined radios. In: Software Defined Radio Forum Technical Conference (2004)
- [9] Broderson, R.W., Wolisz, A., Cabric, D., Mishra, S.M., Willkomm, D.: CORVUS: A Cognitive Radio Approach for Usage of Virtual Unlicensed Spectrum, University of California, Berkeley, Tech. Rep. (2004)
- [10] Mahmoud, Q.H.: Erasure Tolerant Coding for Cognitive Radios. In: Kushwaha, H., Xing, Y., Chandramouli, R., Subbalakshmi, K.P. (eds.) *Cognitive Networks: Towards Self-Aware Networks*, July 25, ch. 13, pp. 315–331 (2007)
- [11] Newman, T.R., Barker, B.A., Wyglinski, A.M., Agah, A., Evans, J.B., Minden, G.J.: Cognitive engine implementation for wireless multicarrier transceivers. *Wireless Commun. Mobile Comput.* 7(9), 1129–1142 (2007)
- [12] Hauris, J.F.: Genetic algorithm optimization in a Cognitive radio for autonomous vehicle communications. In: Proc. Int. Symp. CIRA, Jacksonville, FL, June 20–23, pp. 427–431 (2007)
- [13] Thilakawardana, D., Moessner, K.: A genetic approach to cell-by-cell dynamic spectrum allocation for optimising spectral efficiency in wireless mobile systems. In: Proc. 2nd Int. Conf. CrownCom, Orlando, FL, August 1–3, pp. 367–372 (2007)
- [14] Hu, D., Mao, S., Reed, J.H.: On Video Multicast in Cognitive Radio Networks. In: *INFOCOM 2009*, pp. 2222–2230 (2009)
- [15] Mitola, J., Maguire, G.Q.: Cognitive radio: Making software radios more personal. *IEEE Communication* 6, 13–18 (1999)
- [16] Asareh, A.: A novel reliable broadcasting scheme under cognitive radio environment based on erasure correctable codes'. In: International Conference on Computing, Networking and Communications (ICNC), pp. 257–261 (2012)
- [17] Asareh, A.: Performance evaluation of coding-based cognitive radio for various packet sizes. In: *IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops* (September 2010)
- [18] Wang, X., Chen, W., Cao, Z.: ARCOR: Agile Rateless Coded Relaying for Cognitive Radios. *IEEE Transactions on Vehicular Technology* 60(6) (July 2011)

Performance Analysis of WiMAX Best Effort and rtPS Service Classes for Video Transmission

Hassan Abid¹, Haroon Raja¹, Ali Munir¹, Jaweria Amjad¹, Aliya Mazhar¹,
and Dong-Young Lee^{2,*}

¹National University of Science & Technology

²Myongji College

Abstract. To support different types of data like http, real-time audio and video, VoIP, FTP, there are various classes in WiMax system. In this work, we try to analyze the performance when multimedia contents are transmitted over WiMax network. Due to stringent delay requirement of real-time multimedia data, a separate class is allocated for it. i.e. rtPS. Thus our objective is to find out that how much we gain advantage by transmitting multimedia over this separate class? This requires a thorough analysis while considering all the scenarios. Our contribution in this paper is to build an initial framework for answering the above stated questions. The Network Simulator (ns-2) which is a popular tool for the simulation of computer networks has been used to simulate the results. Standard-compliant implementations have been used to authenticate the results.

Index Terms: WiMAX, Best Effort, Real Time Polling Service.

1 Introduction

Over the past last decade, there has been a major boom in the field of communication networks. A lot of investment has been made in this field and still there is room for improvement. A lot of deployment in terms of infrastructure, development of high performance backbone networks and related fields has enhanced the overall picture of communication. A lot of growth in terms of new services, access mechanisms, quality, etc. has been witnessed for commercial as well as residential users. As the advancements moves on and high speed fiber backbone links are made available, the access to broadband services is made available to the end users. This brings in new service classes that require significant network resources. Broadband Access opened a lot of opportunities for end users based on the availability to a lot of multimedia applications like VoIP, VoD, Video Conferencing, Massively Multiplayer Online Gaming (MMOG), etc. Broadband Wireless Access (BWA) is considered to be a

* This research was supported in part by Higher Education Commission Pakistan grants National Research Program for Universities:1667 and 1668, King Abdul Aziz City for Science and Technology (KACST) grants: NPST-11-INF1688-10 & NPST-10-ELE1238-10 and National ICTRDF Pakistan grant SAHSE-11.

promising solution for broadband access due to many factors like fast deployment resulting in cost savings and the ability to reach very crowded or rural areas, etc. IEEE 802.16 working group is working on the standards for Broadband Wireless Access. Worldwide Interoperability for Microwave Access (WiMax) forum was also founded to promote the technologies based on 802.16. The main objectives include high speed Internet access, cellular backhaul, Wi-Fi hotspot backhaul and services to the private networks. WiMax has evolved from IEEE 802.16 to IEEE 802.16d for fixed wireless access in 2004 and in 2005, to IEEE 802.16e standard, including mobility support.

Multimedia applications/ services over broadband wireless access are the main area of research nowadays. As these applications usually require stringent network guarantees such as reserved bandwidth or bounded delays, so providing Quality of Service (QoS) simultaneously to the services having different requirements and characteristics is the main challenge for Broadband Wireless Access networks these days. As Medium Access Control (MAC) layer is mostly responsible for QoS in wireless networks, so in WiMax, QoS is provided using classification and scheduling of four different types of traffic classes to support different types of data flows. These classes are: **Unsolicited Grant Service (UGS)** designed for real time traffic with constant bit rate (CBR) like VoIP; the bandwidth is assigned once in the start of the transmission and provides fixed size transmission opportunities at regular time interval. **Real-time polling service (rtPS)** designed for variable bit rate (VBR) traffic support like MPEG video. The base station offers periodic request opportunities to the SS. **Non-real-time polling service (nrtPS)** is designed for delay tolerant data services with minimum data rate like FTP. Contention and unicast requests are sent by the SS for bandwidth requests. **Best Effort (BE)** service does not specify any service related requirements. The requests are sent using contention request and unicast request opportunities. A fifth (5th) class is included in mobile WiMax (IEEE 802.16e) named as **extended rtPS**. It provides a scheduling algorithm that builds on the efficiency of both UGS and rtPS. These classes provide service differentiation mechanism by working with admission control and scheduling of different connections depending on the available resources of the network. The figure.1 shows frame structure of WiMax. The uplink and downlink frame constitute a whole frame.

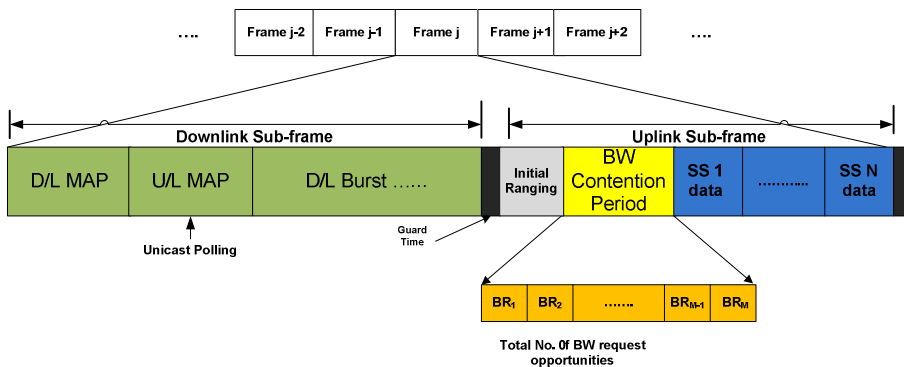


Fig. 1. Downlink and Uplink Sub-frame

When multimedia contents are transmitted over a WiMax network some amount of delay and jitter is introduced. Multimedia content like video, VOIP etc, have very strict delay and jitter requirements as compared to other types of data (web traffic). Due to sensitivity of multimedia content towards incurred delay and jitter, satisfactory performance is obtained mostly by prioritizing multimedia traffic. Due to error prone nature of wireless networks providing satisfactory QoS for real time data is more challenging in wireless networks. WiMax systems are now well established for providing broadband services over a wireless link. Now, it is essential to have an analytical model that characterizes delay and jitter experienced by multimedia content depending on overall system parameters. Thus in this work, we try to analytically model the behavior of WiMax system for providing satisfactory share of bandwidth to each class.

In WiMax system, the base station uses three types of polling to grant bandwidth to SS's. 1) Unicast polling. 2) Multi-cast polling 3) Broadcast polling. The unicast polling is shown in the figure 2. In this paper we aim to make a performance analysis for both ertPS and BE for video streaming. We test our results under different scenarios.

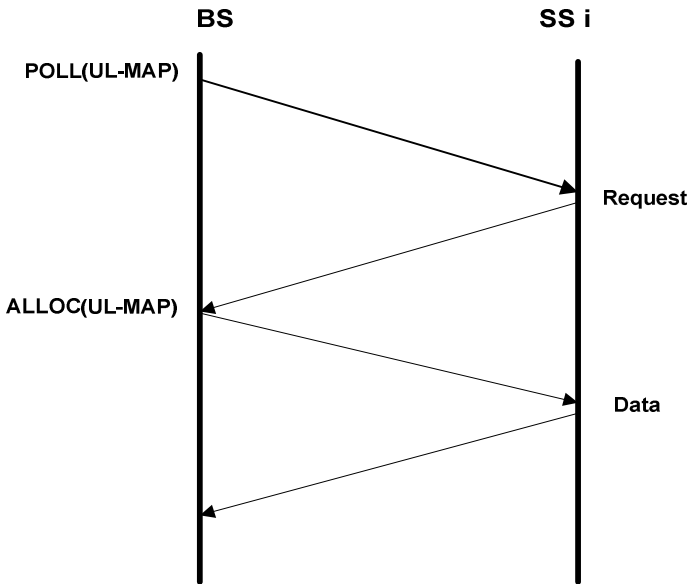


Fig. 2. Unicast Polling (for rtPS and nrtPS class)

2 Problem Definition

In this section, we formally describe the problem statement of our paper. WiMAX has defined five service classes primarily. All of these classes have their own service requirements and performance bounds. We try to make a performance comparison of

Best effort and ertPS. In this paper, we focus on the performance of BE traffic associated with more specific bandwidth allocation schemes. Since BE traffic does not have any specific delay or bandwidth requirement, high utilization and fair bandwidth sharing are the major concerns of BE scheduling. Since ertPS has stringent bandwidth and delay requirements, its scheduling is quite deterministic, and performance relies more on admission control rather than on scheduling [2].

We consider both the saturated and unsaturated mode of operation. For the sake of fair comparison, different flows are considered. Table 1. Shows a list of scenarios considered.

No. of BE flows	No of ertPS flows
Single (video streaming)	X
Multiple (Different applications)	X
Single (video streaming)	Single (video streaming)
Multiple (Different Applications)	Multiple (Different Applications)
X	Single (video streaming)
X	Multiple (Different applications)

All the scenarios are tested under both the saturated and unsaturated network conditions. Main objective is a thorough study of the service classes so that concrete conclusions could be reached.

We have used network simulator version 2 (ns-2) for simulations. Support for WiMAX module was developed in ns-2 in the release 2.28. We use the wireless channel available in the ns-2 environment. We have used C++ for implementation. The basic module design is based upon the Data Over Cable Service Interface Specifications (DOCSIS) [3] simulation module in ns-2. Code reuse is possible due to a lot of similarities in the structure of IEEE 802-16 and DOCSIS. Some modifications in the DOCSIS module code necessary to make it compliant with the IEEE 802.16 standard were (a) implementation of ertPS services for the uplink traffic, (b) implementation of all five types of service for the downlink traffic.

Every Service flow in WiMAX possesses four major features; Classifier, Queue, Allocation table, Finite state machine (FSM). Finite State Machines (FSMs) for BE uplink service flow is based on the state machines proposed by [4] for the DOCSIS simulation module. The FSM for the ertPS uplink service is specified to allow the sending of both data PDUs and bandwidth requests during the periodic grants allocated by the BS.

The QoS provided depends on admission control and scheduling mechanisms implemented in the BS and SSs. The scheduling mechanisms implemented in the SS scheduler and in the BS downlink scheduler follow a Strict Priority discipline. The BS uplink scheduler uses three queues; the low priority, intermediate and high priority queues. The scheduler serves the requests in strict order of priority, from the high priority queue to the low priority one. The low priority queue stores the bandwidth requests of the BE service flow.

3 Simulation Setup

We have used the simulator available at [6] and integrated it with ns-2.30. The topology consists of a BS wire attached to a fixed node through a 100 Mbps link with a 2 ms delay. The BS was located at the center of a 250 x 250 meter area, and the SSs were uniformly distributed around it. The frame duration was 5 ms and the capacity of the channel was 40 Mbps, assuming a 1:1 downlink-to-uplink TDD split. The module uses the wireless channel provided by the ns-2 simulator which has a Direct-Sequence Spread-Spectrum (DSSS) radio interface. We configure the simulated channel to provide the desired capacity. For simulation purpose we have simulated network of 5, 10 and 15 SSs only due to time limitations. Following four cases have been simulated:

- Multimedia traffic over BE only
- Multimedia traffic over ertPS only
- Multimedia traffic over BE and ertPS only
- Traffic of all classes. The data have been generated as follows:

Each SS had one uplink flow and one downlink flow, which were mapped to the same service type. Five types of traffic were considered: voice, voice with silence suppression, video, FTP, and WEB, which were associated with UGS, ertPS, rtPS, nrtPS, and BE services, respectively.

The voice model was an “on/off” one, with exponentially distributed period durations; the “on” and “off” periods lasted for a mean of 1.2 s and 1.8 s. During the “on” periods; packets of 66 bytes were generated every 20 ms [8]. The voice with silence suppression model used the Enhanced Variable Rate Codec (EVRC). Packets were generated every 20 ms using one of four different rates: Rate 1 (171 bits/packet), Rate 1/2 (80 bits/packet), Rate 1/4 (40 bits/packet) or Rate 1/8 (16 bits/packet). The rate selection was governed by a Markov chain. Video traffic was generated by real MPEG traces [7]. We have used baseball.dat video file from [7] for our simulation purpose. The WEB traffic was modeled by a hybrid Lognormal/Pareto distribution. The body of the distribution corresponding to an area of 0.88 was modeled as a Lognormal distribution with a mean of 7247 bytes, and the tail was modeled as a Pareto distribution with a mean of 10,558 bytes. FTP traffic was generated using an exponential distribution with a mean of 512 kBytes.

The interval between data grants for the UGS and ertPS services was 20 ms, since this is the generation rate of the application packets. The interval between unicast request opportunities of the rtPS service was 20 ms and the nrtPS service interval was 1 s. For rtPS service, the delay requirement was 100 ms, with each connection having its own minimum bandwidth requirement varying in accordance with the mean rate of the video transmission. The nrtPS service had a minimum bandwidth requirement of 200 Kbps, and the BE service did not have any specific QoS requirement.

4 Results and Inferences

This section assesses the throughput and end to end delay provided to the real time traffic by the scheduling mechanisms implemented in our module. We undertook a set

of simulations with the number of SSs increasing from 5 to 15 in steps of 5 units. Figs. 3 show the average throughput for downlink connections, with respect to the number of nodes for e rtPS, BE, both BE and ertPS and all the services. As expected, the throughput of all the scenarios increases as the load increases. And the increase remains proportional i.e., none of the connections show a peculiar behavior. An important point to note here is that when all the service classes are used at the same time, the throughput is much lesser than when only BE or ertPS are used. Moreover we can see that behavior of BE and ertPS is more or less the same everywhere.

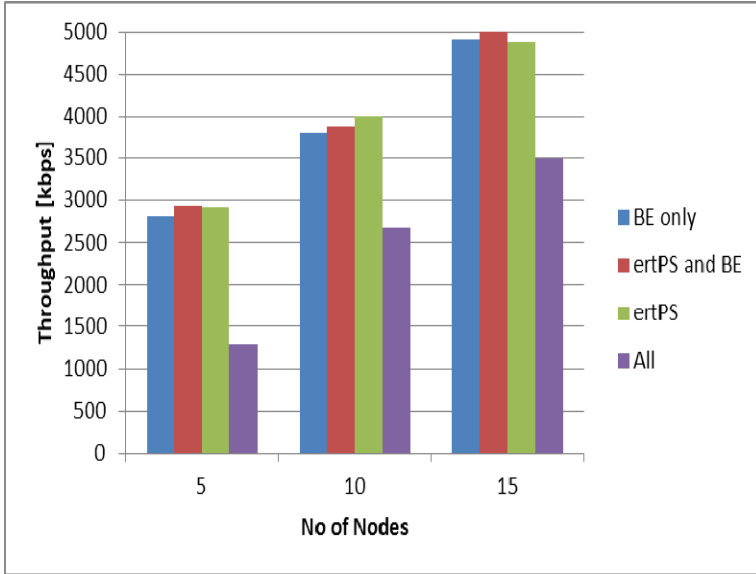


Fig. 3. Throughput analysis of WiMAX classes for different Number of Nodes

We conclude that same level of QoS can be obtained from BE and ertPS both and thus a separate service class for real time multimedia is not required. This conclusion is supported by the fact that the BS allocates resources for the BE service in the downlink direction even when this is competing with higher priority services [5].

In Fig. 4, we analyse the delay experienced by different number of services with respect to different number of nodes. With the increase in nodes the end-to-end delay for BE and ertPS decreases with a gentle slope as the number of SSs increase which shows that schedulers provide data grants at fixed intervals as required by these services. When using a mix of all the services, we see that end to end delay increases sharply as the number of nodes increase.

5 Future Work

We can see that the work presented here is not depicting actual scenarios where there are hundreds of nodes but as we have been able to successfully integrate and

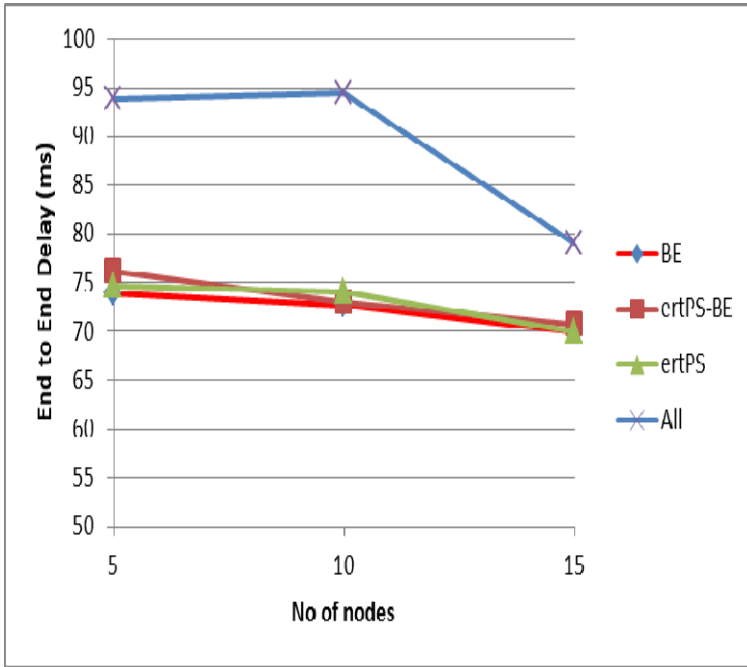


Fig. 4. End-to-end delay Vs. Number of nodes WiMAX QoS classes

understand (partially) the wimax module for ns2, our next step is in-depth study of different QoS classes and what performance gains do they provide.

Acknowledgments. This work is supported by Myongji College with Prof. Dong-Young Lee as corresponding author.

References

- [1] IEEE 802.16e/D12-2005, IEEE Standard for local and metropolitan area networks – part 16: air interface for fixed and mobile broadband wireless access systems - amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1 (October 2005)
- [2] Kim, S., Yeom, I.: Performance Analysis of Best Effort Traffic in IEEE 802.16 Networks
- [3] Cable Television Labs Inc., Data Over Cable Service Interface Specifications – Radio Frequency Interface Specification, SP-RFiv2.0
- [4] Shrivastav, N.: A network simulator model of the DOCSIS protocol and a solution to the bandwidth-hog problem in the cable networks. Master Dissertation, North Carolina State University, EUA (2003)
- [5] Borin, J.F., da Fonseca, N.L.S.: Simulator for WiMAX networks

- [6] http://www.lrc.ic.unicamp.br/wimax_ns2/
- [7] Brady, P.: A model for generating on-off speech patterns in two-way conversations. *Bell System Technical Journal* 48, 2445–2472 (1969)
- [8] Seeling, P., Reisslein, M., Kulapala, B.: Network performance evaluation using frame size and quality traces of single-layer and two-layer video: a tutorial. *IEEE Communications Surveys and Tutorials* 6(2), 58–78 (2004)

Jump Oriented Programming on Windows Platform (on the x86)

Jae-Won Min¹, Sung-Min Jung¹, Dong-Young Lee², and Tai-Myoung Chung¹

¹ Dept. of Computer Engineering, Sungkyunkwan University
300 Cheoncheon-dong, Jangan-gu, Suwon-si, Gyeonggi-do, 440-746, Korea
{jwmin, smjung}@imtl.skku.ac.kr, tmchung@ece.skku.ac.kr

² Dept. of Information and Communication, Myong-ji College
356-1 Hongeun3-Dong, Seodaemun-Gu, Seoul 120-776, Korea
dylee@mjc.ac.kr

Abstract. Non-executable memory pages were deployed in operating systems in order to defend against code injection attacks. However, it was bypassed by reusing codes that already exist in the process memory which have the execute permission. The Return-Oriented Programming (ROP), of the most well-known code reuse attack, has been developed and widely used to exploit systems. ROP hijacks the control flow and returns to the middle of instruction sequences that end with a return instruction. These instruction sequences are called gadgets. Researchers proposed many ROP defense mechanisms which mostly relied on the fact that ROP executes many return instructions. Proposed defenses however, are not fundamental defenses. Researches found that the concept of ROP can be implemented in Linux using jump instructions instead of return instructions, therefore successfully bypassing ROP defenses. However, no research was done on implementing the attack on non-Linux systems. In this paper, we show the possibility of implementing JOP (Jump Oriented Programming) attack model on Windows platform by presenting example gadgets and propose an algorithm for searching JOP gadgets in Dynamic Link Libraries.

Keywords: Security, Jump Oriented Programming, Code Reuse Attack.

1 Introduction

Attackers subvert normal control flows of the program to make the program misbehave and do malicious activities. Traditionally, attackers injected malicious codes and gained control of the instruction pointer by overwriting the return address which is pushed to the stack and executed the injected codes. To prevent such attacks, various countermeasures were proposed. For example, GS option of Visual C++ compiler pushes a random value to the stack before the return address to detect buffer overflows. Another widely used technique, Data Execution Prevention (DEP) prevents code execution in memory pages that are marked as non-executable [1].

To bypass code injection protections, attackers started executing codes without injecting any codes at all. Commonly known as code reuse attack, this type of attack is a widely used exploit technique which uses codes that are already inside the memory space to bypass non-executable memory techniques in the operating system. Solar Designer [2] demonstrated a technique called return to libc, which made the program return to a linked C library function which is executable. In Fig. 1, attacker overflows the vulnerable buffer and overwrites the return address with the address of the system() function. As a result, instead of returning to the legitimate caller, function returns to the system() function.

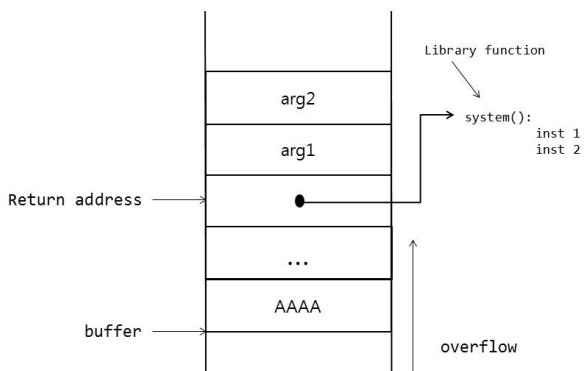


Fig. 1. Return to libc

Return-Oriented Programming (ROP) [3] is a type of code reuse attack similar to return to libc, which executes small sequence of instructions that ends with a return instruction instead of returning to the start of a function. Following assembly codes are an example of a ROP gadget.

```
pop eax
pop ecx
ret
```

By chaining these gadgets together, attackers can execute high level operations like function calls. After the first gadget ends with a return, system pops the return address from the stack, which is the address of the next gadget. When the second gadget finishes execution, it pops the address of the third gadget and this process is repeated until the last gadget is executed. In practice, ROP attacks are used exploit various systems. For example, ROP is used to jailbreak iOS devices and exploit vulnerabilities that exist in applications and operating system kernels.

Thus, security researchers proposed various techniques to protect the system against ROP attacks. Davi et al. [4] and Chen et al. [5] proposed ROP defense systems detecting frequently executed return instructions using dynamic binary instrumentation (DBI) framework. Moreover, Stackghost [6] system used shadow return address stack to protect against ROP attacks.

Later, to bypass ROP defenses, Checkoway et al. [7] proposed a ROP technique without using return instructions. Proposed technique uses indirect jump instructions instead of return instructions. Because return instructions are not used in the attack, it successfully bypasses ROP mitigations. Bletsch et al. [8] extended the research and proposed a similar attack which built a Turing complete gadget set using only GNU C library while Checkoway et al. [7] requires additional libraries to complete Turing complete gadget set. The exploit techniques that uses indirect jump instructions are often called Jump Oriented Programming (JOP).

Both [7,8] have done an innovative research, proposing a new paradigm in code reuse attacks. However there exist some limits. JOP gadget sets were built in GNU C library 2.7 and the attack was tested in Debian 5.0.4 Linux system which is an outdated version. Furthermore, no research is done on non-Linux based operating systems.

Therefore, we describe the possibility of JOP attack in Windows platform (x86 32-bit architecture) and present in detail the proposed algorithm for searching JOP gadgets. Finally, we show the results of the analysis of proposed techniques.

2 Related Work

2.1 Return Oriented Programming

Solar Designer's return to libc attack had limits that although it can call C library functions sequentially, it cannot execute branch operations. Moreover, removal of functions used in the attack from the linked libraries results reduction of capability of the attack. In 2007, Hovav Shacham [3] introduced a new exploit technique that is able to do arbitrary computations without calling any functions. This techniques is called Return Oriented Programming and Fig. 2 shows the building of ROP. The building block of ROP is short code sequences that end with a return instruction.

After the control flow is hijacked by an attacker, control returns to the first gadget and the gadget is executed. The first gadget ends with return instruction which increases the stack pointer and pops the address of the second gadget to the instruction pointer. This process continues. By chaining gadgets, attackers can perform any desired operations. However, not all instruction sequences are useful for ROP attacks. If instructions that change the control flow are placed before the return instruction, then that instruction sequence is useless for the attack.

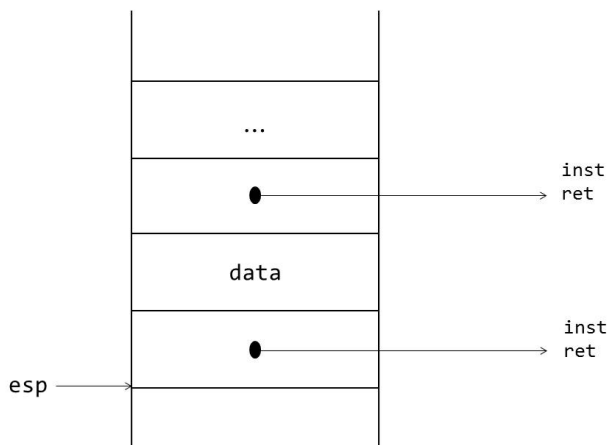


Fig. 2. The ROP model

ROP has some unique characteristics that are detectable. First, return instructions are executed more frequently than ordinary functions. This is because every gadget ends with a return instruction and usually the ROP attack payload contains many gadget addresses. Second, when the ROP attack is executed, the Last-in First-out property is not met. Therefore by keeping a shadow stack and checking the return address, ROP attack can be detected because instruction pointer does not point to the return address that was pushed earlier to the stack.

2.2 Jump Oriented Programming

ROP without returns [7] proved that it is possible to initiate the attack with gadgets that ends with an indirect jump instruction instead of return instruction. Unlike original ROP attack, it requires some mechanism to control the instruction pointer. ROP without returns model make use of `pop x; jmp *x` instruction sequence. Variable `x` can be any general purpose register. This instruction sequence does same job as a return instruction. The attack model is described in Fig. 3.

Memory addresses of gadgets that must be sequentially executed are saved in a data structure called sequence catalog. After the instruction pointer is hijacked, first gadget is executed. End of each gadget, there is a jump instruction that changes the instruction pointer to `pop x; jmp *x` sequence, which chains each gadgets. Data can be pushed to the stack and used by the gadget with a `pop` instruction.

JOP model proposed by Bletsch et al. [8] differs from [7] in the way it controls the instruction pointer. The major drawback of [7] is that `pop x; jmp *x` sequence

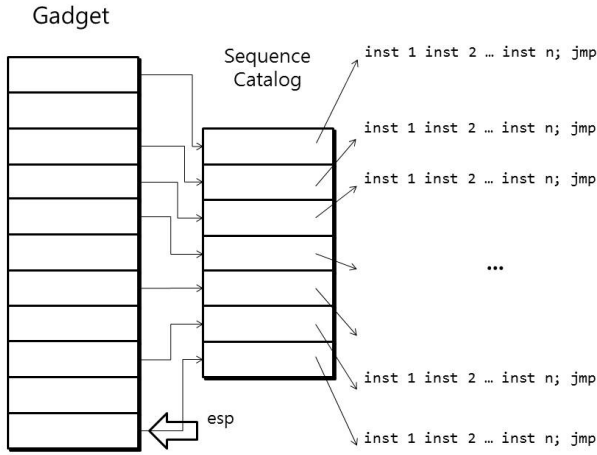


Fig. 3. The Return Oriented Programming without returns model [7]

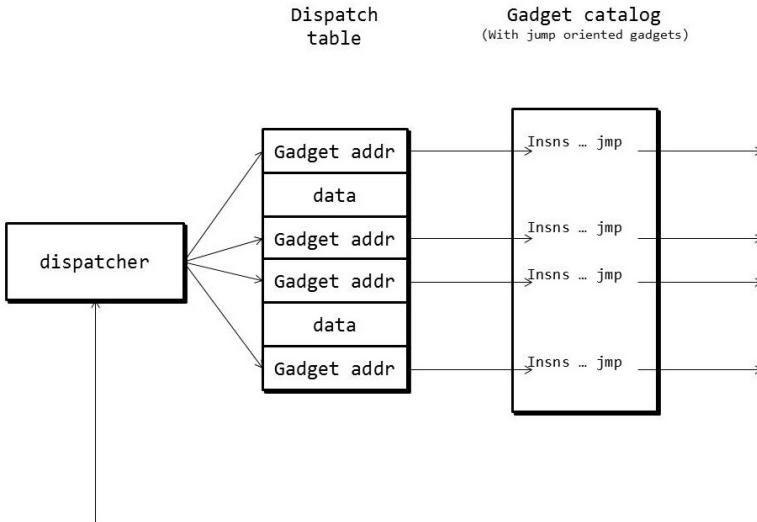


Fig. 4. The JOP model [8]

is quite rare and is not found in GNU C library. Therefore, instead of searching for `pop x; jmp *x` sequence, JOP model uses what is called a dispatcher gadget. This special type of gadget constantly adds a fixed value to a register. This register can be used as a pointer to entries of the dispatch table. Dispatch table is a table of gadget addresses. Advantage of this model is that dispatcher gadget is more frequently found and do not have to rely on stack to save attacker's data. The attack model is described in Fig. 4.

Similar JOP research was also done on ARM architecture [9]. ARM has different instruction set from x86 instruction set. However, using BLX instruction, JOP attack can be initiated successfully.

2.3 Limits

Although ROP is very popular and various platforms are vulnerable to it, a lot of research was done on mitigating the attack making it hard to use ROP to attack target programs. Research done on indirect jump based attack [7,8] opened new possibilities of the code reuse attack. However, it was only shown on Debian Linux 5.0.4 and ARM processor based Android 2.0 [9]. These systems are very outdated (latest Android release is version 4.0) and considering the percentage of Linux or Android users, we cannot say the JOP technique can be widely spread because target operating systems consume less than 10 percent.

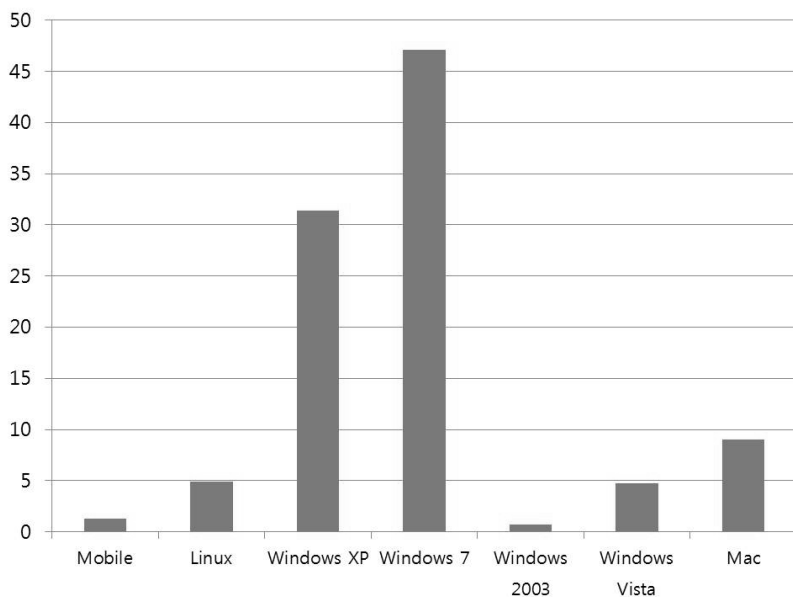


Fig. 5. Operating System statistics [10]

Interest should move on to JOP from ROP and research should be done on both offense and defense. Since Windows is the most widely used operating system, it motivated us to research further. We searched through the library space of the Windows operating system to find valid JOP gadgets.

3 Finding Gadget Sets

Gadgets were searched in the Dynamic Link Library files in the Windows XP Service Pack 3. Kernel32.dll file was mainly searched because it is linked to every program. We wrote a script in Python to automate the process of searching for JOP gadget candidates. Mainly our script uses two Python modules: pefile and pydasm. Pefile is used to parse the Portable Executable file format of the DLL and locate the text section. Pydasm is a disassembler for x86 instruction set architecture. Pydasm is used to disassemble codes in the text section to see whether indirect jump instructions exist. Portable Executable format is depicted in Fig. 6.

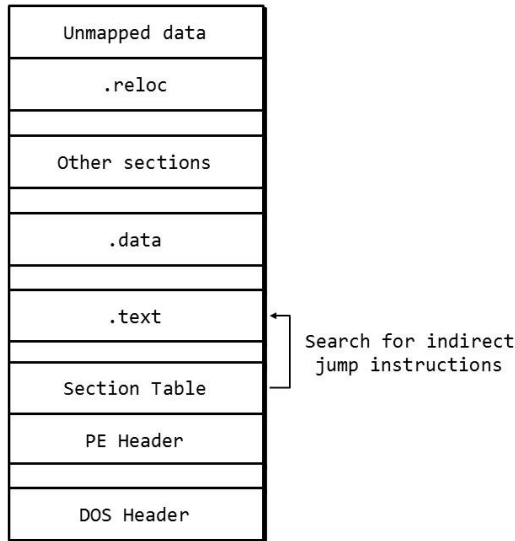


Fig. 6. Portable Executable file format

3.1 Dispatcher Gadget

To construct the JOP attack, we follow the model proposed in [8]. We need a dispatcher gadget to make gadgets execute sequentially. Dispatcher gadget must change the value of the register that is being used by constant amount. Dispatcher gadget acts like a pointer that traverses every entry in the catalog of gadget addresses. Such dispatcher gadgets are found in kernel32.dll and one of them is presented below as an example.

```
add ecx, edi
jmp [ecx+0x7c8856cc]
```

This gadget increases the value of ecx register by the value saved in the edi register every time it is executed. By placing addresses of gadgets separated by the amount of the edi register's value, each gadget can be executed.

3.2 Other Gadgets

Other than the dispatcher gadget, we must find gadgets that actually do certain operations that the attacker wants. One of the basic operations is the data movement operation. We considered three cases in the data movement operation:

1. Loading immediate value to a register
2. Loading memory value to a register
3. Storing register value to a memory space

The first case can be done by popping a immediate value to a register. Example gadget below can be used to load a immediate value in the stack to the edi register. [ecx+0x18] should point to the dispatcher gadget.

```
pop edi
jmp [ecx+0x18]
```

Using mov instruction, certain value can be moved from a memory address to a register, which is the second case. Example JOP gadget is as follows. It copies a value in memory address pointed by esi-0x2 to the ebx register and jumps to the dispatcher gadget to execute the next gadget.

```
mov ebx, [esi-0x2]
jmp [ecx+0x18]
```

The last case, which is storing a value to a memory address from a register, can be similarly done using mov instruction. This gadget we found copies the value of the eax register to memory at esi-0x63611784.

```
mov [esi-0x63611784], eax
stc
jmp [ecx+0x7c8856cc]
```

Typical arithmetic operations like addition and subtraction are done by loading values to the source registers, doing the actual operation and saving the result to the destination memory at the end. Loading values can be done by data movement gadgets. Arithmetic gadgets are commonly found in the Windows DLL files. For example, following gadget was found in kernel32.dll and can be used to implement addition operation.

```
add esi,edi
jmp [ecx+0x18]
```

Gadget below can be used to do a subtraction in the register value. It is very similar to addition gadget. In the case of negation, subtracting a value from zero has same effect.

```
sub esi,ebp
cld
jmp [ecx+0x18]
```

Logical operations are similar to arithmetic operations. Operands are loaded and execution result is saved in memory. Attackers can use other gadgets to load data into the register and then use logical operation gadget to do logical operations with the register. For example, following gadget can be used to execute a exclusive-or instruction.

```
xor edi,ebp
jmp [ecx+0x18]
```

Gadget below is an example which can be used for and operation.

```
and ebx, eax
clc
jmp [ecx+0x7c8856cc]
```

Branch operation is more complex than the other operations. There are two kinds of branch operations: conditional branch and unconditional branch. Unconditional branch can be done by modifying register or memory value that controls the instruction flow [8]. Conditional branch operations execute based on the flag set by previous operations. There are many ways to simulate the conditional branch. One of them is a special instruction called cmov, which can be used to change the instruction pointer in certain conditions [8]. However, there were no usable conditional branch gadgets found in kernel32.dll. Therefore we should continue the research to find working branch gadgets in other DLL files.

Although building block of JOP attack is a short instruction sequence, sometimes calling a complete function might be necessary. Function call can be done by using a technique similar to return-to-libc. Attackers can push function arguments to the stack and jump to the function call gadget. The gadget executes call instruction using previously pushed function arguments. We present an example function call gadget that was found in kernel32.dll file. Call instruction is followed by a jump instruction so the attacker can keep control of the flow of the process after function call returns. We present function call gadget example below.

```
call [eax-0x18]
stc
sar bh, 0x1
jmp [ecx+0x18]
```

By placing required arguments in the stack and storing the address of the target function minus 0x18 in the eax register, attacker can call arbitrary functions.

4 Proposed Algorithm and Scenario

4.1 Proposed Algorithm

We made an algorithm to search for JOP gadgets in Windows DLL files. Algorithm searches the text section of the file for indirect jump instructions. If an jump instruction is found, it traces back constant amount of bytes and start finding every unintended instructions from that position. All the invalid gadgets should be ignored and only valid JOP gadget should be stored. If a valid gadget was found, the algorithm tags offsets of instructions in the gadget. Disassembling position is moved to the next position with no tags and this step repeats until the position reaches the offset of previously found jump instruction. Graphical depiction of the algorithm is presented in the Fig 7. Following are cases when gadget candidates are considered invalid.

1. There is a branch instruction before the jump instruction
2. Instructions change the value of the register which stores the memory address of the dispatcher gadget
3. Bytes does not disassemble to a legal x86 instruction

Eliminating invalid gadgets, we can obtain list of usable gadgets. By carefully chaining memory addresses of found gadgets, we can build an working exploit.

Gadget Searching Algorithm

ALGORITHM Search(C)

```

FOR each position that is an indirect jump in C
  Save the position of the indirect jump instruction
  Step back few bytes
  WHILE current position < position of jump instruction
    IF the current position is not tagged THEN
      Disassemble bytes starting from current position
      IF Disassembled bytes are a valid gadget THEN
        Tag offsets of disassembled instructions
        Save the gadget
      ENDIF
    ENDIF
    INCREMENT current position
  ENDWHILE
ENDFOR

```

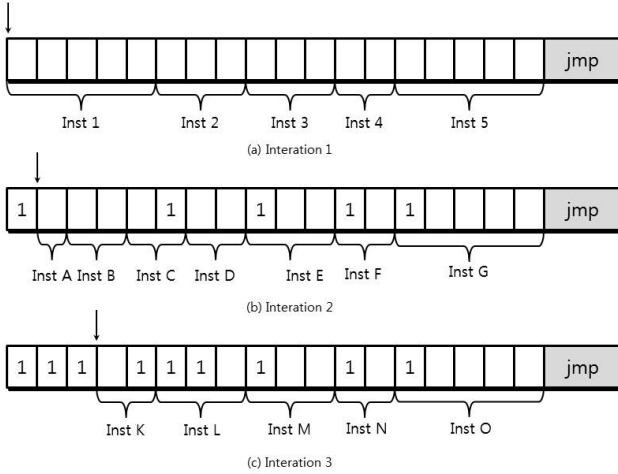


Fig. 7. Graphical depiction of the proposed algorithm

(a) in Fig. 7 shows the initial status of the algorithm. After disassembling Inst 1 to Inst 5, it tags the offset of each instruction as shown in (b). Again, disassembling starts at the next untaged offset and results Inst A to Inst G. It iterates until every offsets are tagged.

4.2 Evaluation

Assume that the number of bytes in the text section is n and amount of bytes stepped back is constant c . Then, the outermost for loop of the algorithm is repeated n times. The while loop depends on the number of tags set on the offsets. The more tags set on each scan, the less number of bytes traversed for finding gadgets. In the worst case, each scan will tag only one offset. Number of bytes traversed while disassembling, will be

$$bytes = \sum_{i=0}^c (c - i) \tag{1}$$

Because this is a constant value, it does not effect the overall time complexity of the algorithm. The complexity depends only on the size of the text section, thus results $O(n)$. [8] proposed a similar algorithm for searching gadgets but there are some important differences. First, [8] follows a right-to-left approach for disassembling potential gadgets which result in making sub-gadgets of already disassembled gadgets as described in (a) of Fig. 8.

There is no need to disassemble sub-gadgets again because they are already included in the larger gadgets. However, gadgets in (b) are considered independent because they have a unique instruction and cause different result when executed. Proposed algorithm in this paper does not produce any sub-gadgets by using tags.

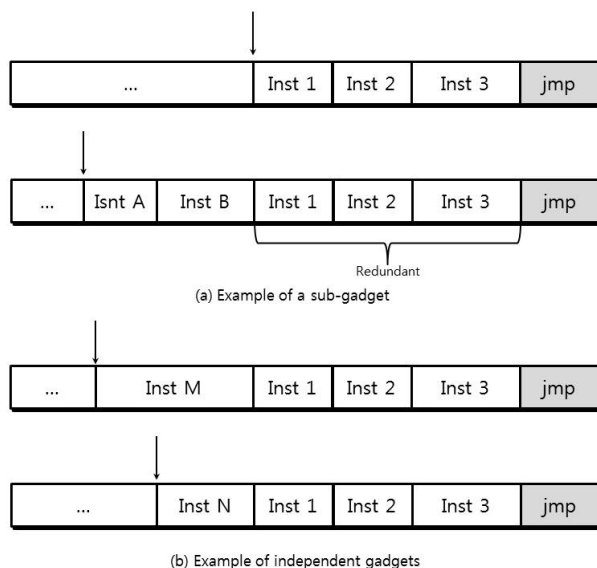


Fig. 8. Example of a sub-gadget and an independent gadget

Second, in the context of bytes traversed, our algorithm is more efficient because it does not have to disassemble every offset while algorithm proposed in [8] disassembles every offset.

4.3 Scenario

There are many known class of software bugs that give the attacker the control of instruction pointer: stack overflow, heap overflow and etc. Primary goal of exploiting these vulnerabilities is controlling the instruction pointer and executing arbitrary code. Any of these known vulnerabilities can be used to trigger the JOP attack payload. Once the instruction pointer points to the address of the first gadget, following gadgets are executed sequentially.

Fig. 9 shows a simple attack scenario that we presented. In the first phase, an attacker finds an exploitable vulnerability in the target program by manual source code auditing or fuzzing. In the next phase, attacker searches the code space of the target program for valid JOP gadgets. Found gadgets are chained together to execute high level operations like function calls. In this scenario, attacker will try to open a reverse shell. After the exploit code is implemented, computer agents searches for hosts with vulnerable target program installed. If vulnerable host is located, attacker launches the exploit.

The victim receives attack payload which triggers the vulnerability inside the target program. Chained gadgets are executed in the victim's system which gives the attacker the root access to the system.

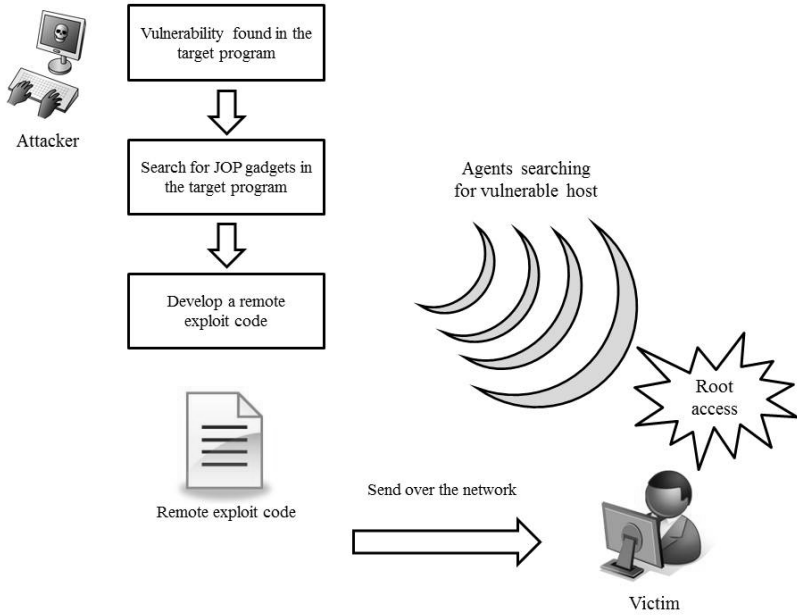


Fig. 9. Attack scenario

4.4 Possible Defenses

Similar to ROP, JOP also has some characteristics that can be used to detect the attack. We list characteristics below.

1. JOP attacks use some kind of trampoline to transfer the control to gadgets.
2. Indirect jump instructions are abnormally frequently used.
3. The stack pointer keeps on moving to higher address of the memory.

First, all JOP has some kind of a trampoline that transfers the control to each chained gadgets. By tracing indirect jump targets, we can decide if it is a legit jump instruction or a trampoline, which is in our case, dispatcher gadget. Another possible way of detecting the attack is checking the number of jump instructions executed. Abnormally frequent execution of indirect jump instructions raises possibility of JOP attack. Tracing the movement of the stack pointer also gives a clue for detecting the attack. Because the stack pointer constantly pops data from the stack without allocating local variables, the pointer most likely moves upward toward the high memory address. However a more fundamental defense is control flow integrity. Drawing graph of legit control flow prevents attackers from hijacking the control flow by code reuse. Some papers were published on this topic [11].

5 Conclusion

In this paper, we analyzed previous researches about JOP and discussed some drawbacks in their work. Furthermore, we showed that useful JOP gadget sets are found in Windows default C Library file, which is kernel32.dll. We built a tool written in Python to statically look for JOP gadgets using algorithm we proposed. By utilizing found gadgets, it is possible for attackers to initiate JOP attacks on Windows XP Service Pack 3 (x86 architecture). For possible defense, we showed that detection based on JOP characteristics is possible; however more fundamental defense is control flow integrity.

In the future, we plan to construct a working JOP exploit that attacks up-to-date software, not exploiting outdated programs or unrealistic concept code. Moreover, analyzing other operating systems such as Windows 7 or mobile operating systems like iOS is an interesting topic. Lastly, researching about 64 bit instruction set architecture might open new possibilities in exploiting techniques because it introduces new registers and instructions. Defenses on these research areas should be followed also.

Acknowledgements. This work was supported by the IT R&D program of MKE/KEIT. [KI001810039260, Integrated dev-environment for personal, biz-customized open mobile cloud service and Collaboration tech for heterogeneous devices on server].

References

1. Sotirov, A., Dowd, M.: Bypassing Browser Memory Protections: Setting back browser security by 10 years, Blackhat (2008)
2. Solar Designer: Getting around non-executable stack (and fix), Bugtraq (August 1997)
3. Shacham, H.: The Geometry of Innocent Flesh on the Bone: Return-into-libc without Function Calls (on the x86). In: Proceedings of the 14th ACM Conference on Computer and Communications Security, pp. 552–561 (2007)
4. Davi, L., Sadephi, A.-R., Winandy, M.: Dynamic integrity measurement and attestation: Towards defense against return-oriented programming attacks. In: Asokan, N., Nita-Rotaru, C., Seifert, J.-P. (eds.) Proceedings of STC 2009, pp. 49–54. ACM Press (2009)
5. Chen, P., Xiao, H., Shen, X., Yin, X., Mao, B., Xie, L.: DROP: Detecting Return-Oriented Programming Malicious Code. In: Prakash, A., Sen Gupta, I. (eds.) ICISS 2009. LNCS, vol. 5905, pp. 163–177. Springer, Heidelberg (2009)
6. Frantzen, M., Shuey, M.: StackGhost: Hardware facilitated stack protection. In: Wallach, D. (ed.) Proceedings of Usenix Security 2001, pp. 55–65. USENIX (2001)
7. Checkoway, S., Davi, L., Dmitrienko, A., Sadeghi, A.-R., Shacham, H., Winandy, M.: Return-oriented programming without returns. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, pp. 559–572. ACM, New York (2010)

8. Bletsch, T., Jiang, X., Freeh, V.W., Liang, Z.: Jump-Oriented Programming: A New Class of Code-Reuse Attack. In: ASIACCS, Boston, vol. 4865, pp. 154–165 (2011)
9. Davi, L., Dmitrienko, A., Sadeghi, A.-R., Winandy, M.: Return-Oriented Programming without Returns on ARM. Technical Report (2010)
10. OS Platform Statistics, http://www.w3schools.com/browsers/browsers_os.asp
11. Bletsch, T., Jiang, X., Freeh, V.: Mitigating Code-Reuse Attacks with Control-Flow Locking. In: Proceedings of the 27th ACSAC (2011)

Cryptanalysis and Improvement of a Biometrics-Based Multi-server Authentication with Key Agreement Scheme

Hakhyun Kim¹, Woongryul Jeon¹, Kwangwoo Lee¹,
Yunho Lee², and Dongho Won^{1,*}

¹Information Security Group,
School of Information and Communication Engineering, Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Korea

²Department of Cyber Security & Police, Gwangju University, 52 Hyoduk-ro,
Nam-gu, Gwangju-si, 503-703, Korea

{hhkim, wrjeon, kwlee, dhwon}@security.re.kr, leeyh@gwangju.ac.kr

Abstract. In 1981, Lamport proposed a password authentication scheme to provide authentication between single user and single remote server. In a smart card based password authentication scheme, the smart card takes password as input, makes a login message and sends it to the server. Many smart card based password authentication schemes with a single server have already been constructed. However it is impossible to apply the authentication methods in single server environment to multi-server environment. Therefore, some smart card based password authentication schemes for the multi-server environment are proposed. In 2010, Yoon et al. proposed a robust biometrics-based multi-server authentication with key agreement scheme for smart cards on elliptic curve cryptosystem. In this paper, however, we show that scheme of Yoon et al. is vulnerable to off-line password guessing attack and propose an improved scheme to prevent the attack.

Keywords: cryptanalysis, key agreement, authentication, biometrics.

1 Introduction

In 1981, Lamport proposed a password authentication scheme[1] to provide authentication between single user and single remote server. In a smart card based password authentication scheme, the smart card takes password as input, makes a login message and sends it to the server. Many smart card based password authentication schemes with a single server have already been constructed [4]. However it is impossible to apply the authentication methods in single server environment to multi-server environment. Therefore, some smart card based password authentication schemes for the multi-server environment are proposed [5-17]. Those schemes can be divided in two types, namely hash-based authentication and public-key based authentication. In 2001, Li *et al.* constructed remote password authentication scheme in

* Corresponding author.

the multi-server environment [5]. However, the scheme is not efficient nor practical, because a smart card cannot execute the computation of neural networks in a short time. In 2004, Tsaur *et al.* improved Li *et al.*'s scheme [18].

However, Tsaur *et al.*'s improved scheme is still insecure, so Juang newly proposed an efficient multi-server user authentication and key agreement based on hashing function and symmetric-key cryptosystem [19]. But Juang's scheme is vulnerable to an online guessing attack [20] and cannot withstand off-line dictionary attacks [7, 21]. Chang and Lee proposed an improved scheme [20], however, the proposed scheme cannot withstand the insider attack, server spoofing attack and registration center spoofing attack [10]. Hu *et al.* proposed an efficient multi-server password authenticated key agreement scheme using smart cards [10]. But it does not satisfy the security requirement in the multi-server environment [17]. Tsai proposed an efficient multi-server authentication scheme based on one-way hash function without verification table [22], however, Tsai's scheme is insecure. Recently, Yoon *et al.* proposed a robust multi-server authentication with key agreement scheme for smart cards using biometrics and elliptic curve cryptosystem [2].

In this paper, we will demonstrate that Yoon *et al.*'s scheme cannot prevent an adversary from off-line password guessing attack and propose an improved scheme to prevent the attack.

The remainder of this paper is organized as follows: In Section 2, we introduce related works. In Section 3, we review of Yoon *et al.*'s scheme. In Section 4, we demonstrate off-line password guessing attack on Yoon *et al.*'s scheme. In Section 5, we propose an improved scheme to prevent off-line password guessing attack. In Section 6, we perform security analysis. Finally, conclusions will be given in Section 7.

2 Related Works

2.1 Biometric Authentication

There are three kinds of approaches for user authentication as follows:

- (1) Password based user authentication ("what you know"): Passwords and PINs are the examples of this approach.
- (2) Token-based user authentication ("what you have"): This method includes physical keys, ATM or smart cards, mobile devices (cell phones, PDA, RFID, sensor nodes) and so on.
- (3) Biometric-based user authentication ("what you possess"): Voice, fingerprints, iris, and keystrokes are included in this method.

Because of their cryptographic capacity and portability, smart cards have been widely used in many network applications. Besides, biometrics hold the promise of fast, easy-to-use, accurate, reliable, and less expensive authentication for variety of applications. Biometric authentication requires comparing a registered biometric information sample against a newly captured biometric information sample, e.g., a fingerprint captured during a login. During registration procedure, a sample of the biometric trait is captured, processed by a computer, and stored for later comparison

(see Fig. 1). For biometric recognition, the biometric system authenticates a person’s claimed identity from their previously registered pattern in verification procedure [23–25] (see Fig. 1). For smart card-based biometrics authentication, a user inserts a smart card in to a device, a simple touch with a finger or a glance at a camera is enough to authenticate the user.

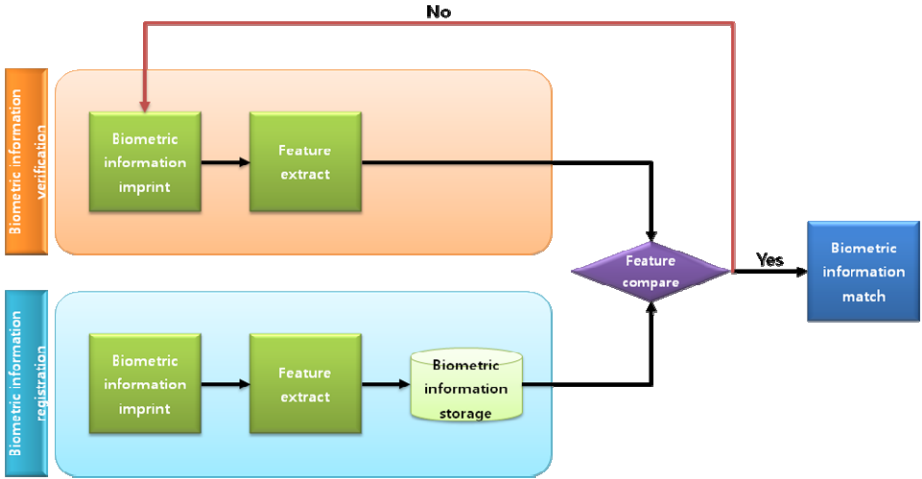


Fig. 1. Biometric information process flow chart

3 Review of Yoon et al.’s Scheme

This section reviews Yoon *et al.*’s multi-server authentication with key agreement scheme briefly. The scheme is composed of four phases; the server registration phase, the user registration phase, the authenticated key agreement phase, and the password and biometrics update phase. Notations used in this scheme are defined as follows.

- U, S_j : the user and the j th server, respectively.
- RC : the registration center.
- ID, PW, B : U ’s identity, password and biometric template, respectively.
- SID_j : S_j ’s identity.
- x : U ’s secret key maintained by the registration center.
- y : S_j ’s secret key maintained by the registration center.
- p : large prime number.
- F_p : finite prime field.
- E : non-super singular Elliptic curve over a finite field F_p , where $E : y^2 = (x^3 + ax + b) \bmod p$ with $a, b \in F_p$ satisfying $(4a^3 + 27b) \bmod p \neq 0$.

- $E(F_p)$: additive group of points on E over a finite field F_p , where $E(F_p) = \{(x,y) : x, y \in F_p \text{ satisfy } y^2 = x^3 + ax + b\} \cup \{O\}$.
- P : generating element (point) of $E(F_p)$ under consideration F_p .
- α, β : session-independent random integer numbers $\in [1, p-1]$ chosen by U and S_j , respectively.
- SK : shared fresh session key computed by U and S_j .
- $d(\cdot)$: symmetric parametric function
- τ : predetermined threshold for biometric verification.
- $h(\cdot)$: secure one-way hash function.
- \oplus : bit-wise exclusive-or(XOR) operation.
- \parallel : concatenation operation.
- $A \rightarrow B: M$: A sends a message M to B .

3.1 Server Registration Phase

The following steps are performed during the server registration phase.

Step 1: $S_j \rightarrow RC : SID_j$

S_j chooses its identity SID_j and transmits it to RC via a secure channel.

Step 2: $RC \rightarrow S_j : h(SID_j \parallel y)$

RC computes $h(SID_j \parallel y)$, where y is the S_j 's secret key maintained by RC , and transmits it to S_j via a secure channel.

3.2 User Registration Phase

The following steps are performed during the user registration phase.

Step 1: $U \rightarrow RC : \{ID, h(PW \parallel B), B\}$

U selects ID and PW , and imprints her biometric information B at the sensor. Then U submits $\{ID, h(PW \parallel B), B\}$ to RC . These private data must be sent in a secure channel.

Step 2: $RC \rightarrow U : \{\text{Smart card containing } (Z, B, h(\cdot), d(\cdot), \tau)\}$

RC computes the user authentication key $R = h(ID \parallel x)$ and $Z = R \oplus h(PW \parallel B)$.

Then, RC stores $\{Z, B, h(\cdot), d(\cdot), \tau\}$ in a smart card and issues it to U via a secure channel.

3.3 Authenticated Key Agreement Phase

The following steps are performed during the authenticated key agreement phase.

Step 1: $U \rightarrow S_j : \{ID, \alpha P, C_1\}$

U inserts her smart card into a card reader, executes the login application software, and imprints biometric B^* at the sensor. Then a biometric verification process of U 's smart card compares the imprinted B^* with the stored B . If $d(B^*, B) < \tau$, then it outputs accept message. If $d(B^*, B) \geq \tau$, then it outputs reject message

which means U does not pass the biometric verification thus the authentication process is terminated. On the contrary, if it outputs accept, U enters her password PW , and then the reader extracts R by computing $Z \oplus h(PW\|B)$, generates a random number $\alpha \in [1, q-1]$, and computes $C_1 = h(R\|\alpha P) = h(h(ID\|x)\|\alpha P)$. Then U transmits $ID, \alpha P$ and C_1 to S_j .

Step 2: $S_j \rightarrow RC : \{ID, \alpha P, C_1, SID_j, \beta P, C_2\}$

S_j generates a random number $\beta \in [1, q-1]$ and computes $C_2 = h(h(SID_j\|y)\|\beta P)$. Then S_j transmits $ID, \alpha P, C_1, SID_j, \beta P$ and C_2 to RC .

Step 3: $RC \rightarrow S_j : \{C_3, C_4\}$

RC computes $C'_1 = h(h(ID\|x)\|\alpha P)$ and $C'_2 = h(h(SID_j\|y)\|\beta P)$ and then RC checks whether $C_1 \stackrel{?}{=} C'_1$ and $C_2 \stackrel{?}{=} C'_2$, respectively. If both equations hold, RC computes the follows:

$$V = h(h(SID_j\|y)\|\beta P\|\alpha P),$$

$$W = h(h(ID\|x)\|SID_j\|\alpha P\|\beta P),$$

$$C_3 = V \oplus W, \text{ and}$$

$$C_4 = h(V\|W),$$

where W is used to the ephemeral secret key between U and S_j . Finally, RC transmits C_3 and C_4 to S_j .

Step 4: $S_j \rightarrow U : \{\beta P, C_5\}$

S_j computes $V' = h(h(SID_j\|y)\|\beta P\|\alpha P)$, and extracts the ephemeral secret key W by computing $C_3 \oplus V' = V \oplus W \oplus V' = W$. Then S_j computes $C'_4 = h(V'\|W)$ and checks whether $C_4 \stackrel{?}{=} C'_4$. If it holds, S_j computes the shared session key $SK = \beta(\alpha P) = \alpha\beta P$, and $C_5 = h(ID\|SID_j\|W\|SK)$. Finally, S_j transmits βP and C_5 to U .

Step 5: $U \rightarrow S_j : \{C_6\}$

U computes the ephemeral secret key W , the shared session key SK and C'_5 as follows:

$$W = h(h(R\|SID_j\|y)\|\alpha P\|\beta P),$$

$$SK = \beta(\alpha P) = \alpha\beta P,$$

$$C'_5 = h(ID\|SID_j\|W\|SK).$$

And U checks whether $C_5 \stackrel{?}{=} C'_5$. If it holds, U computes $C_6 = h(W\|SK\|\beta P)$ and transmits C_6 to S_j .

Step 6: S_j computes $C'_6 = h(W\|SK\|\beta P)$ and checks whether $C_6 \stackrel{?}{=} C'_6$. If it holds, S_j confirms the validity of U .

3.4 Password and Biometrics Update Phase

In this phase, U can liberally and securely change the old password PW to a new password PW^{new} and the old biometrics B to a new biometrics B^{new} , respectively, with own ability.

Step 1: $U \rightarrow U$'s smart card : $\{B^{new}\}$

U inserts smart card into a card reader, executes the password update application software, and imprints biometric B^{new} at the sensor.

Step 2: U 's smart card $\rightarrow U$: {Password input request}

U 's smart card compares the imprinted B^{new} with the stored B . If $d(B^{new}, B) \geq \tau$, it means U does not pass the biometric verification, thus the password and biometrics update phase is terminated. On the contrary, if $d(B^{new}, B) < \tau$, it means U passes the biometrics verification and then U 's smart card sends a password input request message to the user U .

Step 3: $U \rightarrow U$'s smart card : $\{PW, PW^{new}\}$

U enters her old password PW and inputs the new password PW^{new} .

Step 4: U 's token computes new $Z^{new} = Z \oplus h(PW||B) \oplus h(PW^{new}, B^{new})$, and then replaces the old Z and B with Z^{new} and B^{new} , respectively, on the smart card.

4 Cryptanalysis of Yoon *et al.*'s Scheme

This section explains steps for attack on Yoon *et al.*'s multi-server authentication with key agreement scheme. The attack has two assumptions about adversary A as follows:

1. Adversary A has ability to intercept any message between U and S_j during the communication, and
2. Adversary A may steal U 's smart card and extracts information stored on it.

In this attack we perform off-line password guessing attack. Before performing attack, we describe off-line password guessing attack. The concept of off-line password guessing attack is described as follows.

Off-line password guessing attack: An attacker guesses a password and verified his guess off-line. No participation of Server S is required, so S don't notice the attack. If his guess fails the attacker tries again with another password, until he finds the proper one.

In section 4.1, we actually perform off-line password guessing attack on Yoon *et al.*'s scheme.

4.1 Off-Line Password Guessing Attack

Suppose that an adversary A steal U 's smart card. In the beginning of authenticated key exchange phase, Adversary A intercepts a message $\langle ID, \alpha P, C_j \rangle$ sent to S_j by U .

Now A can perform off-line password guessing attack by performing the following steps.

Step 1: A extracts Z, B from smart card.

Step 2: A guesses a random password PW' .

Step 3: Using biometric information B , A computes $E = Z \oplus h(PW' || B)$.

Step 4: Using E and intercepted αP , A computes $C' = h(E || \alpha P)$.

Step 5: If $C_j = C'$, A finds the PW successfully. Otherwise, A starts over with another password.

The overview of Off-line password guessing attack for Yoon *et al.*'s scheme is described in Fig. 2.

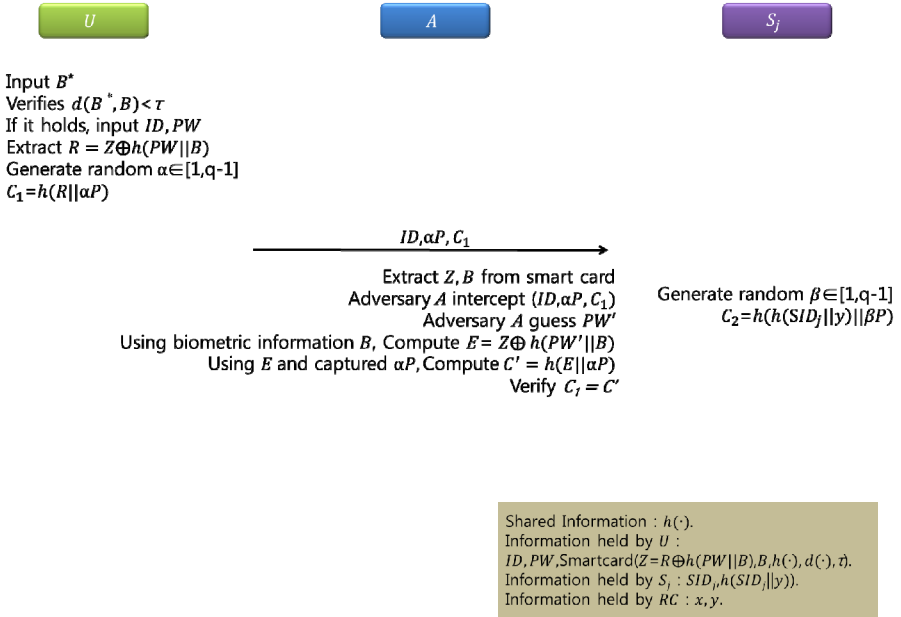


Fig. 2. Off-line password guessing attack on Yoon *et al.*'s scheme

5 Improved Scheme

This section propose improved scheme which is more secure than Yoon *et al.*'s scheme. Before introducing improved scheme, we describe advanced point of improved scheme. In Yoon *et al.*'s scheme, an adversary A can launch off-line password guessing attack because the U's biometric data B is stored to the smart card without concealment and can be obtained by A. A simple way to prevent this attack is to protect B by storing h(B) instead of B. However, due to the nature of hash function, it outputs completely different hash values even if the input biometrics is very close to each other. Therefore, we need more sophisticated method to store h(B) instead of B, such as [3]. For the improved scheme, B should be obtained as follows: Firstly, extract feature vector FV of a user's biometric data. Secondly, apply the one-way transformation (e.g., various Gaussian functions) and quantization to FV. Lastly, concatenate values obtained after quantization yielding B (For more details, see [3]). For simplicity, we define a function h'(·) as the combination of the one-way transformation and the secure hash functions.

From now on, we describe each steps of improved scheme. The scheme is composed of four phases like Yoon *et al.*'s scheme; the server registration phase, the user registration phase, the authenticated key agreement phase, and the password and biometrics update phase. Notations used in improved scheme are defined as follows.

- U, Sj : the user and the jth server, respectively.
- RC : the registration center.

- ID, PW, B : U 's identity, password and biometric template, respectively.
- SID_j : S_j 's identity.
- x : U 's secret key maintained by the registration center.
- y : S_j 's secret key maintained by the registration center.
- p : large prime number.
- F_p : finite prime field.
- E : non-super singular Elliptic curve over a finite field F_p , where $E : y^2 = (x^3 + ax + b) \bmod p$ with $a, b \in F_p$ satisfying $(4a^3 + 27b) \bmod p \neq 0$.
- $E(F_p)$: additive group of points on E over a finite field F_p , where $E(F_p) = \{(x,y) : x, y \in F_p \text{ satisfy } y^2 = x^3 + ax + b\} \cup \{O\}$.
- P : generating element (point) of $E(F_p)$ under consideration F_p .
- α, β : session-independent random integer numbers $\in [1, p-1]$ chosen by U and S_j , respectively.
- SK : shared fresh session key computed by U and S_j .
- $h(\cdot)$: secure one-way hash function.
- $h'(\cdot)$: combination of the one-way transformation and the secure one-way hash function.
- \oplus : bit-wise exclusive-or(XOR) operation.
- \parallel : concatenation operation.
- $A \rightarrow B: M$: A sends a message M to B .

5.1 Server Registration Phase

The following steps are performed during the server registration phase.

Step 1: $S_j \rightarrow RC : SID_j$

S_j chooses its identity SID_j and transmits it to RC via a secure channel.

Step 2: $RC \rightarrow S_j : h(SID_j \parallel y)$

RC computes $h(SID_j \parallel y)$, where y is the S_j 's secret key maintained by RC , and transmits it to S_j via a secure channel.

5.2 User Registration Phase

The following steps are performed during the user registration phase.

Step 1: $U \rightarrow RC : \{ID, h(PW \parallel B), B\}$

U selects ID and PW , and imprints her biometric information B at the sensor. Then U submits $\{ID, h(PW \parallel B), B\}$ to RC . These private data must be sent in a secure channel.

Step 2: $RC \rightarrow U : \{\text{Smart card containing } (Z, h'(B), h(\cdot), h'(\cdot))\}$

RC computes the user authentication key $R = h(ID \parallel x)$ and $Z = R \oplus h(PW \parallel B)$.

Then, RC stores $\{Z, h'(B), h(\cdot), h'(\cdot)\}$ to the memory of U 's smart card and issues it to U via a secure channel.

5.3 Authenticated Key Agreement Phase

The following steps are performed during the authenticated key agreement phase.

Step 1: $U \rightarrow S_j : \{ID, \alpha P, C_1\}$

U inserts the smart card into a card reader, executes the login application software, and imprints biometric B^* at the sensor. Then smart card compares $h'(B)$ with $h'(B^*)$. If $h'(B) = h'(B^*)$, then it outputs accept message. If $h'(B) \neq h'(B^*)$, then it outputs reject message. If it outputs reject, it means U does not pass the biometric verification thus the authentication process is terminated. On the contrary, if it outputs accept, U enters own password PW , and then the reader extracts R by computing $Z \oplus h(PW \| B^*)$, generates a random integer number $\alpha \in [1, q-1]$, and computes $C_1 = h(R \| \alpha P) = h(h(ID \| x) \| \alpha P)$. Then U sends $ID, \alpha P$ and C_1 to S_j .

Step 2: $S_j \rightarrow RC : \{ID, \alpha P, C_1, SID_j, \beta P, C_2\}$

S_j generates a random number $\beta \in [1, q-1]$ and computes $C_2 = h(h(SID_j \| y) \| \beta P)$. Then S_j transmits $ID, \alpha P, C_1, SID_j, \beta P$ and C_2 to RC .

Step 3: $RC \rightarrow S_j : \{C_3, C_4\}$

RC computes $C'_1 = h(h(ID \| x) \| \alpha P)$ and $C'_2 = h(h(SID_j \| y) \| \beta P)$ and then RC checks whether $C_1 \stackrel{?}{=} C'_1$ and $C_2 \stackrel{?}{=} C'_2$, respectively. If both equations hold, RC computes the follows:

$$V = h(h(SID_j \| y) \| \beta P \| \alpha P),$$

$$W = h(h(ID \| x) \| SID_j \| \alpha P \| \beta P),$$

$$C_3 = V \oplus W, \text{ and}$$

$$C_4 = h(V \| W),$$

where W is used to the ephemeral secret key between U and S_j . Finally, RC transmits C_3 and C_4 to S_j .

Step 4: $S_j \rightarrow U : \{\beta P, C_5\}$

S_j computes $V' = h(h(SID_j \| y) \| \beta P \| \alpha P)$, and extracts the ephemeral secret key W by computing $C_3 \oplus V' = V \oplus W \oplus V' = W$. Then S_j computes $C'_4 = h(V' \| W)$ and checks whether $C_4 \stackrel{?}{=} C'_4$. If it holds, S_j computes the shared session key $SK = \beta(\alpha P) = \alpha \beta P$, and $C_5 = h(ID \| SID_j \| W \| SK)$. Finally, S_j transmits βP and C_5 to U .

Step 5: $U \rightarrow S_j : \{C_6\}$

U computes the ephemeral secret key W , the shared session key SK and C'_5 as follows:

$$W = h(h(R \| SID_j \| y) \| \alpha P \| \beta P),$$

$$SK = \beta(\alpha P) = \alpha \beta P,$$

$$C'_5 = h(ID \| SID_j \| W \| SK).$$

And U checks whether $C_5 \stackrel{?}{=} C'_5$. If it holds, U computes $C_6 = h(W \| SK \| \beta P)$ and transmits C_6 to S_j .

Step 6: S_j computes $C'_6 = h(W \| SK \| \beta P)$ and checks whether $C_6 \stackrel{?}{=} C'_6$. If it holds, S_j confirms the validity of U .

5.4 Password and Biometrics Update Phase

In this phase, U can liberally and securely change the old password PW to a new password PW^{new} and the old biometrics B to a new biometrics B^{new} , respectively, with own ability.

Step 1: $U \rightarrow U$'s smart card : $\{B^{new}\}$

U inserts smart card into a card reader, executes the password update application software, and imprints biometric B^{new} at the sensor.

Step 2: U 's smart card $\rightarrow U$: {Password input request}

U 's smart card compares $h'(B^{new})$ which is hashed value from imprinted B^{new} with the stored $h'(B)$. If $h'(B) = h'(B^*)$, then it outputs accept message. If $h'(B) \neq h'(B^*)$, then it outputs reject message. If it outputs reject, it means U does not pass the biometric verification thus the authentication process is terminated. On the contrary, if it outputs accept, U 's smart card sends a password input request message to the user U .

Step 3: $U \rightarrow U$'s smart card : $\{PW, PW^{new}\}$

U enters her old password PW and inputs the new password PW^{new} .

Step 4: U 's token computes new $Z^{new} = Z \oplus h(PW||B) \oplus h(PW^{new}, B^{new})$, and then replaces the old Z and $h'(B)$ with Z^{new} and $h'(B^{new})$, respectively, on the smart card.

6 Security Analysis

In this section, we provide the security analysis between the improved scheme and the previous scheme. Before describing security analysis, we need to define the security terms to conduct an analysis.

1. The Elliptic Curve Discrete Logarithm Problem (ECDLP) is described as follows: Given a public key point $Q = \alpha P$, it's hard to compute secret key α .
2. The Elliptic Curve Diffie-Hellman Problem (ECDHP) is described as follows: Given point elements αP and βP , it's hard to find $\alpha\beta P$.

6.1 Guessing Attacks

In the improved scheme, the undetectable on-line guessing attack is no longer valid against our improved scheme, since after Step 3 in authenticated key agreement phase, RC can authenticate user U . The off-line guessing attack is no longer valid against our improved scheme as well. Because the password PW and the biometric information B are only used for protecting the corresponding smart card. The secret value $Z = R \oplus h(PW||B)$ is stored in U 's smart card. Only legal user U has its own password PW and biometric information B can extract the user authentication key R by computing $Z \oplus h(PW||B)$ and then use its own smart card. As a result, the improved scheme can resist guessing attacks.

For more detail description about prevention of improved scheme about Off-line password guessing attack as follows :

Off-line password guessing attack perform as following steps.

Step 1: A extracts Z, B from smart card.

Step 2: A guesses a random password PW' .

Step 3: Using biometric information B , A computes $E = Z \oplus h(PW' \| B)$.

Step 4: Using E and intercepted αP , A computes $C' = h(E \| \alpha P)$.

Step 5: If $C_j = C'$, A finds the PW successfully. Otherwise, A starts over with another password.

For success of Off-line password guessing attack, attacker should obtain value $R (=h(ID \| x))$. To obtain value R , attacker compute $h(PW' \| B)$ and XOR with value Z .

In attack procedure, value E and value R are exact same if attacker guess correct password. Therefore, attack can be performed. However attack that described above have important assumption which is attacker extracts biometric information(B) from smart card. In improved scheme, on the other hand, smart card store hashed biometric information($h'(B)$). Therefore, attacker is impossible to compute $h(PW' \| B)$ and obtain R . As a result, Off-line password guessing attack cannot be performed in improved scheme.

6.2 Replay Attacks

In the improved scheme, the replay attack is no longer valid against our improved scheme. Because the newness of the messages transmitted in the authentication phase is provided by the session key SK and the random numbers $\alpha P, \beta P$. Only U and S_j , who can gain the session key SK and the shared ephemeral secret key W , can embed the SK and W in the hashed messages C_5, C_6 in authenticated key agreement phase Step 4 and Step 5. As a result, the improved scheme can resist replay attack.

6.3 Stolen-Verifier Attacks

In the improved scheme, S_j and the RC do not store any verification table. Therefore, Stolen-verifier attack is no longer valid against our improved scheme.

6.4 Stolen Smart Card Attacks

In the improved scheme, the stolen smart card attack is no longer valid against our improved scheme. Because Hashed value of User's biometric information is stored in smart card instead of raw biometric information. As a result, attacker cannot extract User's biometric information.

6.5 Insider Attacks

In the improved scheme, the insider attack is no longer valid against our improved scheme. Because User U registers to the RC by presenting $h(PW \| B)$ instead of PW , and during the registration phase we use one-way hash function, insider of RC can't obtain PW .

6.6 Server Spoofing Attacks

In the improved scheme, the server spoofing attack is no longer valid against our improved scheme. Because none of the servers store any user authentication key $h(ID\|x)$ in it, none of the servers can authenticate U . When S_j wants to authenticate, S_j must be authenticated by the RC first, and then obtain the user authentication key $h(ID\|x)$ from RC . When an attacker wishes to deceive RC , they must have $h(SID_j\|y)$.

6.7 Registration Center Spoofing Attacks

In the improved scheme, the registration center spoofing attack is no longer valid against our improved scheme. Because every server S_j has a $h(SID_j\|y)$. S_j can use $h(SID_j\|y)$ to prove the identity of RC .

6.8 Impersonation Attacks

In the improved scheme, the impersonation attack is no longer valid against our improved scheme. Basically, all authentication messages between S_j and U are protected by $h(ID\|x)$. Therefore, attacker have to get $h(ID\|x)$ to generate authentication messages. To avoid S_j get $h(ID\|x)$, RC sends the ephemeral secret key $W = h(h(ID\|x)\|SID_j\|\alpha P\|\beta P)$ to S_j . The ephemeral secret key W is generated by the random point elements αP and βP , so this key is different in each authentication process. As a result, the improved scheme can resist impersonation attack.

6.9 Mutual Authentication

In the improved scheme, the goal of mutual authentication is to generate an agreed session key SK between U and S_j for i th session. In Step 3 of authenticated key agreement phase, after RC receiving the message ID , αP , C_1 , SID_j , βP , and C_2 from S_j , he/she will check if two hash values C_1 and C_2 are match. Because of each random nonce NC and NR are hashed with the user authentication key $h(ID\|x)$ shared between U and RC and the server authentication key $h(SID_j\|y)$ shared between S_j and RC , respectively, RC will believe the i th random points αP and βP was originally sent from U and S_j , respectively. In Step 4 of authenticated key agreement phase, after S_j receiving the message C_3 and C_4 from RC , he/she will check if the hash value C_4 is correct. Since the hashed message included the shared secret value $h(SID_j\|y)$, S_j will believe C_3 and C_4 was originally sent from RC . In Step 5 of authenticated key agreement phase, after U receiving the message βP and C_5 from S_j , he/she will check if the hash value C_5 is correct. Since the hashed message included $h(ID\|x)$ and SK , U will believe βP and C_5 was originally sent from S_j . In Step 6 of authenticated key agreement phase, after S_j receiving the message C_6 from U , he/she will check if the hash value C_6 is correct. Since the hashed message included W and SK , U will believe C_6 was originally sent from U . As a result, the improved scheme provides the mutual authentication.

6.10 Session Key Security

A session key SK is generated from $W = h(h(ID\|x)\|SID_j\|\alpha P\|\beta P)$, αP and βP . These parameter values are different in each session, and each is only known by S_j and U . Whenever the communication closes between U and S_j , the key will immediately self-destruct and will not be reused. When U re-enters the system, a brand new session key will be generated for encrypting all the messages between S_j and RC . Therefore, assuming the attacker has obtained a session key, U cannot use this session key to decode the information in other communication processes. Because the random point elements βP and αP are both generated randomly and are protected by the ECDLP (Elliptic Curve Discrete Logarithm Problem), ECDHP (Elliptic Curve Diffie-Hellman Problem), and the secure one-way hash function, a known session key is unable to be used to calculate the value of the next session key.

In addition, since the values α and β of the random point elements are very large, attackers are unable to directly guess the values α and β of the random point elements to generate session key SK . As a result, the improved scheme provides session key security.

6.11 Security of Ephemeral Secret Key

The ephemeral secret key $W = h(h(ID\|x)\|SID_j\|\alpha P\|\beta P)$ is used to assist S_j authenticate U . When U re-enters the system, the ephemeral secret key W is re-generated during the authentication process. Therefore, assuming the attacker has obtained an ephemeral secret key W by cracking S_j , an attacker cannot use this key to authenticate U in other authentication processes. Because the random point element αP is generated randomly and very large, it is impossible to use a known ephemeral secret key W to calculate the value of the next ephemeral secret key. As a result, the improved scheme provides security of the ephemeral secret key.

6.12 Security of Known-Key

Known-key security means that each execution of an authentication and key agreement protocol between two communication entities (the client and the server) ought to produce unique secret keys; such keys are called session keys. In the improved scheme, knowing a session key $SK = \alpha\beta P$ and the random point elements α and β is useless for computing the other session keys $SK = \alpha\beta P$, because without knowing α and β it is impossible to compute the session key SK . As a result, the improved scheme provides known-key security.

6.13 Perfect Forward Secrecy

In the improved scheme, a disclosed long-lived secret keys x and y including password PW and B cannot derive the session key $SK = \alpha\beta P$ used before because without getting the used random integers α and β , nobody can compute the used session key SK . If an attacker wiretaps all conversations of the medium and derives some used random point

elements αP and βP , he/she could not compute the used session key SK . This problem is the Elliptic Curve Diffie–Hellman key exchange algorithm based on ECDLP (Elliptic Curve Discrete Logarithm Problem) and ECDHP (Elliptic Curve Diffie–Hellman Problem). As a result, the improved scheme provides perfect forward secrecy.

6.14 Secure Password and Biometrics Update Protocol

In the improved scheme, every user can select their own password whenever they want. Therefore, the user can remember the password easily. Moreover, a password update protocol for users to change their passwords and biometric information is provided. It is impossible for a user to change a password and biometric information off-line when the system can resolve the smart card lost problem. If the off-line changing of passwords and the biometric information stored in the smart card is compromised, any attacker may easily guess the password and change the password if he/she wangles a smart card. Moreover, the proposed password and biometrics update protocol allows explicitly a pre-password and biometric information check, where the smart card tests if the old password and biometric information was contained within the smart card by checking the correctness of $h'(B)$. As a result, the improved scheme provides a secure password update protocol.

We compare security properties between related scheme and improved scheme in next page (see Table 1.). As you can see in Table 1., our improved scheme provides prevention of guessing attack that Yoon *et al.*'s scheme doesn't provide. Finally, the improved scheme provides enhanced security.

Table 1. Security properties comparisons between the improved scheme and related schemes

<i>Security properties</i>	<i>Improved scheme</i>	<i>Yoon et al.'s scheme</i>	<i>Tsai's scheme</i>
No verification table	Yes	Yes	Yes
Single registration	Yes	Yes	Yes
Prevention of guessing attack	Yes	No	Yes
Prevention of replay attack	Yes	Yes	Yes
Prevention of stolen-verifier attack	Yes	Yes	Yes
Prevention of stolen smart card attack	Yes	Yes	Yes
Prevention of insider attack	Yes	Yes	No
Prevention of server spoofing attack	Yes	Yes	No
Prevention of RC spoofing attack	Yes	Yes	Yes
Prevention of impersonation attack	Yes	Yes	Yes
Mutual authentication	Yes	Yes	Yes
Session key security	Yes	Yes	Yes
Known-key security	Yes	Yes	Yes
User friendly	Yes	Yes	Yes
Dynamic user authentication key	Yes	Yes	Yes
Providing of perfect forward secrecy	Yes	Yes	N/A
Providing of secure password update	Yes	Yes	N/A
Providing of biometrics authentication	Yes	Yes	N/A

* N/A : Not Available

7 Conclusion

In this paper, we showed that Yoon *et al.*'s scheme is insecure against off-line password guessing attack. We also presented an improved scheme that can withstand such attack.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0026023).

References

1. Lamport, L.: Password authentication with insecure communication. *Communication of ACM* 24, 28–30 (1981)
2. Yoon, E.J., Yoo, K.Y.: Robust biometrics-based multi-server authentication with key agreement scheme for smart cards on elliptic curve cryptosystem. *Journal of Supercomputing* (2010), doi:10.1007/s11227-010-0512-1
3. Sutcu, Y., Sencar, T., Memon, N.: A secure biometric authentication scheme based on robust hashing. In: *ACM MMSEC Workshop*, pp. 111–116 (2005)
4. Leung, K.C., Cheng, L.M., Fong, A.S., Chang, C.K.: Cryptanalysis of a modified remote user authentication scheme using smart cards. *IEEE Trans. Consum. Electron* 49(4), 1243–1245 (2003)
5. Li, L., Lin, I., Hwang, M.: A remote password authentication scheme for multi-server architecture using neural networks. *IEEE Trans. Neural Netw.* 12(6), 1498–1504 (2001)
6. Fan, L., Xu, C.X., Li, J.H.: User authentication scheme using smart cards for multi-server environments. *Chinese Journal of Electronics* 13(1), 179–181 (2004)
7. Hwang, R.-J., Shiau, S.-H.: Password authenticated key agreement protocol for multi-servers architecture In: *International Conference on Wireless Networks Communications and Mobile Computing*, pp. 279–284 (2005)
8. Chang, C.-C., Kuo, J.-Y.: An efficient multi-server password authenticated key agreement scheme using smart cards with access control. In: *Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA 2005)*, vol. 2, pp. 257–260 (2005)
9. Cao, Z.-F., Sun, D.-Z.: Cryptanalysis and improvement of user authentication scheme using smart cards for multi-server environments. In: *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, pp. 2818–2822 (2006)
10. Hu, L., Niu, X., Yang, Y.: An efficient multi-server password authenticated key agreement scheme using smart cards. In: *International Conference on Multimedia and Ubiquitous Engineering (MUE 2007)*, pp. 903–907 (2007)
11. Lee, Y., Won, D.: Security weaknesses in Chang and Wu's key agreement protocol for a multi-server environment. In: *IEEE International Conference on e-Business Engineering*, pp. 304–308 (2008)
12. Geng, J., Zhang, L.: A dynamic ID-based user authentication and key agreement scheme for multi-server environment using bilinear pairings. In: *Workshop on Power Electronics and Intelligent Transportation System*, pp. 33–37 (2008)

13. Lim, M.-H., Lee, S., Lee, H.: An efficient multi-server password authenticated key agreement scheme revisited. In: Third International Conference on Convergence and Hybrid Information Technology, pp. 396–400 (2008)
14. Liao, Y.-P., Wang, S.-S.: A secure dynamic ID based remote user authentication scheme for multi-server environment. *Computer Standards & Interfaces* 31, 24–29 (2009)
15. Chen, Y., Huang, C.-H., Chou, J.-S.: A novel multi-server authentication protocol. *Cryptology ePrint Archive* (2009), <http://eprint.iacr.org/2009/176>
16. Zhu, H., Liu, T., Liu, J.: Robust and simple multi-server authentication protocol without verification. In: Ninth International Conference on Hybrid Intelligent Systems, pp. 51–56 (2009)
17. Yoon, E.-J., Yoo, K.-Y.: Robust multi-server authentication scheme. In: Sixth IFIP International Conference on Network and Parallel Computing, pp. 197–203 (2009)
18. Tsauro, W.J., Wu, C.C., Lee, W.B.: A smart card-based remote scheme for password authentication in multi-server Internet services. *Computer Standards & Interfaces* 27, 39–51 (2004)
19. Juang, W.-S.: Efficient multi-server password authenticated key agreement using smart cards. *IEEE Transactions on Consumer Electronics* 50(1), 251–255 (2004)
20. Chang, C.C., Lee, J.S.: An efficient and secure multi-server password authentication scheme using smart cards. In: International Conference on Cyber worlds (CW 2004), pp. 417–422 (2004)
21. Lee, J.H., Lee, D.H.: Efficient and secure remote authenticated key agreement scheme for multi-server using mobile equipment. In: Proceedings of International Conference on Consumer Electronics, pp. 1–2 (2008)
22. Tsai, J.L.: Efficient multi-server authentication scheme based on one-way hash function without verification table. *Computers & Security* 27(3–4), 115–121 (2008)
23. Chen, J., Yang, Y.: Temporal dependency based checkpoint selection for dynamic verification of temporal constraints in scientific workflow systems. *ACM Trans. Softw. Eng. Methodol* (June 17, 2009), <http://www.swinflow.org/papers/TOSEM.pdf> (in press, accepted)
24. Wang, M., Kotagiri, R., Chen, J.: Trust-based robust scheduling and runtime adaptation of scientific workflow. *Concurr. Comput. Pract. Exp.* 21(16), 1982–1998 (2009)
25. Chen, J., Yang, Y.: Activity completion duration based checkpoint selection for dynamic verification of temporal constraints in grid workflow systems. *Int. J. High Perform. Comput. Appl.* 22(3), 319–329 (2008)
26. Nam, J., Kim, S., Won, D.H.: Secure Group Communications over Combined Wired and Wireless Networks. In: Katsikas, S.K., López, J., Pernul, G. (eds.) *TrustBus 2005*. LNCS, vol. 3592, pp. 90–99. Springer, Heidelberg (2005)
27. Lee, K., Won, D., Kim, S.: A Secure and Efficient E-Will System Based on PKI. *Information - An International Interdisciplinary Journal*, *International Information Institute* 14(7), 2187–2206 (2011)
28. Park, N., Kim, S., Won, D.H., Kim, H.W.: Security Analysis and Implementation Leveraging Globally Networked RFIDs. In: Cuenca, P., Orozco-Barbosa, L. (eds.) *PWC 2006*. LNCS, vol. 4217, pp. 494–505. Springer, Heidelberg (2006)

Rate-Distortion Optimized Transcoder Selection for Multimedia Transmission in Heterogeneous Networks

Haroon Raja and Saad Bin Qaisar

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology, Islamabad, Pakistan

Abstract. In this paper we propose a solution for selection of appropriate transcoding nodes in a network operating in ad-hoc mode. The heterogeneity present in today's networked devices necessitates different quality of video for different end users. One possible solution for this heterogeneity is to transcode the video stream as per user demand. In this work, we define significant parameters to facilitate decision on selection of transcoding nodes within a wireless access network. We formulate the problem as a rate-distortion optimization to achieve conflicting objectives of high quality and minimum time of delivery to an end user. Unlike past works which have focused on transcoding to develop efficient distributed transcoders, our aim is to come up with methods for placement of these parallel transcoding nodes in a heterogeneous network, keeping in view the constraints of timely delivery of video and minimal distortion.

1 Introduction

Advancement in communication technologies over last few years has led to the development of number of access mechanisms like 802.11 standard for wireless access, Ethernet, 802.15.4 etc. Most of the devices like laptops, PDAs etc. manufactured today have the capability to support multiple access technologies. These advancements move us even closer to the future promised by pervasive/ubiquitous computing.

The heterogeneity in present networks puts demands on encoders to supply with data that can be decoded at all the nodes. This objective of providing data that is decodable by all the nodes has been eluding researches for some time now. In case of lack of ability of such an encoder or due to weakness at the decoder end one possible solution is to transcode the data at some intermediate location. The transcoding is a complex operation as it involves decoding and re-encoding of data. According to [8], time taken for transcoding 1 second of video stream is in range of 5 to 12 sec for different coding rates. One possible solution for transcoding in real-time applications is to perform it in a distributed manner. Works in [9], [5], [11] and [13] focused on distributed transcoding for peer-to-peer networks. In [5] and [13], focus is more on defining overlay protocols

for supporting transcoding. In contrast, [9] and [11] pay more attention towards affect of using more than one transcoder, in parallel. According to their findings, by increasing the number of transcoders, we can reduce transcoding time but with a tradeoff that as we go on increasing number of transcoders, distortion of received video keeps on increasing due to lack of information about previous frames that are being transcoded by other transcoder.

In peer-to-peer and ad-hoc networking paradigms, users collaborate to fulfill the data rate demands of other users present in the network, similarly the concepts of grid and mobile computing help users to distribute computation load among other peers. In these cases, as the load on infrastructure is reduced, it results in reduction of cost by means of decrease in infrastructure expansion and maintenance costs. One specific example of using peer-to-peer networking for IPTV and VOD is [8]. In [8], hybrid network (mixture of ad-hoc and infrastructure mode) is developed. To reduce the load on infrastructure, data is temporarily stored on intermediate nodes (the equipment present at user premises) and the central control room builds a peer-to-peer network among devices for VOD delivery from these intermediate nodes. In case of [8], the targeted users were assumed to have equal capabilities and were connected via wired link. The method proposed in this work is the one possible extension to the system proposed in [8] where introduction of low capability devices is handled using distributed transcoding and best set for transcoding operation is selected.

Our work is inspired from ad-hoc networking and mobile computing and aim is to provide enabling technologies for video streaming services to client nodes in a heterogeneous network environment. The problem is solved at access network level and the proposed solution can be helpful to gain new insights into using core wireless networking concepts for further advancement of pervasive computing concepts. This work also gives useful insights for developing practical solutions for transmitting multimedia to the users (as an example of [8] is already mentioned).

A variety of current day wireless networks provide support for multiple data rates at PHY layer. To derive full advantage of these rates, a good link adaptation algorithm is needed. Cross-layer design has also been employed to the good effect for transmission of data belonging to different applications having different QoS demands. Cross-layer based algorithms adjust different parameters across the protocol stack to achieve finest point of operation depending on the specific user demands. In this work, we focus on finding suitable parameters at application, link (MAC) and physical (PHY) layer for enhanced quality of video streaming to nodes having different decoder requirements than the other nodes. At application level, the parameter used is number of transcoders, at MAC retry limit is adjusted and at PHY modulation and coding rate are adapted. In this work, a method for selection of transcoders among different options (different nodes capable of transcoding have different link quality resulting in different parameters at PHY, MAC and application layer) present in the network is proposed. Precisely, in the above context, we make following contributions:

1. A rate-distortion function is obtained on the basis of application, MAC and PHY parameters, where rate is defined with the help of channel transmission rate and the retry count and distortion is characterized with the help of packet error rate induced by using these parameters. The formation of RD function is explained with detail in section 3.
2. Once RD function is obtained, an RD optimization is solved which gives us operational point on RD-curve. And by decoupling this point's properties we will get answers to the questions; number of transcoders to use, retry limit at MAC and transmission rate of the selected transcoders.

Rest of our work is organized as follows: In section 2, we provide an overview of the underlying system. We formulate our problem as an optimization problem and presented its solution in section 3. In section 4 results are presented and finally in section 6 paper is concluded.

2 System Overview

2.1 Distortion Due to Packet Loss

We consider a wireless medium between transcoders and the final destination nodes. Wireless channel is error prone and packet loss is contributed by low SNR of link as well as collisions while accessing the channel. Details related to different types of losses and the total loss suffered by transmitted data are explained in the next sections.

In case of video streaming applications, loss of data results in increase in distortion of the received video. [4] and [3], among others, have defined distortion of video as a function of packet losses. In these works, authors have assumed additive increase in distortion as a function of frame loss. If frame f of sequence is lost, it will have affect on future frames as well before next I frame is received.

$$D(f) = \sum_{j=1}^n D(f_j) \quad (1)$$

Where n is the number of frames left before receiving the next I frame. According to [4] and [3], the distortion function depending on transmission policy is given by:

$$D(\psi) = \sum_{f=1}^G D(f) \epsilon(\psi_f) \quad (2)$$

Here $\epsilon(\psi_f)$ is the probability of packet loss when transmission policy ψ is applied. $D(\psi)$ is the expected value of distortion when policy ψ is used and $D(f)$ is the distortion due to loss of frame f in GOP (group of picture) and G is the size of GOP.

In [6], video distortion model has been proposed and again, additive effect due to packet losses has been proposed. [6] provided more detailed analysis and effect of packet loss on distortion depending on frame position in GOP is also included in mathematical model.

We assume additive effect due to packet loss on video distortion. Additive model for distortion might not hold true in some cases, even then it is a good assumption [4].

2.2 System Architecture

In this section, we explain the system architecture under consideration. We assume a network at access level which is operating in an infrastructure mode but end nodes present in the network have the capability to operate in ad-hoc mode or in p2p fashion as well (like provision of mesh topology in WiMAX and capability of developing ad-hoc networks in WLANs).

We assume receiver driven video transmission, the client starts off when mobile/destination station requests access point (AP) for the video stream. AP decides whether transcoding is required. If transcoding is not needed, normal routine is followed i.e. video stream from access point is directly streamed to the ultimate destination. In case transcoding is required, the algorithm proceeds with selection of suitable transcoders.

The next task for an AP is to randomly select a set of potential transcoder nodes J from nodes present in the access network. In this work, it is assumed that all nodes have equal processing resources available. All the nodes in set J transmit probe packets to destination node for computation of received SNR from each node. Value of SNR is used to select modulation and coding rate from each of the potential transcoders to destination node. Aim of the algorithm is to select optimal number of transcoders having best transmission characteristics. The transmission characteristics we are interested in are modulation scheme and the retry limit. These two parameters have direct impact on packet loss and delay of the link.

3 Network Architecture

We consider a wireless access network in this work. The network is characterized by two features, its probability of packet loss and delay experienced by packet within network.

3.1 Probability of Packet Loss

Let γ be SNR of signal at the receiver end and the number of nodes contending for wireless medium be n . The reason that transmitted packet is received in error will be result of either of collision or bad channel. The probability of bit error due to bad channel in case of AWGN channel is given as a function of E_b/N_0 , [12]:

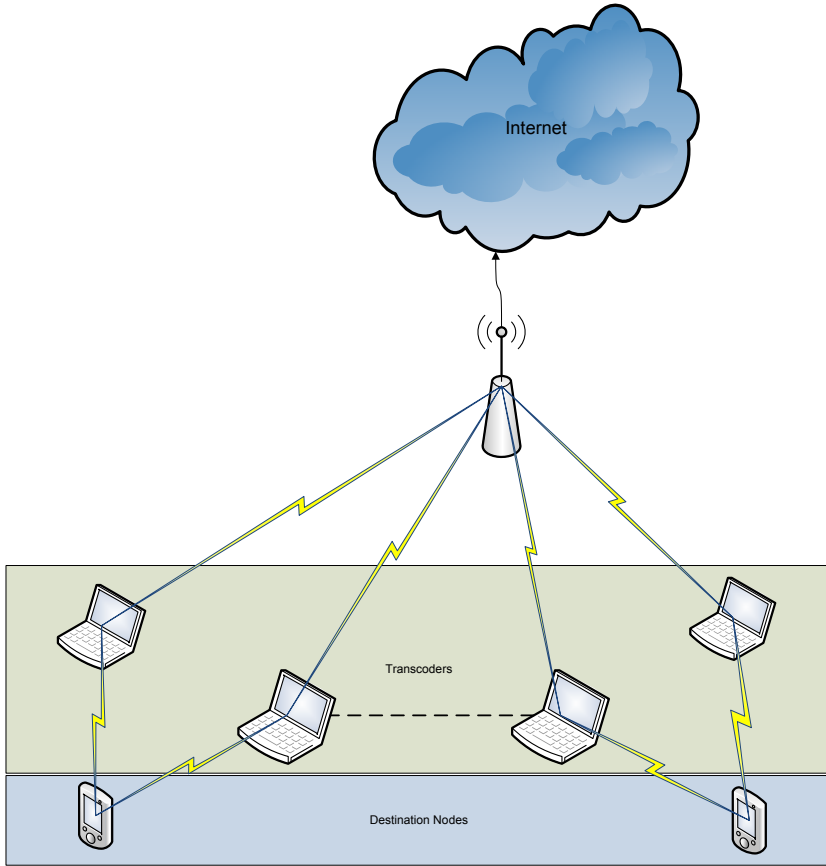


Fig. 1. System Architecture

$$BER = \frac{2(1 - L^{-1})}{\log_2(L)} Q \left(\sqrt{\left(\frac{3 \log_2 L}{L^2 - 1} \right) \frac{2E_b}{N_0}} \right) \quad (3)$$

If packet of length b is transmitted then the probability of receiving a packet in error is given by:

$$P_{err}^{ch} = 1 - (1 - BER)^b \quad (4)$$

Bianchi has developed an analytical model for CSMA-CA protocol over single-hop wireless networks in [2]. According to [2], probability of successfully transmitting a packet when n nodes are contending for a shared medium in CSMA-CA case is given by:

$$P_s = \frac{n\tau(1 - \tau)^{n-1}}{1 - (1 - \tau)^n} \quad (5)$$

Here τ is the probability that station will transmit in any random time slot. Probability that the packet will suffer a collision is given as complement of the probability of successful transmission:

$$P_{err}^{col} = 1 - P_s \quad (6)$$

Overall probability of receiving a packet in error due to channel SNR or collision is given by:

$$P_{err} = 1 - (1 - P_{err}^{col})(1 - P_{err}^{ch}) \quad (7)$$

In case of multimedia streaming applications, packet loss is due to two factors; packet is dropped at the MAC layer as it exceeds its retry limit or a packet is received but it has already passed its delay deadline. Hence the probability of packet loss is:

$$P_{loss} = 1 - (1 - P_{drop})(1 - P_{late}) \quad (8)$$

Here P_{drop} is the probability that packet has been dropped at MAC layer as retry counter exceeds the limit and P_{late} is the probability that packet is received after the deadline. If L is the retry limit, probability that packet is dropped at MAC layer can be presented by following equation:

$$\epsilon = (P_{err})^{L+1} \quad (9)$$

3.2 Delay Constraints

Delay experienced by a video received at the destination node is accumulation over different delay sources. The delay stages consist of transcoding delay at the intermediate node T_{trans} , channel access delay at the MAC layer T_{Acc} and the transmission delay T_{Tx} dependent on channel conditions. For communication of multimedia data, delay deadline must be met. Let $T_{Deadline}$ be the deadline for the transmitted packet, as deadline must be met for packet to be useful the constraint equation is:

$$T_{trans} + T_{Acc} + T_{Tx} \leq T_{Deadline} \quad (10)$$

For retry limit L , if packet can be transmitted L times before it is received successfully the constraint equation is:

$$\sum_{i=1}^L T_{Tx} + T_{Acc_i} + T_{Ttrans} \leq T_{Deadline} \quad (11)$$

Rearranging the equation, we get upper bound on value of retry limit L as a function of number of transcoders N and modulation scheme m .

$$L(N, m) \leq \frac{T_{Deadline} - E[T_{Ttrans(N)}]}{E[T_{Tx(m)} + T_{Acc}]} \quad (12)$$

Let us define $T_N = T_{T_x} + T_{Acc}$ as the total time taken by MAC and PHY for transmission of packet and σ be the slot size. Using MAC model for CSMA proposed in [2], we can write TN as:

$$T_N = (1 - P_{tr})\sigma + P_{tr}P_sT_s + P_{tr}(1 - P_s)T_c \quad (13)$$

Here T_s and T_c are the time taken for transmission of packet successfully and in error, respectively.

4 Transcoder Selection Problem and Solution

In first part of this section we will be developing the transcoder selection problem in terms of rate distortion problem and the solution of problem is also proposed. In second part, the naive approach for selection of transcoding nodes is explained.

4.1 Rate-Distortion Optimized Solution

The problem under consideration is a two tier problem, in first step local optimal points for each frame are computed and in second step distortion rate function is optimized for whole video sequence. Each of these steps is explained in subsections below.

Optimizing Transmission of Single Frame

1. **Error-Cost Relationship** When talking about transmission of a single packet, we can represent distortion by means of normalized distortion or simply, channel error rate and normalized rate (cost) can be used instead of absolute rate as well. This leads towards calculation of error-cost function for a single frame transmission. Let $\epsilon(\psi_f)$ be the transmission cost per frame of a video sequence. We define cost in terms of number of transmissions or the retry limit. Let, the frame f of sequence is transmitted using policy ψ_f and B_f is the size in bytes of the frame f of the video sequence. Then the expected rate $R(\psi)$ will be:

$$R(\psi) = \sum_{f=1}^F B_f \rho(\psi_f) \quad (14)$$

Here ψ is a policy vector and each element of the vector defines the policy for f_{th} frame of the sequence, $\psi[\psi_1, \psi_2, \dots, \psi_F]$. And the policy for each individual frame is a three tuple defined as follows:

$$\psi_f = [m_f, L_f, n_f] \quad (15)$$

In this equation m , L and n are modulation scheme, retry limit and number of transcoders respectively. We have already defined error function ϵ in equation (9) and now cost function has also been defined. The convex hull

of the solution space is achieved through quick hull algorithm [1]. The function of lower convex hull of error-cost is divided into different regions. Here each region will define the number of transcoders used. As cost is increased, number of transcoders will increase as more number of transcoders result in lesser computational complexity resulting in more number of retransmissions. Region containing optimal point selected after solving optimization problem will give us the number of transcoders used. If we assume perfect synchronization between transcoders then we will have no distortion in case of using more than one transcoders, then in this case the regions defining number of transcoders will be increasing as a function of increasing number of retransmissions only.

2. **Solving for Individual Frames** The first step involves the computation of local optima for each frame transmission. According to [6] it is possible to enumerate all the possible scenarios and find the optimal point using exhaustive search, but it is sufficient to find the points on the lower convex hull of error-cost function, once the lower convex hull is obtained we can use Lagrangian to find the optimal point.

$$J_\psi = \epsilon_\psi + \lambda \rho_\psi \quad (16)$$

This equation can be solved by using dynamic programming as in [7], as well as branch and bound algorithm proposed in [10]. In equation (16) ϵ_ψ and ρ_ψ are the expected values of error and cost respectively when policy ψ is used. Optimizing transmission of video sequence The second step towards finding an optimal policy for transmission of a video sequence is to find global minima for distortion i.e. varying local optima's to reach a point having minimum distortion value for the whole video.

$$J' = D + \lambda' R \quad (17)$$

In [10] authors have shown that selection of transmission policy on single QoS driven network is an NP-hard problem which gives way to heuristic based algorithms for policy computation. Hence, after finding an operating point on error-cost function the iterative algorithm is applied to come up with the optimal point on rate-distortion curve. Heuristics based algorithm proposed in [6] has been used to find the optimal policy.

4.2 Naive Approach

We use a naive method for the selection of transcoders which will be used as a reference system to compare performance of proposed system. In this method the access point randomly selects a set of transcoders and video stream is directed towards the destination nodes by passing through these randomly selected transcoding nodes.

5 Results and Discussions

The network used for simulation consists of 802.11g links, the video is transmitted from transcoder to the destination node over one of these links. The background traffic is generated by adding 10 more users to the network each user is generating a CBR traffic at the rate of 1Mbps and packet size of 500 bytes is used for these transmissions. For the selection of transcoders from the potential set the optimization problem is solved offline and the static channel is simulated for the entire video transmission. For testing video transmission Akiyo and Foreman sequences with CIF parameters is used in simulations. The sequences are encoded using H.264 encoder at the frame rate of 30 Hz with GOP size of 30. Results in figure 2 show the decrease in number of frame drops for video sequence when retry limit is increased, in this case the transmission rate is kept constant and the same channel conditions are simulated for each case. Figure 3 shows the PSNR corresponding to frame errors in figure 2. As we can see for the same channel conditions Akiyo sequence has much better reception rate and the video quality, the reason for higher frame loss rate is the increase in source rate for using foreman sequence, as due to motion the compression efficiency is lesser for foreman sequence resulting in more video data, while the service rate of channel is still the same. And the drastic decrease in video quality of foreman sequence can also be attributed to the more motion in foreman sequence as compared to Akiyo, theciteore loss of one frame for foreman will cause more distortion as compared to the Akiyo sequence.

Figures 4 shows the comparison between received video quality using rate-distortion optimized selection of transcoders and the nave method for selection of transcoders. The x-axis shows the frame indices and the quality of received frame in terms of PSNR is shown on y-axis. The results indicate that we can achieve gains in terms of video quality enhancement by selecting transcoders in rate-distortion optimized and channel aware manner as compared to random selection of transcoding nodes by nave method. In case of figure same transmission rate is assigned to both the algorithms, hence the only variable left behind is the retry count parameter at MAC layer, as nave approach selects the parameters randomly theciteore the distribution of retry limit value for nave approach will be uniform and correspondingly the PSNR value achieved is the average over all the values achieved by using different retry limit values. And the overall result indicates that using rate-distortion selection of transcoders has given us significant gains as compared to the random selection of transcoders even when only one parameter (i.e. retry limit) is selected using a systematic approach than a nave one. Figure 5 also shows the comparison between instantaneous quality of received video, for figure 5 foreman sequence is used. For the same channel conditions foreman sequence is received with a marked difference in quality as compared to the Akiyo sequence, but the important point to note is improvement in video quality with the change in parameters and here also rate-distortion optimized streaming outperforms the nave method.

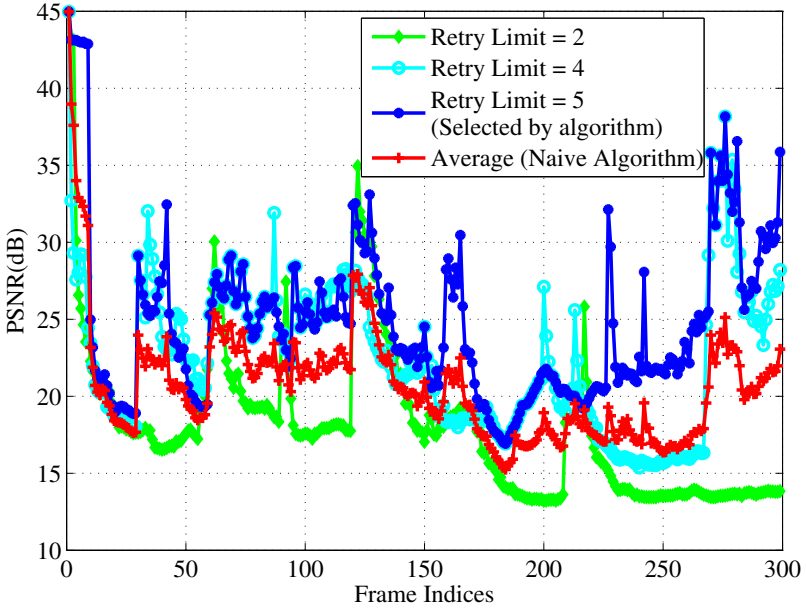
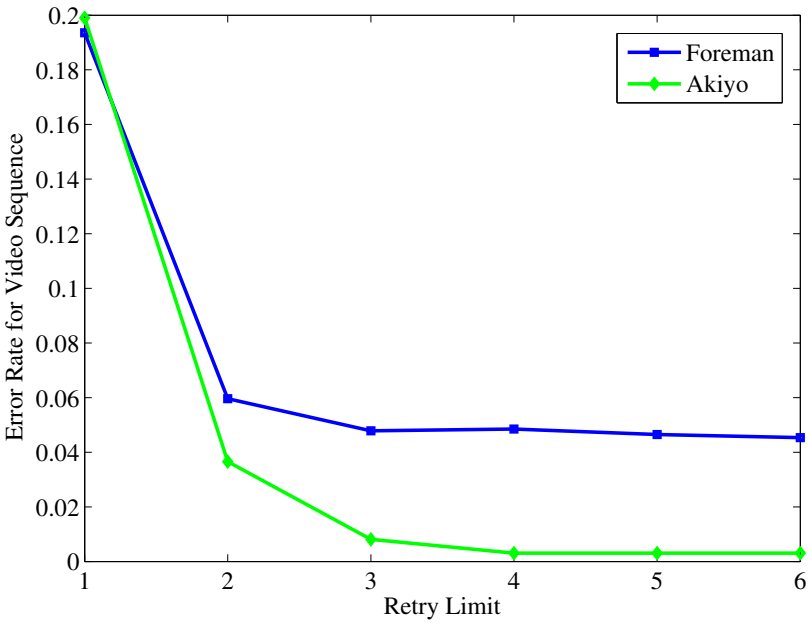
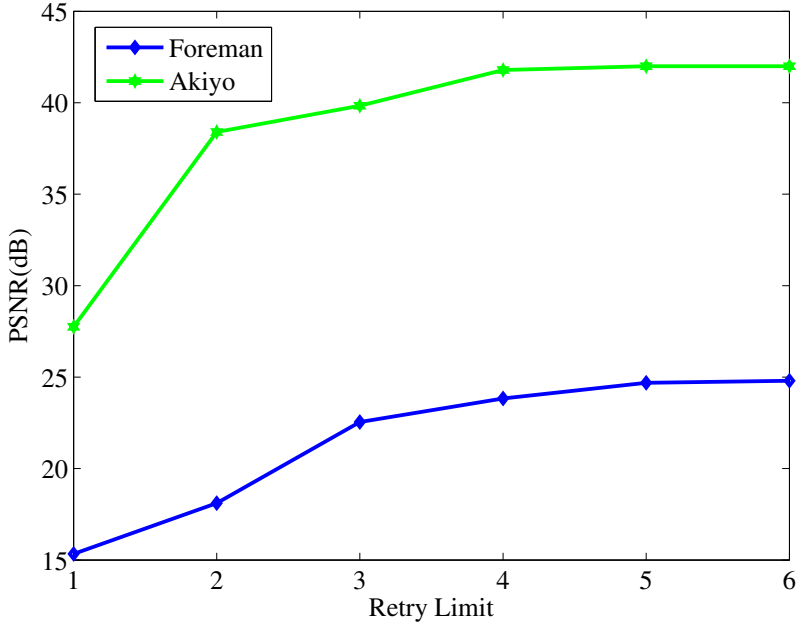


Fig. 2. Variation in video PSNR with variation in time.





6 Conclusions

In this paper, we have proposed a solution for assigning a transcoding task to the nodes present in the network, proposed solution decides on the number of transcoders to be used for operation as well as the transmission characteristics of the selected nodes (retry limit and transmission rate adjustments). The results presented here suggest that there is a need for an algorithm to allocate the transcoders in the network, as the proposed scheme has outperformed naive approach for transcoders selection, hence rendering the naive approach as insufficient.

References

1. Barber, C., Dobkin, D., Huhdanpaa, H.: The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* 22(4), 469–483 (1996)
2. Bianchi, G.: Performance analysis of the ieee 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications* 18(3), 535–547 (2000)
3. Chakareski, J., Apostolopoulos, J., Girod, B.: Low-complexity rate-distortion optimized video streaming. In: *International Conference on Image Processing of ICIP 2004*, vol. 3, pp. 2055–2058. *IEEE* (2004)
4. Chakareski, J., Chou, P.: Radio edge: Rate-distortion optimized proxy-driven streaming from the network edge. *IEEE/ACM Transactions on Networking* 14(6), 1302–1312 (2006)

5. Chen, F., Repantis, T., Kalogeraki, V.: Coordinated media streaming and transcoding in peer-to-peer systems. In: Proceedings of 19th IEEE International Parallel and Distributed Processing Symposium, pp. 56b. IEEE (2005)
6. Choi, L., Ivrlac, M., Steinbach, E., Nossek, J.: Analysis of distortion due to packet loss in streaming video transmission over wireless communication links. In: IEEE International Conference on Image Processing, ICIP 2005, vol. 1, pp. 1–189. IEEE (2005)
7. Chou, P., Miao, Z.: Rate-distortion optimized streaming of packetized media. *IEEE Transactions on Multimedia* 8(2), 390–404 (2006)
8. Gopalakrishnan, V., Bhattacharjee, B., Ramakrishnan, K., Jana, R., Srivastava, D.: Cpm: Adaptive video-on-demand with cooperative peer assists and multicast. In: IEEE INFOCOM 2009, pp. 91–99. IEEE (2009)
9. Noh, J., Makar, M., Girod, B.: Streaming to mobile users in a peer-to-peer network. In: Proceedings of the 5th International ICST Mobile Multimedia Communications Conference, p. 24. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2009)
10. Roder, M., Cardinal, J., Hamzaoui, R.: On the complexity of rate-distortion optimal streaming of packetized media. In: Proceedings of Data Compression Conference, DCC 2004, pp. 192–201. IEEE (2004)
11. Sambe, Y., Watanabe, S., Yu, D., Nakamura, T., Wakamiya, N.: High-speed distributed video transcoding for multiple rates and formats. *IEICE Transactions on Information and Systems E Series D* 88(8), 1923 (2005)
12. Sklar, B.: *Digital communications: fundamentals and applications*. Prentice-Hall, Inc. (1988)
13. Wu, J., Huang, P., Yao, J., Chen, H.: A collaborative transcoding strategy for live broadcasting over peer-to-peer iptv networks. *IEEE Transactions on Circuits and Systems for Video Technology* (99), 1 (2010)

Formal Probabilistic Analysis of Cyber-Physical Transportation Systems

Atif Mashkoor¹ and Osman Hasan²

¹ Software Competence Center Hagenberg,
Hagenberg, Austria
`atif.mashkoor@scch.at`

² School of Electrical Engineering and Computer Science,
National University of Sciences and Technology,
Islamabad, Pakistan
`osman.hasan@seecs.nust.edu.pk`

Abstract. Formal specification and verification of cyber-physical transportation systems is inherently a complex task. A fail-safe specification of such systems not only includes intricate formalizations of assumptions and requirements but also a fine-grained analysis of their unpredictable and random components, at times at different levels of abstraction. Traditional techniques of verification and validation, such as simulation or model checking, do not cope very well with the posed challenges. In fact, sometimes it becomes merely impossible to guarantee certain properties, such as liveness, under all possible scenarios. We propose an approach based on higher-order logic for formal modelling and reasoning of cyber-physical transportation systems. In this approach, we express the unpredictable elements of the model by appropriate random variables. Instead of guaranteeing absolute correctness, these randomized models can then be used to formally reason about the probability or expectation of the system meeting its required specification. For illustration purposes, the paper presents a simple analysis of a vehicle platoon control algorithm.

1 Introduction

Automation is widely being practised nowadays in all modes of transportation, be it aviation, railway or automotive. The automation of such systems involves extensive data acquisitions from their environment, communication mechanisms to interact with the other members of the domain and enormous computations within their real-time embedded components for generating the required control signals. These specifications make them one of the most complex cyber-physical systems to design and verify.

The complexity related to transportation systems coupled with the enormous involvement of continuous and unpredictable physical aspects makes their verification a great challenge. Traditionally, their analysis is done using computer-aided design tools like Matlab. The main idea is to create a software model, capturing all the continuous and unpredictable details of the system, expressed

in terms of Matlab functions. However, computer arithmetic cannot support infinite precision so the continuous physical realities are approximated by their closest discrete counter parts in terms of floating or fixed point numbers. Similarly, true randomness cannot be attained in computers and thus the random physical realities are approximated using pseudo random numbers in the Matlab models of the system. Once the system is modelled, the Matlab functions are analysed using computer simulation techniques, which deduce a property to be true by checking it for a number of test cases. This kind of testing is mainly utilized because exhaustive simulation, or testing for all possible combinations, is not feasible in terms of computational time and complexity for cyber-physical transportation systems. Thus, the simulation based analysis of Matlab models cannot be termed as accurate due to the inherent nature of simulation, and the usage of floating or fixed point numbers and pseudo random number generator based random variables in the Matlab models.

Transportation systems are among the top most operational safety-critical systems in the modern world to date. A faulty transportation system could result in disastrous consequences and may lead to the loss of human lives in worst cases. For example, the 2002 mid-air collision in Überlingen caused due to the human-controller interaction flaw, the 2008 frontal train collision caused due to an error in the warning system, and the 2009 crash of Air France Flight 447 that resulted in killing all 216 passengers and 12 crew members happened due to a flaw in the automatic reporting system, are the historic events that no one would like to be repeated. Therefore, given such a safety-critical nature of cyber-physical transportation systems, inaccurate analysis techniques, like simulation, should not be completely relied upon for their verification.

Formal methods are capable of conducting precise system analysis and thus can overcome the above mentioned limitations of simulation in the context of analysing cyber-physical transportation systems. The main principle behind formal analysis of a system is to construct a computer based mathematical model of the given system and formally verify, within a computer, that this model meets rigorous specifications of intended behaviour using mathematical reasoning.

Two of the most commonly used formal verification methods are model checking [4] and higher-order logic theorem proving [20]. Model checking is an automatic verification approach for systems that can be expressed as a finite-state machine. Higher-order logic theorem proving, on the other hand, is an interactive approach but is more flexible in terms of tackling a variety of systems. Both model checking and theorem proving have been successfully used for the precise functional correctness of a broad range of engineering and scientific systems, including some aspects of cyber-physical transportation systems. These days, certification authorities explicitly demand transportation systems to be safe, dependable and correctly implemented. In this realm, formal approaches to assure system safety, dependability and correctness are finding their way into being used as a certification argument (cf. DO-178C, IEC-61508, SIL-4).

However, in most of the existing work in the formal verification of cyber-physical transportation systems, a high level model of the actual system is

considered such that the continuous and random physical realities are abstracted away. These analysis, even though are accurate due to the inherent soundness of the formal methods yet, cannot be considered as complete as the left out details could lead to potential failures. One of the main reasons for not considering the complete models of the cyber-physical systems is that most of the formal methods based analysis in this domain have been conducted by using either model checking or theorem proving with a decidable logic. These techniques, even though are automatic and thus user friendly yet, are not expressive enough to completely capture the continuous and random physical realities.

To address the above mentioned expressiveness problem, we propose to use higher-order logic theorem proving [16] for analysing cyber-physical transportation systems in this paper. Higher-order logic is a system of deduction with a precise semantics and due to its high expressiveness it can be used to model any system that can be expressed in a closed mathematical form. Higher-order logic has also been successfully used to develop some of the classical mathematical theories. Interactive theorem proving is the field of computer science and mathematical logic concerned with computer based formal proof tools that require some sort of human assistance.

The core of theorem provers usually consists of a handful of axioms and primitive inference rules. Soundness is assured as every newly added construct must adhere them. Powerful mathematical techniques such as induction and abstraction are the strengths of theorem proving and make it a very flexible verification technique. These distinguishing characteristics make higher-order logic theorem proving a very flexible verification technique that can be used to analyse any system with all of its continuous and physical details. For example, higher-order logic theorem proving has been used to successfully analyse continuous systems, such as optical waveguides [22], fractional order systems [35], real-time systems such as Stop-and-Wait protocol [7], systems with unpredictable and random elements such as reconfigurable memory arrays in the presence of stuck-at and coupling faults [23], and the wireless sensor network scheduling [12].

We propose to leverage upon the expressiveness of higher-order logic theorem proving to analyse cyber-physical transportation systems in this paper. In particular, the paper describes a framework for analysing their randomized aspects. The main idea is to model the randomness found in such systems in terms of formalized random variables in the higher-order logic model. These models can then be used to formally reason about the interesting probabilistic and statistical characteristics of the given system within the sound core of a theorem prover. Due to the undecidable nature of the underlying logic, the proofs would involve interaction, a cost that we pay to attain accurate probabilistic analysis results. However, these added efforts are justifiable given the safety-critical nature of transportation systems. Moreover, from past experiences, we expect the formal reasoning efforts to decrease with the availability of higher-order logic formalizations of cyber-physical systems as these available formal models and theorems can be re-utilized to formalize other variants of the same domain.

We have utilized HOL [21], a higher-order logic theorem prover for our work. Main reasons of its utilization are the availability of the underlying mathematical theories of probability and a possibility of a real analysis using its libraries.

The rest of the paper is organized as follows: After reviewing the related work in Section 2, we proceed by presenting some preliminaries including a brief introduction to higher-order logic theorem proving and the HOL theorem prover in Section 3. Next, Section 4 describes the proposed theorem proving based probabilistic analysis approach for the cyber-physical systems. This is followed by the simple analysis of a vehicle platoon control algorithm in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

Due to inaccuracies introduced by the simulation based analysis methods, many researchers around the world are exploring the usage of formal methods for probabilistic analysis. Generally, probabilistic model checking is employed to assess the quantitative aspects of systems' safety and reliability. For example, probabilistic model checker PRISM [29] has been quite frequently used to evaluate the dependability and safety features of various systems (e.g., [28,15]). Probabilistic model checking involves the construction of a precise state-based mathematical model of the given probabilistic system. It is then subject to exhaustive analysis to formally verify that it satisfies a set of formally represented probabilistic properties. However, it can be used to analyse systems that can be expressed as probabilistic finite state machines only. Another major limitation of the probabilistic model checking approach is state space explosion. The state space of a probabilistic system can be very large, or sometimes even infinite. Thus, at the outset, it is impossible to explore the entire state space with limited resources of time and memory. Similarly, we cannot reason about mathematical expressions in probabilistic model checking. Thus, probabilistic model checking cannot be used to formally verify probability distributions or statistical characteristics, which are widely used parameters to assess the probabilistic correctness of a property.

A statistical model checker has been recently utilized to analyze some aspects of cyber-physical systems [9]. However, this approach also suffers from the classical model checking issues, like the state-space explosion and inability to reason about mathematical relations. Thus, the probabilistic model checking approach, even though is capable of providing exact solutions, is quite limited in terms of handling a variety of probabilistic analysis problems. Whereas higher-order logic theorem proving is capable of overcoming all the above mentioned problems, though at the cost of significant user interaction.

Formal methods, specifically B [1] and Event-B [2], have been extensively used in the development of transportation systems (e.g. [5,3]). Recently, these methods have been extended to allow the modelling and verification of probabilistic features. These extensions primarily use a probabilistic choice operator to probabilistically reason about certain termination conditions [17] and provide

semantics of a Markov process to reason about some reliability issues [36]. However, these initiatives have a very limited scope. For example, they cannot be used to reason about generic mathematical expressions for probabilistic or statistical properties. Similarly, such formalisms are not mature enough to model and reason about all different kinds of continuous probability distributions. Given the continuous random nature of the transportation systems, both the probabilistic model checking and Event-B based techniques cannot be used to capture their behaviour and thus, the use of probabilistic and quantitative assessment for the formal verification of transportation systems is still few and far between.

The foremost criteria for conducting the formal probabilistic analysis in a theorem prover is to be able to express probabilistic notions, such as probability of an event and random variables, in higher-order logic and reason about the probability distribution and statistical properties of random variables in a higher-order logic theorem prover. A formalized probability theory provides the foundations for expressing probabilistic notions.

A number of authors, including Hurd [27], Mhamdi [32] and Hölze [26], reported higher-order logic based formalizations of probability theory. The recent works by Mhamdi [32] and Hölzl [26] are based on extended real numbers (including $\pm\infty$) and provide the formalization Lebesgue integral for reasoning about advanced statistical properties. This way, they are more mature than Hurd's [27] formalization of measure and probability theories, which is based on simple real numbers. However, these recent formalizations do not support a particular probability space like the one presented in Hurd's work. Due to this distinguishing feature, Hurd's formalization [27] has been utilized to verify sampling algorithms of a number of commonly used discrete [27] and continuous random variables [24] based on their probabilistic and statistical properties [24]. Due to the availability of a particular probability space as well as the formalization of probability and statistical properties, we build upon Hurd's formalization of measure and probability theories in this paper to analyse cyber-physical transportation systems.

Recently, a probabilistic kernels based mathematical approach [25] has been proposed for formalizing certain probabilistic safety claims. The mathematical framework has been illustrated using an example of a conflict detection system for an aircraft. Our proposed measure theoretic framework for probabilistic reasoning about cyber-physical systems is much more powerful in terms of handling a larger set of problems. Moreover, to the best of our knowledge, the mathematical framework of [25] has not been formalized in a theorem prover yet and thus the corresponding system analysis cannot be considered as completely sound. Our proposed formalization is based on the higher-order-logic formalization of measure and probability theories in HOL and thus the analysis are carried within the sound core of HOL theorem prover.

3 Preliminaries

In this section, we give a brief introduction to theorem proving in general and the HOL theorem prover in particular. The intent is to introduce the main ideas

behind this technique to facilitate the understanding of this paper for the cyber-physical system community.

3.1 Theorem Proving

Theorem proving [16] is a widely used formal verification technique. The system that needs to be analysed is mathematically modelled in an appropriate logic and the properties of interest are verified using computer based formal tools. The use of formal logics as a modelling medium makes theorem proving a very flexible verification technique as it is possible to formally verify any system that can be described mathematically. The core of theorem provers usually consists of some well-known axioms and primitive inference rules. Soundness is assured as every new theorem must be created from these basic or already proved axioms and primitive inference rules.

The verification effort of a theorem in a theorem prover varies from trivial to complex depending on the underlying logic [18]. For instance, first-order logic [13] is restricted to propositional calculus and terms (constants, function names and free variables) and is semi-decidable. A number of sound and complete first-order logic automated reasoners are available that enable completely automated proofs. More expressive logics, such as higher-order logic [6], can be used to model a wider range of problems than first-order logic, but theorem proving for these logics cannot be fully automated and thus involves user interaction to guide the proof tools. For probabilistic analysis, we need to formalize (mathematically model) random variables as functions and their characteristics such as probability distribution properties and expectation, by quantifying over random variable functions. Henceforth, first-order logic does not support such formalization and we need to use higher-order logic to formalize probabilistic analysis.

3.2 HOL Theorem Prover

HOL is an interactive theorem prover developed by Mike Gordon at the University of Cambridge for conducting proofs in higher-order logic. It utilizes the simple type theory of Church [8] along with Hindley-Milner polymorphism [33] to implement higher-order logic. HOL has been successfully used as a verification framework for both software and hardware as well as a platform for the formalization of pure mathematics.

Secure Theorem Proving. In order to ensure secure theorem proving, the logic in the HOL system is represented in the strongly-typed functional programming language ML [34]. An ML abstract data type is used to represent higher-order logic theorems and the only way to interact with the theorem prover is by executing ML procedures that operate on values of these data types. The HOL core consists of only 5 basic axioms and 8 primitive inference rules, which are implemented as ML functions.

Terms. There are four types of HOL terms: constants, variables, function applications, and lambda-terms (denoted function abstractions). Polymorphism, types containing type variables, is a special feature of higher-order logic and is thus supported by HOL. Semantically, types denote sets and terms denote members of these sets. Formulas, sequences, axioms, and theorems are represented by using terms of Boolean types.

Theories. A HOL theory is a collection of valid HOL types, constants, axioms and theorems, and is usually stored as a file in computers. Users can reload a HOL theory in the HOL system and utilize the corresponding definitions and theorems right away. The concept of HOL theory allows us to build upon existing results in an efficient way without going through the tedious process of regenerating these results using the basic axioms and primitive inference rules.

HOL theories are organized in a hierarchical fashion. Any theory may inherit types, definitions and theorems from other available HOL theories. The HOL system prevents loops in this hierarchy and no theory is allowed to be an ancestor and descendant of a same theory. Various mathematical concepts have been formalized and saved as HOL theories by the HOL users. These theories are available to a user when he first starts a HOL session. We utilized the HOL theories of Booleans, lists, sets, positive integers, *real* numbers, measure and probability in our work. In fact, one of the primary motivations of selecting the HOL theorem prover for our work was to benefit from these built-in mathematical theories.

Writing Proofs. HOL supports two types of interactive proof methods: forward and backward. In forward proof, the user starts with previously proved theorems and applies inference rules to reach the desired theorem. In most cases, the forward proof method is not the easiest solution as it requires the exact details of a proof in advance. A backward or a goal directed proof method is the reverse of the forward proof method. It is based on the concept of a *tactic*; which is an ML function that breaks goals into simple sub-goals. In the backward proof method, the user starts with the desired theorem or the main goal and specifies tactics to reduce it to simpler intermediate sub-goals. Some of these intermediate sub-goals can be discharged by matching axioms or assumptions or by applying built-in decision procedures. The above steps are repeated for the remaining intermediate goals until we are left with no further sub-goals and this concludes the proof for the desired theorem.

The HOL theorem prover includes many proof assistants and automatic proof procedures [18] to assist the user in directing the proof. The user interacts with a proof editor and provides it with the necessary tactics to prove goals while some of the proof steps are solved automatically by the automatic proof procedures.

4 Proposed Approach

A cyber-physical transportation system is composed of several interacting components which effect the travel demands within a given area and the services to

satisfy these demands. A vast majority of the underlying components of transportation are nondeterministic. For example, the number of passengers, traffic at a given time and the speed of the individual vehicles are all random quantities. In the higher-order logic theorem proving based analysis, we propose to formalize the behaviour of the given transportation system including its randomized and unpredictable components in higher-order logic. The randomized behaviours would be captured in these formal models by using appropriate random variables.

The second step in theorem proving based probabilistic analysis is to utilize the formal model of the cyber-physical transportation system to express desired system properties as higher-order logic goals. The prerequisite for this step is the ability to express probabilistic and statistical properties related to both discrete and continuous random variables in higher-order logic. All probabilistic properties of discrete and continuous random variables can be expressed in terms of their *Probability Mass Functions* (PMFs) and *Cumulative Distribution Function* (CDFs), respectively. Similarly, most of the commonly used statistical properties can be expressed in terms of the expectation and variance characteristics of the corresponding random variable.

We require the formalization of mathematical definitions of PMF, CDF, expectation and variance for both discrete and continuous random variables in order to be able to express the given system's reliability characteristics as higher-order logic theorems. The third and the final step for conducting the formal probabilistic analysis in a theorem prover is to formally verify the higher-order logic goals developed in the previous step using a theorem prover. For this verification, it would be quite handy to have access to a library of some pre-verified theorems corresponding to some commonly used properties regarding probability distribution functions, expectation and variance. Since, we can build upon such a library of theorems and thus speed up the verification process.

Next, the higher-order logic formalization details associated with the above mentioned prerequisites are briefly described.

4.1 Discrete Random Variables and the PMF

A random variable is called discrete if its range, i.e., the set of values that it can attain, is finite or at most countably infinite. Discrete random variables can be completely characterized by their PMFs that return the probability that a random variable X is equal to some value x , i.e., $Pr(X = x)$.

Discrete random variables can be formalized in higher-order logic as deterministic functions with access to an infinite Boolean sequence B^∞ ; an infinite source of random bits with data type (*natural* \rightarrow *bool*) [27]. These deterministic functions make random choices based on the result of popping bits in the infinite Boolean sequence and may pop as many random bits as they need for their computation. When the functions terminate, they return the result along with the remaining portion of the infinite Boolean sequence to be used by other functions. Thus, a random variable that takes a parameter of type α and ranges over values of type β can be represented by the function

$$\mathcal{F} : \alpha \rightarrow B^\infty \rightarrow (\beta \times B^\infty)$$

For example, a *Bernoulli*($\frac{1}{2}$) random variable that returns 1 or 0 with probability $\frac{1}{2}$ can be modelled as

$$\vdash \text{bit} = \lambda s. (\text{if shd } s \text{ then } 1 \text{ else } 0, \text{stl } s)$$

where the variable s represents the infinite Boolean sequence and the functions `shd` and `stl` are the sequence equivalents of the list operations 'head' and 'tail'. A function of the form $\lambda x.t$ represents a lambda abstraction function that maps x to $t(x)$. The function `bit` accepts the infinite Boolean sequence and returns a pair with the first element equal to either 0 or 1 and the second element equal to the unused portion of the infinite Boolean sequence.

The higher-order logic formalization of probability theory [27] also consists of a probability function \mathbb{P} from sets of infinite Boolean sequences to *real* numbers between 0 and 1. The domain of \mathbb{P} is the set \mathcal{E} of events of the probability. Both \mathbb{P} and \mathcal{E} are defined using the Carathéodory's Extension theorem, which ensures that \mathcal{E} is a σ -algebra: closed under complements and countable unions. The formalized \mathbb{P} and \mathcal{E} can be used to formally verify all basic axioms of probability. Similarly, they can also be used to prove probabilistic properties for random variables. For example, we can formally verify the following probabilistic property for the function `bit`, defined above,

$$\vdash \mathbb{P} \{s \mid \text{fst} (\text{bit } s) = 1\} = \frac{1}{2}$$

where the function `fst` selects the first component of a pair and $\{x \mid C(x)\}$ represents a set of all elements x that satisfy the condition C .

The above mentioned infrastructure can be utilized to formalize most of the commonly used discrete random variables and verify their corresponding PMF relations [27]. For example, the formalization and verification of Bernoulli and Uniform random variables can be found in [27] and of Binomial and Geometric random variables can be found in [24].

4.2 Continuous Random Variables and the CDF

A random variable is called continuous if it ranges over a continuous set of numbers that contains all real numbers between two limits. Continuous random variables can be completely characterized by their CDFs that return the probability that a random variable X is exactly less than or equal to some value x , i.e., $Pr(X \leq x)$.

The sampling algorithms for continuous random variables are non-terminating and hence require a different formalization approach than discrete random variables, for which the sampling algorithms are either guaranteed to terminate or satisfy probabilistic termination, meaning that the probability that the algorithm terminates is 1. One approach to address this issue is to utilize the concept of the non-uniform random number generation [11], which is the process of obtaining arbitrary continuous random numbers using a Standard Uniform

random number generator. The main advantage of this approach is that we only need to formalize the Standard Uniform random variable from scratch and use it to model other continuous random variables by formalizing the corresponding non-uniform random number generation method.

Based on the above approach, a methodology for the formalization of all continuous random variables for which the inverse of the CDF can be represented in a closed mathematical form is presented in [24]. The first step in this methodology is the formalization of the Standard Uniform random variable, which can be done by using the formalization approach for discrete random variables and the formalization of the mathematical concept of limit of a *real* sequence [19]:

$$\lim_{n \rightarrow \infty} (\lambda n \cdot \sum_{k=0}^{n-1} (\frac{1}{2})^{k+1} X_k) \quad (1)$$

where X_k denotes the outcome of the k^{th} random bit; *True* or *False* represented as 1 or 0, respectively. The formalization details are outlined in [24].

The second step in the methodology for the formalization of continuous probability distributions is the formalization of the CDF and the verification of its classical properties. This is followed by the formal specification of the mathematical concept of the inverse function of a CDF. This definition along with the formalization of the Standard Uniform random variable and the CDF properties, can be used to formally verify the correctness of the Inverse Transform Method (ITM) [11]. The ITM is a well known non-uniform random generation technique for generating non-uniform random variables for continuous probability distributions for which the inverse of the CDF can be represented in a closed mathematical form. Formally, it can be verified for a random variable X with CDF F using the Standard Uniform random variable U as follows [24].

$$Pr(F^{-1}(U) \leq x) = F(x) \quad (2)$$

The formalized Standard Uniform random variable can now be used to formally specify any continuous random variable for which the inverse of the CDF can be expressed in a closed mathematical form as $X = F^{-1}(U)$. Whereas, its CDF can be verified based on simple arithmetic reasoning, using the formally verified ITM, given in Equation (2). This approach has been successfully utilized to formalize and verify Exponential, Uniform, Rayleigh and Triangular random variables [24].

4.3 Statistical Properties for Discrete Random Variables

In probabilistic analysis, statistical characteristics play a major role in decision making as they tend to summarize the probability distribution characteristics of a random variable in a single number. Due to their widespread interest, the computation of statistical characteristics has now become one of the core components of every contemporary probabilistic analysis framework.

The expectation for a function of a discrete random variable, which attains values in the positive integers only, is defined as follows [30]

$$Ex_fn[f(X)] = \sum_{n=0}^{\infty} f(n)Pr(X = n) \tag{3}$$

where X is the discrete random variable and f represents a function of X . The above definition only holds if the associated summation is convergent, i.e., $\sum_{n=0}^{\infty} f(n)Pr(X = n) < \infty$. The expression of expectation, given in Equation (3), has been formalized in [24] as a higher-order logic function using the probability function \mathbb{P} . The expected value of a discrete random variable that attains values in positive integers can now be defined as a special case of Equation (3)

$$Ex[X] = Ex_fn[(\lambda n.n)(X)] \tag{4}$$

when f is an identity function. In order to verify the correctness of the above definitions of expectation, they are utilized in [24] to formally verify various properties of expectation like its linearity and Markov’s inequality. These properties not only verify the correctness of the above definitions but also play a vital role in verifying the expectation characteristics of discrete random components of probabilistic systems.

Variance of a random variable X describes the difference between X and its expected value and thus is a measure of its dispersion.

$$Var[X] = Ex[(X - Ex[X])^2] \tag{5}$$

The above definition of variance has been formalized in higher-order logic in [24] by utilizing the formal definitions of expectation, given in Equations (3) and (4). This definition is then formally verified to be correct by proving its classical properties like linearity and Chebyshev’s inequality.

These results allow us to reason about expectation, variance and tail distribution properties of any formalized discrete random variable that attains values in positive integers, e.g., the formal verification for Bernoulli, Uniform, Binomial and Geometric random variables is presented in [24].

4.4 Statistical Properties for Continuous Random Variables

The most commonly used definition of expectation, for a continuous random variable X , is the probability density-weighted integral over the real line.

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx \tag{6}$$

The function f in the above equation represents the *Probability Density Function* (PDF) of X and the integral is the well-known Reimann integral. The above definition is limited to continuous random variables that have a well-defined

PDF. A more general, but not so commonly used, definition of expectation for a random variable X , defined on a probability space (Ω, Σ, P) [14], is as follows:

$$E[X] = \int_{\Omega} X dP \tag{7}$$

This definition utilizes the Lebesgue integral and is general enough to cater for both discrete and continuous random variables. The reason behind its limited usage in the probabilistic analysis domain is the complexity of solving the Lebesgue integral, which takes its foundations from the measure theory that most engineers and computer scientists are not familiar with.

The obvious advantage of using Equation (6) for formalizing expectation of a continuous random variable is the user familiarity with Riemann integral that usually facilitates the reasoning process regarding the expectation properties in the theorem proving based probabilistic analysis approach. On the other hand, it requires extended real numbers, $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, whereas all the foundational work regarding theorem proving based probabilistic analysis, outlined above, has been built upon the standard real numbers \mathbb{R} , formalized by Harrison [19]. The expectation definition given in Equation (7) does not involve extended real numbers, as it accommodates infinite limits without any ad-hoc devices due to the inherent nature of the Lebesgue integral. It also offers a more general solution. The limitation, however, is the compromise on the interactive reasoning effort, as it is not a straightforward task for a user to build on this definition to formally verify the expectation of a random variable.

We have formalized the expectation of a continuous random variable as in Equation (7) by building on top of a higher-order logic formalization of Lebesgue integration theory [?]. Starting from this definition, two simplified expressions for the expectation are verified that allow us to reason about expectation of a continuous random variable in terms of simple arithmetic operations [24]. The first expression is for the case when the given continuous random variable X is bounded in the positive interval $[a, b]$

$$E[X] = \lim_{n \rightarrow \infty} \left[\sum_{i=0}^{2^n-1} (a + \frac{i}{2^n}(b-a))P \left\{ a + \frac{i}{2^n}(b-a) \leq X < a + \frac{i+1}{2^n}(b-a) \right\} \right] \tag{8}$$

and the second one is for an unbounded positive random variable [14].

$$E[X] = \lim_{n \rightarrow \infty} \left[\sum_{i=0}^{n2^n-1} \frac{i}{2^n} P \left\{ \frac{i}{2^n} \leq X < \frac{i+1}{2^n} \right\} + nP(X \geq n) \right] \tag{9}$$

Both of the above expressions do not involve any concepts from Lebesgue integration theory and are based on the well-known arithmetic operations like summation, limit of a real sequence, etc. Thus, users can simply utilize them, instead of Equation (7), to reason about the expectation properties of their random variables and gain the benefits of the original Lebesgue based definition. The formal verification details for these expressions are given in [24]. These

expressions are further utilized to verify the expected values of Uniform, Triangular and Exponential random variables [24]. The above mentioned definition and simplified expressions will also prove to very helpful in formalizing variance and verifying its corresponding properties in a theorem prover, which to the best of our knowledge has not been done so far.

The above mentioned formalization plays a pivotal role for the formal probabilistic analysis of cyber-physical transportation systems. It is a general framework that can be built upon to formally reason about any kind of a system and property. Besides that, there are numerous other advantages of our proposed approach compared to the proof-based languages like Event-B. A couple of worth mentioning features include: (1) absolute correctness of results due to the inherent strong typed nature and soundness of higher-order logic theorem proving, and (2) the ability to verify all sorts of properties, including the temporal ones, due to the expressiveness of the underlying higher-order logic.

In the next section, we illustrate the usefulness and practical effectiveness of the proposed approach by analysing a control algorithm of autonomous vehicles moving as a platoon using HOL.

5 Vehicle Platoon Control Algorithm

A platoon is a set of self-operating vehicles moving in a convoy. It can be seen as a road-train where cars are linked by software, instead of hardware. Platooning has several potential uses in an urban mobility system: augmenting throughput, herding unused cars to stations, or running transient buses, for instance.

Homologous to other transportation systems, vehicle platoons also exhibit many unpredictable characteristics, like the platoon speed, inter-platoon gaps, intra-platoon headways and the inter-arrival time between consecutive platoons. In this study, we concentrate only on one randomized aspect, i.e., headways. The gaps between the vehicles in different platoons significantly impact the performance of an un-signalized intersection. Our formal modelling is based on a dichotomized headway model [10]. This model distinguishes the free vehicles (platoon leaders that are travelling without interacting with the vehicles ahead) and the bunched vehicles (platoon leader followers) in terms of their headway distribution. It assumes that the free vehicles with a proportion, α , of all platoons follow the displaced exponential headway distribution while the bunched vehicles $(1 - \alpha)$ have the same constant headway t_m .

We formalized the underlying continuous random variable of the dichotomized headway model described above in HOL as follows:

Definition 1: *Dichotomized Headway Random Variable*

$$\vdash \forall l \ a \ t_m \ s. \text{headway_rv } l \ a \ t_m \ s = (\lambda x. -\frac{1}{t} (\ln (1 - x))/a + t_m) (\text{std_unif_cont } s)$$

where the variables l , a , t_m , and s , denote the decay constant, variable α , variable t_m , and the infinite boolean sequence, formalized in [27], respectively. The function `std_unif_cont` represents the standard uniform random variable

and has been used to facilitate the inverse transform method based formalization of the given random variable, as described in the previous section.

Based on Definition 1, we also formally verified the following useful Cumulative Distribution Function (CDF) property.

Theorem 1: *CDF for the Dichotomized Headway Random Variable*

$$\vdash \forall l \ a \ t_m \ t. (0 < l) \wedge (0 < a) \Rightarrow \\ \text{cdf} (\lambda s. \text{headway_rv } l \ a \ t_m \ s) \ t = \\ \text{if } t \leq t_m \text{ then } 0 \text{ else } (1 - a (\exp (-1 (t - t_m))))$$

where `cdf` represents the HOL function for the CDF. The verification of Theorem 1 was done interactively and the formal reasoning relied heavily upon the probabilistic analysis related formalization [27,24], transcendental functions and real analysis. It is important to note that the above mentioned theorem is universally quantified for all the parameters, which means that these parameters can be specialized to obtain any specific results. Moreover, the CDF allows us to reason about any probabilistic property of the system. For example, we used it to reason about the probability of the headway being greater than $x + t_m$ seconds as follows:

Theorem 2: *Probabilistic Property of Platoon Headway*

$$\vdash \forall l \ a \ t_m \ x. (0 < l) \wedge (0 < a) \wedge (0 < x) \Rightarrow \\ \mathbb{P} \{s \mid (\text{headway_rv } l \ a \ t_m \ s) > x + t_m\} = a (\exp (-1 x))$$

The above exercise illustrates the fact that interactive theorem proving is capable of conducting probabilistic analysis of cyber-physical transportation systems with at least the same degree of accuracy as the analytical proof techniques usually carried out using paper-and-pencil proof methods; a novelty that cannot be achieved by any other computer based techniques, such as simulation or model checking. As aforementioned, simulation-based techniques rely on many approximations and thus can never achieve accurate results. Similarly, due to the inherent limitations of the state-based formal methods, discussed in Section 2 they cannot evaluate the values as precisely as we have attained using the proposed approach.

6 Conclusions

In this paper, we have advocated the unification of formal analysis with quantitative reasoning for the verification of cyber-physical transportation systems. The main idea is that in addition to analysing the absolute correctness, which at times is not possible, a probabilistic analysis can provide a better insight about the model. In this fashion, it is easier for stakeholders to obtain a probability of occurrence of a hazard in terms of the likelihood of components failures. The paper presents an introduction to the available framework for formal probabilistic analysis in higher-order logic and proposes to use it to analyse the probability aspects of considered systems. To demonstrate our approach, we have used

a small-scale example of inter-platoon headway property of a platooning algorithm. To the best of our knowledge, this is the first time that higher-order logic theorem proving has been proposed to be used in this context in the open literature.

In future, we intend to expand the boundaries of our work to a full-scale platooning system [37] and the transport domain model [31]. Our aim is to bring their Event-B specifications into higher-order logic in HOL and integrate them with associated random variable models, like the one given in Definition 1 and reason about more advanced probabilistic properties.

References

1. Abrial, J.R.: *The B Book*. Cambridge University Press (1996)
2. Abrial, J.R.: *Modeling in Event-B: System and Software Engineering*. Cambridge University Press (2010)
3. Badeau, F., Amelot, A.: Using B as a High Level Programming Language in an Industrial Project: Roissy VAL. In: Treharne, H., King, S., Henson, M., Schneider, S. (eds.) ZB 2005. LNCS, vol. 3455, pp. 334–354. Springer, Heidelberg (2005)
4. Baier, C., Katoen, J.: *Principles of Model Checking*. MIT Press (2008)
5. Behm, P., Benoit, P., Faivre, A., Meynadier, J.-M.: *Météor: A Successful Application of B in a Large Project*. In: Wing, J.M., Woodcock, J. (eds.) FM 1999. LNCS, vol. 1708, pp. 369–387. Springer, Heidelberg (1999)
6. Brown, C.: *Automated Reasoning in Higher-order Logic*. College Publications (2007)
7. Cardell-Oliver, R.: *The Formal Verification of Hard Real-time Systems*. PhD Thesis, University of Cambridge, UK (1992)
8. Church, A.: A Formulation of the Simple Theory of Types. *Journal of Symbolic Logic* 5, 56–68 (1940)
9. Clarke, E.M., Zuliani, P.: Statistical Model Checking for Cyber-Physical Systems. In: Bultan, T., Hsiung, P.-A. (eds.) ATVA 2011. LNCS, vol. 6996, pp. 1–12. Springer, Heidelberg (2011)
10. Cowan, R.J.: Useful Headway Models. *Transportation Research* 9, 371–375 (1975)
11. Devroye, L.: *Non-Uniform Random Variate Generation*. Springer (1986)
12. Elleuch, M., Hasan, O., Tahar, S., Abid, M.: Formal Analysis of a Scheduling Algorithm for Wireless Sensor Networks. In: Qin, S., Qiu, Z. (eds.) ICFEM 2011. LNCS, vol. 6991, pp. 388–403. Springer, Heidelberg (2011)
13. Fitting, M.: *First-Order Logic and Automated Theorem Proving*. Springer (1996)
14. Galambos, J.: *Advanced Probability Theory*. Marcel Dekker Inc. (1995)
15. Gomes, A., Mota, A., Sampaio, A., Ferri, F., Buzzi, J.: Systematic Model-Based Safety Assessment Via Probabilistic Model Checking. In: Margaria, T., Steffen, B. (eds.) ISoLA 2010, Part I. LNCS, vol. 6415, pp. 625–639. Springer, Heidelberg (2010)
16. Gordon, M.: Mechanizing Programming Logics in Higher-Order Logic. In: *Current Trends in Hardware Verification and Automated Theorem Proving*, pp. 387–439. Springer (1989)
17. Hallerstede, S., Hoang, T.S.: Qualitative Probabilistic Modelling in Event-B. In: Davies, J., Gibbons, J. (eds.) IFM 2007. LNCS, vol. 4591, pp. 293–312. Springer, Heidelberg (2007)

18. Harrison, J.: Formalized Mathematics. Technical Report 36, Turku Centre for Computer Science, Finland (1996)
19. Harrison, J.: Theorem Proving with the Real Numbers. Springer (1998)
20. Harrison, J.: Handbook of Practical Logic and Automated Reasoning. Cambridge University Press (2009)
21. Harrison, J., Slind, K., Arthan, R.D.: HOL. In: Wiedijk, F. (ed.) The Seventeen Provers of the World. LNCS (LNAI), vol. 3600, pp. 11–19. Springer, Heidelberg (2006)
22. Hasan, O., Afshar, S.K., Tahar, S.: Formal Analysis of Optical Waveguides in HOL. In: 22nd International Conference on Theorem Proving in Higher-Order Logics, Munich, Germany. Springer (2009)
23. Hasan, O., Tahar, S., Abbasi, N.: Formal Reliability Analysis using Theorem Proving. IEEE Transactions on Computers 59(5), 579–592 (2010)
24. Hasan, O., Tahar, S.: Formal Probabilistic Analysis: A Higher-Order Logic Based Approach. In: Frappier, M., Glässer, U., Khurshid, S., Laleau, R., Reeves, S. (eds.) ABZ 2010. LNCS, vol. 5977, pp. 2–19. Springer, Heidelberg (2010)
25. Herencia-Zapana, H., Hagen, G., Narkawicz, A.: Formalizing Probabilistic Safety Claims. In: Bobaru, M., Havelund, K., Holzmann, G.J., Joshi, R. (eds.) NFM 2011. LNCS, vol. 6617, pp. 162–176. Springer, Heidelberg (2011)
26. Hölzl, J., Heller, A.: Three Chapters of Measure Theory in Isabelle/HOL. In: van Eekelen, M., Geuvers, H., Schmaltz, J., Wiedijk, F. (eds.) ITP 2011. LNCS, vol. 6898, pp. 135–151. Springer, Heidelberg (2011)
27. Hurd, J.: Formal Verification of Probabilistic Algorithms. PhD Thesis, University of Cambridge, UK (2002)
28. Kwiatkowska, M., Norman, G., Parker, D.: Controller Dependability Analysis by Probabilistic Model Checking. Control Engineering Practice 15(11), 1427–1434 (2007)
29. Kwiatkowska, M., Norman, G., Parker, D.: PRISM: Probabilistic Symbolic Model Checker. In: Field, T., Harrison, P.G., Bradley, J., Harder, U. (eds.) TOOLS 2002. LNCS, vol. 2324, pp. 200–204. Springer, Heidelberg (2002)
30. Levine, A.: Theory of Probability. Addison-Wesley (1971)
31. Mashkoor, A., Jacquot, J.P.: Utilizing Event-B for Domain Engineering: A Critical Analysis. Requirements Engineering 16(3), 191–207 (2011)
32. Mhamdi, T., Hasan, O., Tahar, S.: On the Formalization of the Lebesgue Integration Theory in HOL. In: Kaufmann, M., Paulson, L.C. (eds.) ITP 2010. LNCS, vol. 6172, pp. 387–402. Springer, Heidelberg (2010)
33. Milner, R.: A Theory of Type Polymorphism in Programming. Journal of Computer and System Sciences 17, 348–375 (1977)
34. Paulson, L.: ML for the Working Programmer. Cambridge University Press (1996)
35. Siddique, U., Hasan, O.: Formal Analysis of Fractional Order Systems in HOL. In: Formal Methods in Computer Aided Design, pp. 163–170 (2011)
36. Tarasyuk, A., Troubitsyna, E., Laibinis, L.: Towards Probabilistic Modelling in Event-B. In: Méry, D., Merz, S. (eds.) IFM 2010. LNCS, vol. 6396, pp. 275–289. Springer, Heidelberg (2010)
37. Yang, F., Jacquot, J.-P.: Scaling Up with Event-B: A Case Study. In: Bobaru, M., Havelund, K., Holzmann, G.J., Joshi, R. (eds.) NFM 2011. LNCS, vol. 6617, pp. 438–452. Springer, Heidelberg (2011)

DEM Reconstruction of Coastal Geomorphology from DInSAR

Maged Marghany

Universiti Teknologi Malaysia, Institute for Geospatial Science and Technology (INSTeG)
Johor Bahru, 81310 UTM, Skudai, Malaysia
magedupm@hotmail.com,
maged@utm.my

Abstract. The paper is focused on Digital Elevation Model (DEM) reconstruction from differential interferometry synthetic aperture radar (DInSAR). In doing so, conventional DInSAR procedures are implemented to three repeat passes of RADARSAT-1 SAR fine mode data (F1). Further, the multichannel MAP height estimator is implemented with phase unwrapping technique. Consequently, the multichannel MAP height estimator is used to eliminate the phase decorrelation impact from the interferograms. The study shows the performance of DInSAR method using the multichannel MAP height estimator is better than DInSAR technique which is validated by a lower range of error (0.01 ± 0.11 m) with 90% confidence intervals. In conclusion, integration of the multichannel MAP height estimator with phase unwrapping produce accurate 3-D coastal geomorphology reconstruction.

Keywords: DInSAR, fringe, interferogram, The multichannel MAP height estimator algorithm, coastal geomorphology, spit, Digital Elevation Model (DEM).

1 Introduction

Interferometric synthetic aperture radar (InSAR) satellite data have been intensively used to study the Earth surface deformation [1][2]. InSAR uses two or more single look complex synthetic aperture radar (SAR) images to produce maps of surface deformation [3] or digital elevation [4][5]. Over time-spans of days to years, InSAR can detect the centimetre-scale of deformation changes [5]. InSAR technique involves of two complex SAR data that acquired from slightly different flight path [2][4]. The phase image is produced by multiplied the complex SAR image by the coregistered complex conjugate pixels of the other SAR image. In this regard, the phase difference of the two images is processed to attain height and/or motion information of the Earth's surface [6]. Further, the precision DEMs with of a couple of ten meters can produce from InSAR technique compared to conventional remote sensing methods. Nevertheless, the availability of the precision DEMs may a cause of two-pass InSAR; regularly 90 m SRTM data may be accessible for numerous territories[5]. InSAR, consequently, provides DEMs with 1-10 cm accuracy, which can be improved to millimetre level by DInSAR. Even so, alternative datasets must acquire at high latitudes or in areas of rundown coverage[6][13]. However, the

baseline decorrelation and temporal decorrelation make InSAR measurements unfeasible [8][9][10][11]. In this context, Gens [12] reported the length of the baseline designates the sensitivity to height changes and sum of baseline decorrelation. Further, Gens [12] stated the time difference for two data acquisitions is a second source of decorrelation. Indeed, the time differences while comparing data sets with a similar baseline length acquired one and 35 days apart suggests only the temporal component of the decorrelation. Therefore, the loss of coherence in the same repeat cycle in data acquisition are most likely because of baseline decorrelation. According to Roa et al. [7], uncertainties could arise in DEM because of limitation InSAR repeat passes. In addition, the interaction of the radar signal with troposphere can also induce decorrelation. This is explained in several studies [3][8][15].

Commonly, the propagation of the waves through the atmosphere can be a source of error exist in most interferogram productions. When the SAR signal propagated through a vacuum it should theoretically be subjected to some decent accuracy of timing and cause phase delay [3]. A constant phase difference between the two images caused by the horizontally homogeneous atmosphere was over the length scale of an interferogram and vertically over that of the topography. The atmosphere, however, is laterally heterogeneous on length scales both larger and smaller than typical deformation signals [9]. In other cases the atmospheric phase delay, however, is caused by vertical inhomogeneity at low altitudes and this may result in fringes appearing to correspond with the topography. Under this circumstance, this spurious signal can appear entirely isolated from the surface features of the image, since the phase difference is measured other points in the interferogram, would not contribute to the signal [3]. This can reduce seriously the low signal-to-noise ratio (SNR) which restricted to perform phase unwrapping. Accordingly, the phases of weak signals are not reliable. According to Yang et al., [11], the correlation map can be used to measure the intensity of the noise in some sense. It may be overrated because of an inadequate number of samples allied with a small window [9]. Weights are initiated to the correlation coefficients according to the amplitudes of the complex signals to estimate accurate reliability [11].

Baselice et al., [21], Ferraiuolo et al., [22], Ferraiuolo et al., [23] and Ferret et al., [24] have developed multichannel MAP height estimator based on a Gaussian Markov Random Field (GMRF) to solve the uncertainties of DEM reconstruction from InSAR technique. They found that multichannel MAP height estimator have managed the phase discontinuities and improved the DEM profile. Taking advantage of the fact that the multichannel MAP height estimator for solving uncertainty problem because of decorrelation and the low signal-to-noise ratio (SNR) in data sets. This work hypothesises that integration of The multichannel MAP height estimator algorithm with phase unwrapping can produce accurately digital elevation of object deformation. The aim of this paper is to explore the precision of the digital elevation models (DEM) derived from RADARSAT-1 fine mode data (F1) and, thus, the potential of the sensor for mapping coastal geomorphologic feature changes. Depending on the results, a wider application of F1 mode data for the study of Kuala Terengganu mouth river landscapes is envisaged.

2 DInSAR Data Processing

Two methods are involved to perform DInSAR from RADARSAT-1 SAR F1 mode data (i) conventional DInSAR procedures; and (ii) Multichannel MAP height estimator.

2.1 Conventional DInSAR Method

The DInSAR technique measures the block displacement of land surface caused by subsidence, earthquake, glacier movement, and volcano inflation to cm or even mm accuracy [10]. According to Lee [9], the surface displacement can estimate using the acquisition times of two SAR images S_1 and S_2 . The component of surface displacement thus, in the radar-look direction, contributes to further interferometric phase (ϕ) as

$$\phi = \frac{4\pi}{\lambda}(\Delta R + \zeta) \quad (1)$$

where ΔR is the slant range difference from satellite to target respectively at different time, λ is the RADARSAT-1 SAR fine mode wavelength which is about 5.6 cm for C_{HH} -band. According to Lee [9], for the surface displacement measurement, the zero-baseline InSAR configuration is the ideal as $\Delta R = 0$, so that

$$\phi = \phi_d = \frac{4\pi}{\lambda} \zeta \quad (2)$$

In actual fact, zero-baseline, repeat-pass InSAR configuration is hardly achievable for either spaceborne or airborne SAR. Therefore, a method to remove the topographic phase as well as the system geometric phase in a non-zero baseline interferogram is needed. If the interferometric phase from the InSAR geometry and topography can strip of from the interferogram, the remnant phase would be the phase from block surface movement, providing the surface maintains high coherence [5]. Zebker et al. [20] used the three-pass method to remove topographic phase from the interferogram. This method requires a reference interferogram, which is promised to contain the topographic phase only. The three-pass approach has the advantage in that all data is kept within the SAR data geometry while DEM method can produce errors by misregistration between SAR data and cartographic DEM [9]. The three-pass approach is restricted by the data availability. The three-passes DInSAR technique uses another InSAR pair as a reference interferogram that does not contain any surface movement event as

$$\phi' = \frac{4\pi}{\lambda} \Delta R' \quad (3)$$

Incorporating equations 2 and 3 gives the phase difference, only from the surface displacement as

$$\phi_d = \phi - \frac{\Delta R}{\Delta R'} \phi' = \frac{4\pi}{\lambda} \zeta . \tag{4}$$

For an exceptional case where $\frac{\Delta R}{\Delta R'}$ in equation 4 there is a positive integer number, phase unwrapping may not be necessary [1]. However, this is not practical and it is difficult to achieve from the system design for a repeat-pass interferometer. From equation 4 the displacement sensitivity of DInSAR is given as

$$\frac{\partial \phi_d}{\partial \zeta} = \frac{4\pi}{\lambda} . \tag{5}$$

2.2 DEM Reconstruction Using Multichannel MAP Height Estimator

The multichannel MAP height estimator is used to solve the decorrelation problem with InSAR and DInSAR methods. This algorithm is adopted from the study of Baselice et al., [21]. Following Baselice et al., [21], The interferometric phase signal can be expressed by the following mathematical formula [21],

$$\phi_{sn} = \left\langle \left(\frac{4\pi}{\lambda R_0 \sin \theta} \right) B_{\perp n} h_s + \alpha \right\rangle_{2\pi} , \quad n=1,2,\dots,N; \quad s=1,2,\dots,S \tag{6}$$

Where s is the pixel position inside the SAR image, n is considered interferogram channel, λ is sensor wavelength, R_0 is the distance between the center of the scene and the master antenna, $B_{\perp n}$ is the orthogonal baseline, h_s is height value, α is the phase decorrelation noise, and is θ incident angle. Further, $\langle \cdot \rangle_{2\pi}$ represents the “ modulo - 2π ”.

Assume N is independent interferogram channels, then the problem involves in modeling the height values h_s is starting from the $S \times N$ estimated wrapped phase ϕ_{sn} . Following Ferraiuolo et al. [22], the problem of modeling height can be solved using a MAP height estimation approach. In this regard, multichannel likelihood function F_{mc} is considered and is given by

$$F_{mc}(\phi_s | \zeta_s) = \prod_{n=1}^N f(\phi_{sn} | \zeta_s) \tag{7}$$

Where $F(\phi_{sn} | \zeta_s)$ is the signal channel likelihood function, ϕ_s is measured wrapped phase data referred to the pixel s , $\phi_s = [\phi_{s1}, \phi_{s2}, \dots, \phi_{sN}]^T$, and ζ_s is collected vector height values where $\zeta_s = [\zeta_1, \zeta_2, \dots, \zeta_S]^T$. Following Baselice et al., [21] and Ferraiuolo et al. [22] a MAP height estimation can be given by

$$\hat{\zeta}_{MAP} = \arg_{\zeta} \max \ln \left[\left(\prod_{s=1}^S F_{mc}(\phi_s \mid \zeta_s) g(\zeta; \hat{\sigma}) \right) \right] \quad (8)$$

Where $g(\cdot)$ is a prior probability density function (pdf) which is adopted by using Gaussian Markov Random Field and $\hat{\sigma}$ is the hyperparameter vector which is not a prior known. According to Baselice et al., [21], it has to be estimated starting from the measured interferograms. This is accomplished by considering sub-bands, corresponding to different azimuth looks. In this regard, a Gaussian Markov Random Field (GMRF) as a-priori model, whose expression is:

$$g(\zeta, \hat{\sigma}) = \frac{1}{Z(\hat{\sigma})} e^{-\sum_{s=1}^{SuV} \sum_{k \in N_s} \left[\frac{(\zeta_s - \zeta_k)^2}{2\hat{\sigma}_{sk}^2} \right]} \quad (9)$$

where N_s is the neighborhood system of sth pixel, and s are the so-called hyperparameters, which are representative of the local characteristics of the image h , $\hat{\sigma}$ is the hyperparameter vector collecting all s values, and $Z(\hat{\sigma})$ is the partition function [9] necessary to normalize the pdf [21][22][23][24]. Finally, the normalized reconstruction square error is given by

$$\mathcal{E} = \frac{\left\| \hat{\zeta} - \zeta \right\|^2}{\left\| \zeta \right\|^2} \quad (10)$$

Where ζ is height which is derived from equation 5 and the elevation derived from equation 8 ($\hat{\zeta}$). Although the reconstruction, considering the limited number of available data (four channels), is good, we want to improve its quality, particularly on the discontinuities. Then, algorithm is implemented based on the introduction of ground elevation data.

2.3 Ground Survey

The GPS survey used to: (i) to record exact geographical position of shoreline; (ii) to determine the cross-sections of shore slopes; (iii) to corroborate the reliability of DInSAR data co-registered; and finally, (iv) to create a reference network for future surveys. The geometric location of the GPS survey was obtained by using the new satellite geodetic network, IGM95. After a careful analysis of the places and to identify the reference vertexes, we thickened the network around such vertexes to perform the measurements for the cross sections (transact perpendicular to the coastline). The GPS data collected within 20 sample points scattered along 400 m coastline. The interval distance of 20 m between sample location is considered. In every sample location, Rec-Alta (Recording Electronic Tacheometer) was used to

acquire the spit elevation profile. The ground truth data were acquired on 23 December 2003 March 26, 2005, during satellite passes.

3 Results and Discussion

In the present study, DInSAR methods are implemented on RADARSAT-1 SAR data sets of 23 November 1999 (SLC-1), 23 December 2003 (SLC-2) and March 26, 2005, (SLC-3) of Fine mode data (F1) are implemented (Fig.1). These data are C-band and had the lower signal-to-noise ratio owing to their HH polarization with wavelength of 5.6 cm and frequency of 5.3 GHz. The Fine beam mode is intended for applications which require the best spatial resolution available from the RADARSAT-1 SAR system. The azimuth resolution is 8.4 m, and range resolution ranges between 9.1 m to 7.8 m. Originally, five Fine beam positions, F1 to F5, are available to cover the far range of the swath with incidence angle ranges from 37° to 47° . By modifying timing parameters, 10 new positions have been added with offset ground coverage. Each original Fine beam position can either be shifted closer to or farther away from Nadir. The resulting positions are denoted by either an N (Near) or F (Far). For example, F1 is now complemented by F1N and F1F [19]. Finally, RADARSAT-1 requires 24 days to return to its original orbit path. This means that for most geographic regions, it will take 24 days to acquire exactly the same image (the same beam mode, position, and geographic coverage). However, RADARSAT's imaging flexibility allows images to be acquired on a more frequent basis [19].

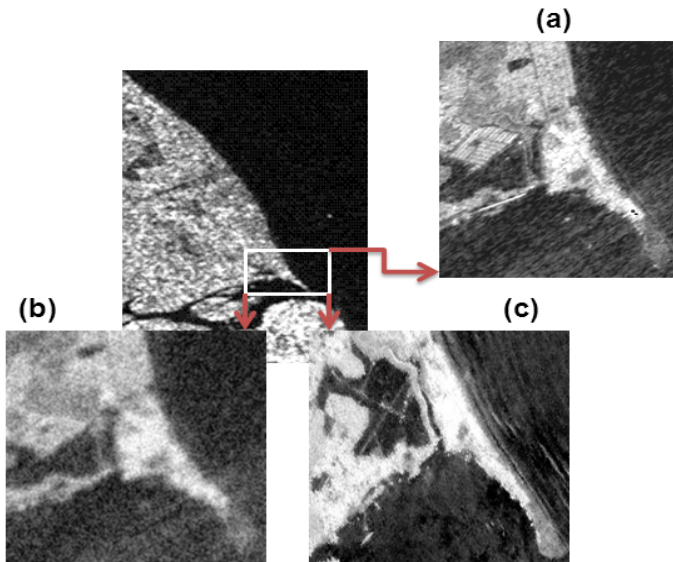


Fig. 1. RADARSAT-1 SAR fine mode data acquisition (a) master data, (b) data slave 1 data and (c) slave 2 data

The spit is located across the largest hydrological communications between the estuary and the South China Sea i.e. mouth river of Kuala Terengganu [15]. In addition, Sand materials make up the entire of this beach [14][16]. Further, the geological structure of the Kuala Terengganu River is composed of alluvium soil, and carbonate rock with decay living organism [17][18].

Fig. 2 shows that urban zone dominated with higher coherence of 0.8 than vegetation and sand areas. Since three F1 mode data acquired in wet north-east monsoon period, there is an impact of wet sand on radar signal penetration which causing weak penetration of radar signal because of dielectric. Figure 2 shows the ratio coherence image, clearly the total topographic decorrelation effects along the radar-facing slopes are dominant and highlighted as bright features of 3 over a grey background. This is caused by the micro-scale movement of the sand particles driven by the coastal hydrodynamic, and wind continuously changes the distribution of scatterers resulting in rapid temporal decorrelation which has contributed to decorrelation in the spit zone.

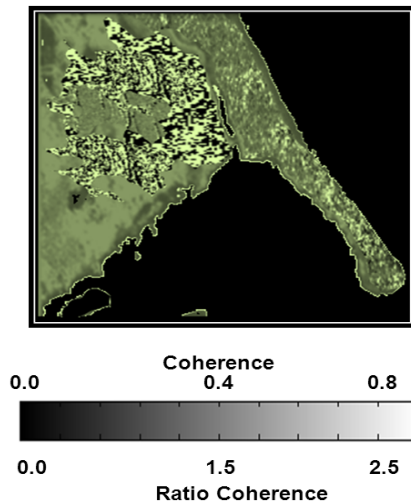


Fig. 2. Composite result of coherence and ratio coherence in F1 mode data

Clearly, the random changes in the surface scatterer locations among data acquisitions with a wavelength of 5.6 cm for C-band are sufficient to decorrelate the interferometric signal. Under this circumstance, it will be visible in the coherence data (Fig.2). Since vegetation and wet sand changes may also reduce the coherence because the estuary area has tides and water lines that are so highly variable, this can be defined in probabilistic terms. The geomorphology feature of spit is rendered meaningless or unreliable in the long term because of their high variability. The overall scene is highly incoherent, not only because of the meteorological conditions and the vegetation cover at the time but also because of ocean surface turbulent changes. This decorrelation caused poor detection of spit which induce large

ambiguities because of poor coherence and scattering phenomenology. The ground ambiguity and ideal assumption that volume-only coherence can be acquired in at least one polarization. This assumption may fail when vegetation is thick, dense, or the penetration of electromagnetic wave is weak. This is agreed with study of Lee [9].

Fig. 3 shows the interferogram created from F1 data. For three data sets, only small portion of the scene processed because of temporal decorrelation. According to Luo et al.,[10], the SAR interferogram is considered to be difficult to unwrap because of its large areas of low coherence, which caused by temporal decorrelation. These areas of low coherence segment the interferogram into many pieces, which creates difficulties for the unwrapping algorithms (Fig.3). In this context, Lee [9] reported that when creating an interferogram of surface deformation by using InSAR, it is not always true that an interference pattern (fringes) of an initial interferogram directly shows surface deformation. Indeed, the difference in phase between two observations is influenced by things outside surface deformation.

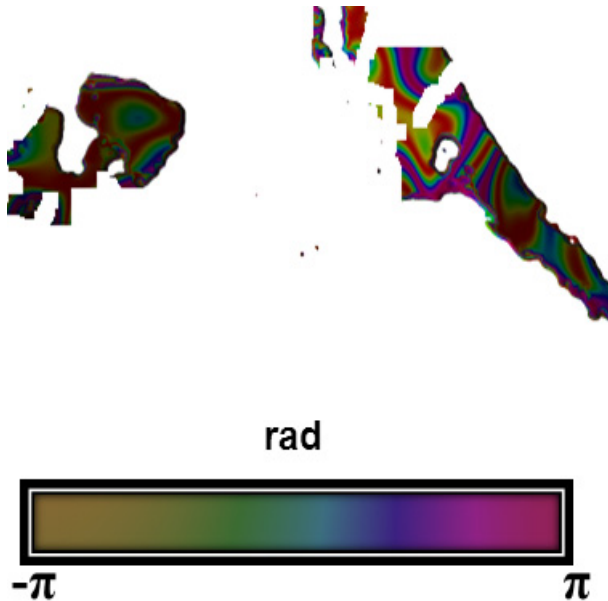


Fig. 3. Interfeorgram generated from F1 mode data

Fig. 4 shows the interferogram created using multichannel MAP height estimator. The full color cycle represents a phase cycle, covering range between $-\pi$ to π . In this context, the phase difference given module 2π ; is color encoded in the fringes. Seemingly, the color bands change in the reverse order, indicating that the center has a great deformation along the spit. This shift corresponds to 0.4 centimetres (cm) of coastal deformation over the distance of 500 m. The urban area dominated by deformation of 2.8 cm. Fig. 5 represents 3-D spit reconstruction using multichannel MAP height estimator with the maximum spit 's elevation is 3 m with gentle slope of 0.86 m.

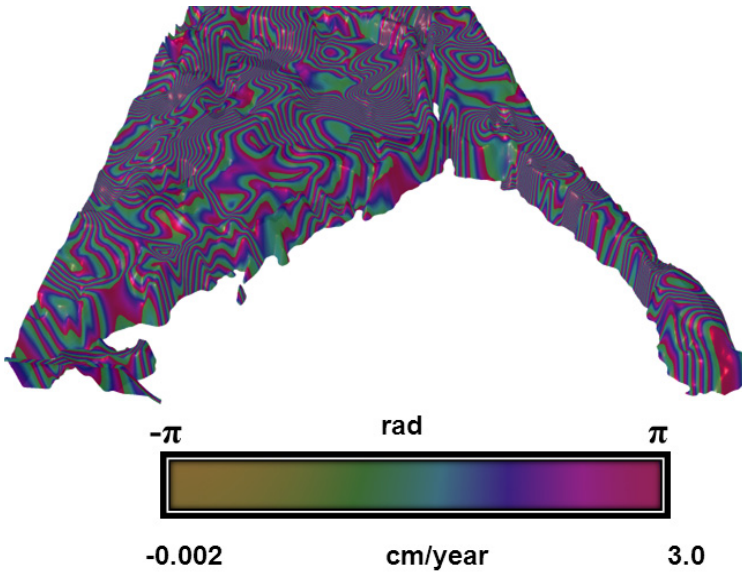


Fig. 4. Fringe Interferometry generated by multichannel MAP height estimator

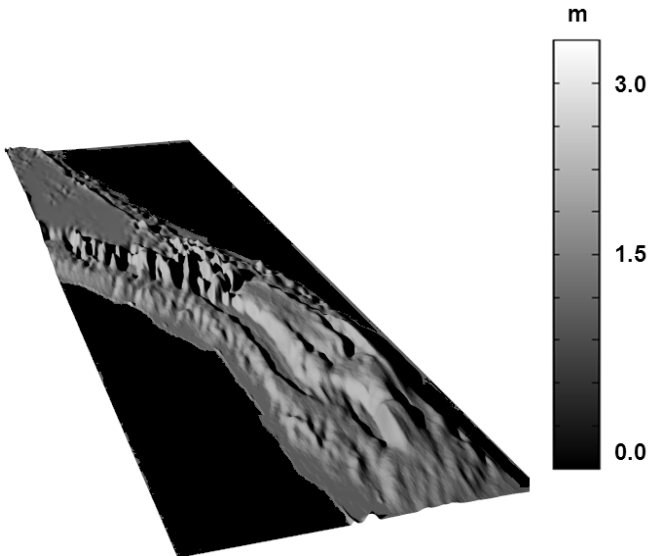


Fig. 5. DEM of coastal spit

Table 1 shows the statistical comparison between the simulated DEM from the DInSAR, real ground measurements and with using multichannel MAP height estimator. This table represents the bias (averages mean the standard error, 90 and 95% confidence intervals, respectively). Evidently, the DInSAR using multichannel MAP height estimator has bias of -0.03 m, lower than ground measurements and the

DInSAR method. Therefore, multichannel MAP height estimator has a standard error of mean of ± 0.023 m, lower than ground measurements and the DInSAR method. Overall performances of DInSAR method using multichannel MAP height estimator is better than DInSAR technique which is validated by a lower range of error (0.01 ± 0.11 m) with 90% confidence intervals.

Table 1. Statistical Comparison between DInSAR and DInSAR- Multichannel MAP height estimator

Statistical parameters	DInSAR techniques			
	DInSAR	Multichannel MAP Height Estimator		
Bias	2.5	-0.03		
Standard error of the mean	1.5	0.02		
	Lower	Upper	Lower	Upper
90%(90% confidence interval)	1.2	2.6	0.01	0.12
95%(95% confidence interval)	0.98	2.4	0.02	0.11

Multichannel MAP height estimator produced perfect pattern of fringe interferometry compared with one produced by DInSAR technique (Fig. 4). It shows there are many deformations of over several centimetres. In these deformations, it is known the deformation in spit because of coastline sedimentation. The other deformations, however, are caused not by the movement of the coastal sediment but the spatial fluctuation of water vapour in the atmosphere. In addition, the growths of urban area induces also land cover changes. Further, it can be noticed that Multichannel MAP height estimator detailed edges with discernible fringes. Indeed, Fig 4. shows smooth interferogram, in terms of spatial resolution maintenance, and noise reduction, compared to conventional methods [2][5][8][10].

This has been contributed since the local estimation of hyperparameters is very powerful, because it allows one to localize the flat regions and the discontinuities of the image, taking into account the local characteristics of the profile and generating a hyperparameter map for the profile. In this context, the estimation from the measured interferograms is typically solved iteratively, using the *Expectation-Maximization*. This leads to a powerful and general model, well suited to represent a wide class of height profiles (Fig.4). In addition, fringe discontinuities that shown in Fig.3, have been corrected using by a Gaussian Markov Random Field (GMRF) that provides large number of interpolated samples over corrupted fringe data.

This study confirms the work done by Baselice et al., [21], Ferraiuolo et al., [22], Ferraiuolo et al., [23] and Ferret et al., [24]. Moreover, Multichannel MAP height estimator can also used to reconstruct the DEM of coastal geomorphology features in

presence of very high discontinuities and very low coherence in addition to its main purpose to determine 3D reconstruction urban areas as discussed by Baseline et al., [21].

4 Conclusions

This work has demonstrated the 3-D spit reconstruction from DInSAR using three C-band SAR images acquired by RADARSAT-1 SAR F1 mode data. The conventional method of DInSAR used to create 3-D coastal geomorphology reconstruction. Nevertheless, it was difficult to generate phase and interferogram using conventional DInSAR because of decorrelation impact. The result shows that spit and vegetation zone have poor coherence of 0.25 as compared to the urban area. In addition, only small portion of the F1 mode scene was processed because of decorrelation effect. Finally, The multichannel MAP height estimator used to reconstruct fringe pattern, and 3-D from decorrelate unwrapped phase. The fringe pattern shows the deformation of 0.4 cm along spit and 1.4 cm in urban area. Further, the maximum 3-D spit elevation is 3 m with standard error of mean of ± 0.02 m. In addition, the multichannel MAP height estimator is better than conventional DInSAR technique with lower range of error (0.01 ± 0.11 m). In conclusion, the multichannel MAP height estimator could be an excellent tool for 3-D coastal geomorphology reconstruction from SAR data the under circumstance of very low coherence and discontinuities.

References

1. Massonnet, D., Feigl, K.L.: Radar interferometry and its application to changes in the earth's surface. *Rev. Geophys.* 36, 441–500 (1998)
2. Burgmann, R., Rosen, P.A., Fielding, E.J.: Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. *Ann. Rev. of Earth and Plan. Sci.* 28, 169–209 (2000)
3. Hanssen, R.F.: *Radar Interferometry: Data Interpretation and Error Analysis*. Kluwer Academic, Dordrecht (2001)
4. Zebker, H.A., Rosen, P.A., Hensley, S.: Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps. *J. Geophys. Res.* 102, 7547–7563 (1997)
5. Askne, J., Santoro, M., Smith, G., Fransson, J.E.S.: Multitemporal repeat-pass SAR interferometry of boreal forests. *IEEE Trans. Geosci. Remote Sens.* 41, 1540–1550 (2003)
6. Nizalapur, V., Madugundu, R., Shekhar Jha, C.: Coherence-based land cover classification in forested areas of Chattisgarh, Central India, using environmental satellite—advanced synthetic aperture radar data. *J. Appl. Rem. Sens.* 5, 059501–1-059501-6 (2011)
7. Rao, K.S., Al Jassar, H.K., Phalke, S., Rao, Y.S., Muller, J.P., Li, Z.: A study on the applicability of repeat pass SAR interferometry for generating DEMs over several Indian test sites. *Int. J. Remote Sens.* 27, 595–616 (2006)
8. Rao, K.S., Al Jassar, H.K.: Error analysis in the digital elevation model of Kuwait desert derived from repeat pass synthetic aperture radar interferometry. *J. Appl. Remote Sens.* 4, 1–24 (2010)

9. Lee, H.: Interferometric Synthetic Aperture Radar Coherence Imagery for Land Surface Change Detection. Ph.D theses, University of London (2001)
10. Luo, X., Huang, F., Liu, G.: Extraction co-seismic Deformation of Bam earthquake with Differential SAR Interferometry. *J. New Zea. Inst. of Surv.* 296, 20–23 (2006)
11. Yang, J., Xiong, T., Peng, Y.: A fuzzy Approach to Filtering Interferometric SAR Data. *Int. J. of Remote Sens.* 28, 1375–1382 (2007)
12. Gens, R.: The influence of input parameters on SAR interferometric processing and its implication on the calibration of SAR interferometric data. *Int. J. Remote Sens.* 2, 11767–11771 (2000)
13. Anile, A.M., Falcidieno, B., Gallo, G., Spagnuolo, M., Spinello, S.: Modeling uncertain data with fuzzy B-splines. *Fuzzy Sets and Syst.* 113, 397–410 (2000)
14. Marghany, M.: Simulation of 3-D Coastal Spit Geomorphology Using Differential Synthetic Aperture Interferometry (DInSAR). In: Padron, I. (ed.) *Recent Interferometry Applications in Topography and Astronomy*, pp. 83–94. InTech - Open Access Publisher, University Campus STeP Ri, Croatia (2012)
15. Spagnolini, U.: 2-D phase unwrapping and instantaneous frequency estimation. *IEEE Trans. Geosci. Remote Sensing* 33, 579–589 (1995)
16. Davidson, G.W., Bamler, R.: Multiresolution phase unwrapping for SAR interferometry. *IEEE Trans. Geosci. Remote Sensing* 37, 163–174 (1999)
17. Marghany, M., Sabu, Z., Hashim, M.: Mapping coastal geomorphology changes using synthetic aperture radar data. *Int. J. Phys. Sci.* 5, 1890–1896 (2010)
18. Marghany, M.: Three-dimensional visualisation of coastal geomorphology using fuzzy B-spline of dinsar technique. *Int. J. of the Phys. Sci.* 6, 6967–6971 (2011)
19. RADARSAT International, RADARSAT application (online), <http://www.rsi.ca> (accessed March 20, 2011)
20. Zebker, H.A., Werner, C.L., Rosen, P.A., Hensley, S.: Accuracy of Topographic Maps Derived from ERS-1 Interferometric Radar. *IEEE Geosci. Remote Sens.* 2, 823–836 (1994)
21. Baselice, F., Ferraioli, G., Pascazio, V.: DEM Reconstruction in Layover Areas From SAR and Auxiliary Input Data. *IEEE Geosci. Rem. Sensing Letters* 6, 253–257 (2009)
22. Ferraiuolo, G., Pascazio, V., Schirinzi, G.: Maximum a Posteriori Estimation of Height Profiles in InSAR Imaging. *IEEE Geosci. Rem. Sensing Letters*, 66–70 (2004)
23. Ferraiuolo, G., Meglio, F., Pascazio, V., Schirinzi, G.: DEM Reconstruction Accuracy in Multichannel SAR Interferometry. *IEEE Trans. on Geosci. and Rem.* 47, 191–201 (2009)
24. Ferretti, A., Prati, C., Rocca, F.: Multibaseline Phase Unwrapping for INSAR Topography Estimation. *Il Nuov. Cimento.* 24, 159–176 (2001)

Three-Dimensional Coastal Front Visualization from RADARSAT-1 SAR Satellite Data

Maged Marghany

Universiti Teknologi Malaysia, Institute for Geospatial Science and Technology (INSTeG)
Johor Bahru, 81310 UTM, Skudai, Malaysia
magedupm@hotmail.com,
maged@utm.my

Abstract. Three-dimensional (3D) computer visualization has tremendous demands for complex phenomena studies. Coastal waters are considered as complex system because of they are dominated by complex system. In this regard, this study aims to present a method that is based on fuzzy B-spline to reconstruct 3D of coastal water phenomena such as front from two-dimensional RADARSAT-1 SAR data. In doing so, fuzzy B-spline algorithm is integrated with Volterra model and velocity bunching model. Volterra algorithm is used to determine the sea surface current along the front zone while velocity bunching model implemented to acquire the information about significant wave height. fuzzy B-spline reconstructed 3-D front with smooth graphic feature. Indeed, fuzzy B-spline tracked the smooth and rough surface. Finally, fuzzy B-spline algorithm can keep track of uncertainty with representing spatially clustered gradient of flow points across the front. In conclusion, the fuzzy B-spline algorithm can be used for 3-D front reconstruction with integration of velocity bunching and Volterra algorithm.

Keywords: Fuzzy B-spline algorithm, 3D reconstruction, Front, RADARSAT-1 SAR, velocity bunching, Volterra model, and 3-D.

1 Introduction

Natural phenomena that are imaged using remote sensing satellite data can be reconstructed in 3-D. This process can be accomplished either by active or passive methods. The active methods interfere with the reconstructed phenomena, either mechanically or radiometrically [4]. The radiometric methods reconstruct the 3-D from the reflected or backscattered information about the specific objects or phenomena. However, passive methods use a sensor to measure the radiance reflected or emitted by the object's surface to infer its 3-D structure [3][5]. 3-D reconstruction of natural phenomena plays tremendous role to understand a complex system such as the dynamic processes of coastal waters [1][6][24].

An ocean front is a boundary separating two masses into water of different densities, and is the primary cause of gradient change of physical ocean properties. The water masses separated by a front usually differ in temperature and salinity.

Fronts occur on a wide range of scales, starting with those formed within an estuary between inflowing water and the estuary water. Bowden [5] stated that the foam line is located at the surface convergence, the detritus line where buoyant objects are trapped by currents moving in opposite directions at the surface and near the interface and the colour front where upwelled light undergoes a distinct spectral shift approximately the steeply descending isopycnals [5][19].

According to above mention, remote sensing techniques are able to image front locations in large-scale ocean. Both thermal and microwave remote sensing techniques are good tools to identify front locations. For instance, satellite infrared imagery can image front locations because of their strong thermal signatures. Likewise, satellite visible bands are also cable to image fronts based on imaging different colors of the two water masses. Besides that, synthetic aperture radar (SAR) is also able to identify front as a result of abruptly changes of surface wave pattern across front led to exceedingly change cross backscatter of SAR data. In this regard, SAR images can sometimes be used to interpret frontal dynamics, including growth and decay of meanders [17][18][19]. Recently, Jiang et al., [8] exploited various remote sensing data. Satellite images obtained from the Advanced Very High Resolution Radiometer (AVHRR), the Moderate Resolution Imaging Spectroradiometer (MODIS), the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) and RADARSAT-1 SAR S1 mode data to study coastal water plume and front which is also captured in S1 mode data[9].

Consistent with Klemas [12] remote sensors utilize their dissimilarities in turbidity, color, temperature, or salinity from surrounding water environments, to detect and map fronts and plumes. Various remote sensors exploit to study fronts, which involve multispectral and hyperspectral imagers, thermal infrared (TIR) radiometers, microwave radiometers, and Synthetic Aperture Radar (SAR). These sensors are mounted on satellites and aircraft provide the spatial/temporal resolution and coverage needed for tracking plumes and fronts, including their high temporal and spatial variability. In this paper, we address how 3D front can be reconstructed from single SAR data (namely the RADARSAT-1 SAR) using integration of Volterra kernel [7], velocity bunching [20][21][22][23] and Fuzzy B-spline models [16][17]. There are about three hypothesis that examined are: (i) the use of Volterra model to detect front flow pattern in RADARSAT-1 SAR C_{HH} band; (ii) the use of velocity bunching model to acquire significant wave height from RADARSAT-1 SAR data; and (iii) to utilize fuzzy B-spline to remodel 3-D of front surface.

2 3-D Model of Front

Three algorithms are involved for 3-D front reconstruction: (i) velocity bunching; (ii) Volterra; and (iii) Fuzzy B-spline. Significant wave heights are simulated from RADARSAT-1 SAR image by using velocity bunching model. Fuzzy B-spline used significant wave height information to reconstruct 3-D front. Moreover, front flow pattern is modeled by Volterra model.

2.1 Velocity Bunching Model

The velocity bunching modulation transfer function (MTF) is the dominant component of the linear MTF for the ocean waves with an azimuth wave number (k_x). According to Alpers et al.,[2] and Vachon et al.,[20][21][22], the velocity bunching can contribute to linear MTF based on the following equation

$$M_v = \frac{R}{V} \omega \left[\frac{k_x}{k} \sin \theta + i \cos \theta \right] \tag{1}$$

where R/V is the scene range to platform velocity ratio, which is 111 s in the case of RADARSAT-1SAR image data, θ is RADARSAT-1 SAR image incidence angel (35^0-49^0) and ω is wave spectra frequency which equals to $2\pi/K$.

Estimation of Significant Wave Height from Velocity Bunching Spectra based on the azimuth cut-off arising from the velocity-bunching model [13], equation (1), the azimuth cutoff could be scaled by the standard deviation of the azimuth shift. Vachon et al.,[20][21][22] introduced a relationship between the variance of the derivate of displacement along the azimuth direction $\rho_{\zeta\zeta}$ and the standard deviation of the azimuth shift σ which were estimated from the velocity bunching spectra. This relationship was given by

$$\sigma = \sqrt{\rho_{\zeta\zeta}} \tag{2}$$

The relation between standard deviation of the azimuth shift σ and significant wave height H_s can be given by

$$\sigma = \left(\frac{R}{V}\right) \left(1 - \frac{\sin^2(\theta)}{2}\right)^{0.5} \left(\frac{k_x g}{8}\right)^{0.5} H_s \tag{3}$$

where k_x is the azimuth wave number, θ is RADARSAT-1 SAR image incident angle, R/V is the scene range to platform velocity ratio and g is the acceleration due to the gravity. Note that the mean wave period T_0 is equal to $2\pi(\langle\langle k_x \rangle\rangle g)^{-0.5}$.

According to Vachon et al., [22] the significant wave height H_s can be obtained:

$$H_s = 0.6(\rho_{\zeta\zeta})^{0.5} \left[\frac{1 + \theta^2 / 4}{R/V}\right] T_0 \tag{4}$$

where θ is the RADARSAT-1SAR incidence angle and equation 4 is used to estimate the significant wave height which is based on the standard deviation of the azimuth shift σ .

2.2 Volterra Model

In refereeing to Inland and Garello[7], Volterra series can be used to model nonlinear imaging mechanisms of surface current gradients by RADARSAT-1 SAR image. As result of that Volterra linear kernel is contained most of RADARSAT-1 SAR energy which used to simulate current flow along range direction. In reference to Inland and Garello [7], the inverse filter $G(v_x, v_y)$ is used since the kernel $H_{1y}(v_x, v_y)$ has a zero for (v_x, v_y) which indicates the mean current velocity should have a constant offset [7][15]. The inverse filter $G(v_x, v_y)$ can be given as

$$G(v_x, v_y) = \begin{cases} [H_{1y}(v_x, v_y)]^{-1} & \text{If } (v_x, v_y) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The range current velocity [23] $U_y(0, y)$ can be estimated by

$$U_y(0, y) = I_{RADARSAT-1SAR} \cdot G(v_x, v_y) \tag{6}$$

where $I_{RADARSAT-1SAR}$ is the frequency domain of RADARSAT-1 SAR image acquired by applying 2-D Fourier transform on RADARSAT-1 SAR image.

2.3 The Fuzzy B-Splines Method

Fuzzy B-spline concept has adopted from Anile et al., [2] and Anile [1] which shows excellent 3-D reconstruction stated by Marghany et al., [17] Considering significant wave height modeled by using velocity bunching and radar backscatter cross section across front, fuzzy numbers are created. Let us consider a function $f : h \rightarrow h$, of N fuzzy variables h_1, h_2, \dots, h_n . Where h_n is the global minimum and maximum values of significant wave heights. Additionally, the following must hold for each pair of confidence interval which define a number: $\mu \succ \mu' \Rightarrow h \succ h'$.

The construction begins with the same preprocessing to compress the measured significant wave height values into an uniformly spaced grid of cells. Then, a membership function is defined for each pixel which incorporates the degrees of certainty of radar cross backscatter.

In doing so, two basic notions of confidence interval and presumption level are considered [10]. A confidence interval is a real values interval which provides the sharpest enclosing range for significant wave height values. An assumption level μ - level is an estimated truth value in the [0,1] interval of significant wave height changes [1][2][17]. The 0 value suits to minimum knowledge of significant wave heights, and 1 to the maximum of significant wave height. A fuzzy number is then prearranged in the confidence interval set, each one related to an assumption level μ [0,1].

3 Data Set and Study Area

The RADARSAT-1 SAR fine mode data were acquired on March 26, 2004, over the coastline of Kuala Terengganu, Malaysia ($103^{\circ} 5' E$ to $103^{\circ} 9'E$ and $5^{\circ} 20' N$ to $5^{\circ} 27' N$) (Fig.1). The RADARSAT-1 SAR fine mode data are acquiring information using C band HH polarized of frequency 5.3 GHz. The swath width of RADARSAT-1 SAR fine mode sensor is 50 km, with the range resolution of 8-9 km. There are two numbers of looks for The RADARSAT-1 SAR and the incident angle of 35° - 49° [18].

The study area is situated in the South China Sea between $5^{\circ}21' N$ to $5^{\circ}25' N$, east coast of Peninsular Malaysia. Consistent with Marghany et al., [17] there are four seasons: the two monsoons and the two transitional inter-monsoon periods. The monsoon winds and tidal effects [17] affect the seas around Malaysia. The winds during the Northeast monsoon are normally stronger than the Southwest monsoon [16]. The Accompanying waves are with a height that exceeds 3 m [16]. The bathymetry near the area has gentle slopes with 40 m water depth. A clear feature of this area is the primary hydrologic communications between the estuary and the South China Sea (Fig.1). As stated by Marghany et al., [18] this estuary is the largest estuary along the Terengganu coastline.

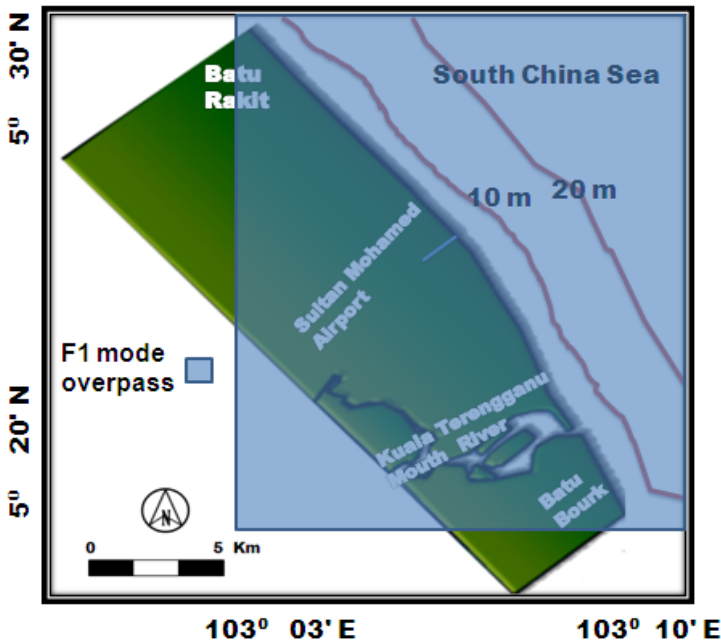


Fig. 1. RADARSAT-1 SAR F1 mode data passover study area

4 Results and Discussion

The methods described above have tested on single RADARSAT-1 SAR F1 mode data that is acquired along the coastal water of Kuala Terengganu Malaysia. The RADARSAT-1 F1 mode backscatter cross-section of front (Fig.2) has a maximum value of -21.25 dB. It is known the maximum backscatter value of 0.33 dB is found across the brightness frontal line. Moreover, the variation of radar backscatter cross-section is due to the current boundary gradient. According to Vogelzang et al. [23], ocean current boundaries are often accompanied by the changes in the surface roughness that can be detected by SAR. These surface roughness changes are due to the interaction of surface waves directly with surface current gradients. These interactions can cause an increase in the surface roughness and radar backscatter [13][17].

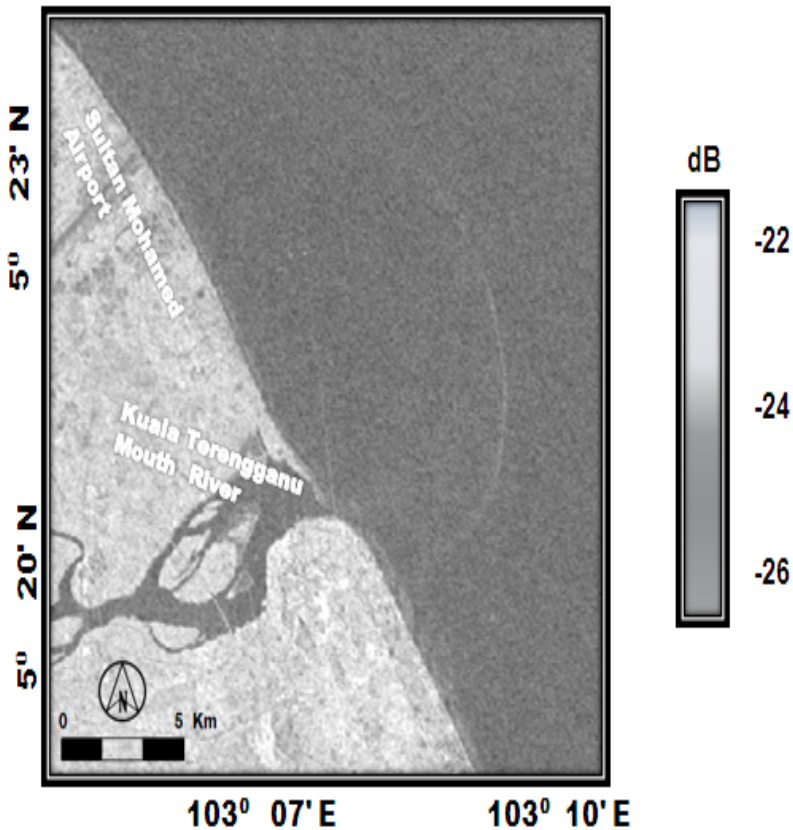


Fig. 2. Backscatter variations in F1 mode data

Fig. 3 shows 3-D front reconstruction with significant wave heights, and current variations cross front. Fig. 4 shows that significant wave variation cross front with maximum significant wave height of 1.2 m and gradient current of 0.9 m/s. March represents the northeast monsoon period as coastal water currents in the South China Sea tend to move from the north direction [17]. Nevertheless, Fig. 3 shows a meander current with southward direction. In fact, this current is created because of the water inflow from Kuala Terengganu Mouth River.

Furthermore, Marghany et al., [17] quoted that strong tidal current is a dominant feature in the South China Sea with maximum velocity of 1.5 m/s. Clearly, 3-D front coincides with water depth range between 10 to 20 m (Fig. 4). This indicates shallow water where the strong tidal stream (Fig. 3) that causes vertical mixing.

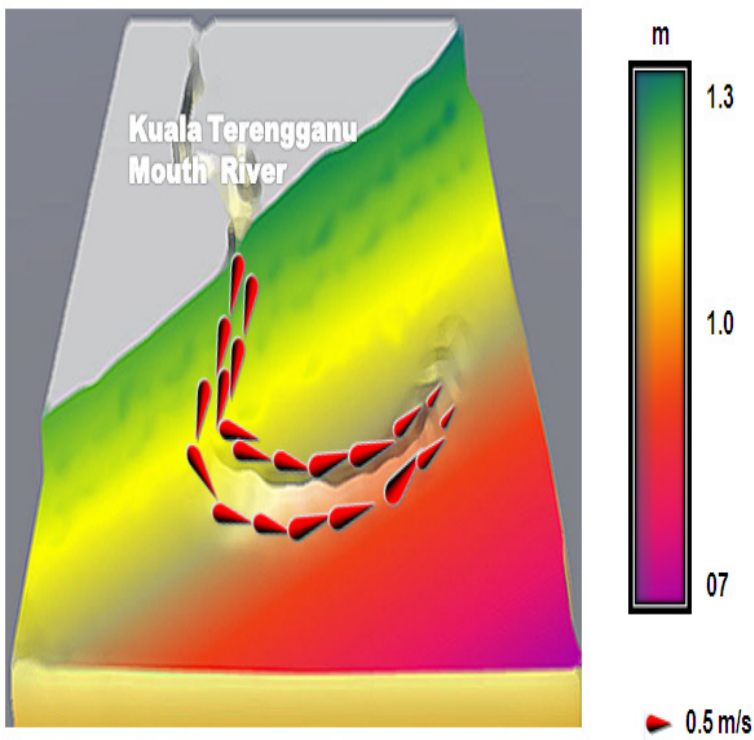


Fig. 3. 3-D Front Reconstruction with Significant Wave Height (H_s) and Surface Current Variations (U_y)

The visualization of 3-D front is sharp with the RADARSAT-1 SAR C_{HH} band because of each operations on a fuzzy number becomes a sequence of corresponding operations on the respective μ and μ' -levels, and the multiple occurrences of the same fuzzy parameters evaluated as a result of the function on fuzzy variables [2]

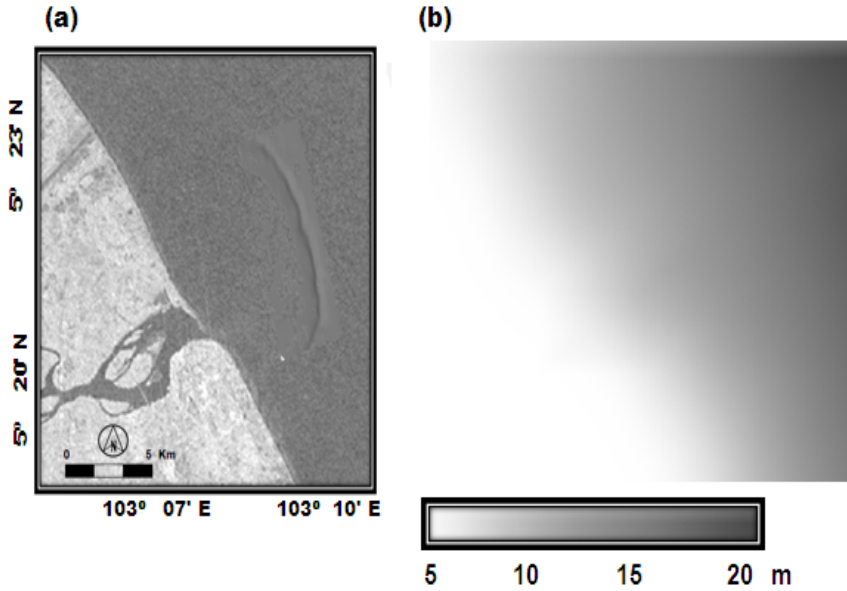


Fig. 4. F1 mode data for (a) 3-D front and (b) coastal bathymetry

[17]. Typically, in computer graphics, two objective quality definitions for Fuzzy B-splines were used: triangle-based criteria and edge-based criteria. Triangle-based criteria follow the rule of maximization or minimization, respectively, the angles of each triangle. The so-called max-min angle criterion prefers short triangles with obtuse angles. In addition, the fuzzy B-spline depicts optimize a locally triangulation between two different points [1][4][11][14]. This corresponds to the feature of deterministic strategies of finding only sub-optimal solutions usually which overcomes uncertainties. In this context, the spatial cluster of gradient flow at each triangulation points can simulated (Fig. 3).

Consequently, triangle-based criteria follow the rule of maximization or minimization, respectively, of the angles of each triangle [14] which prefers short triangles with obtuse angles. Further, edge-based criteria prefer edges are closely related. This study confirms the previous studies of Anile et al., [2]; Fuchs et al., [14]; Marghany et al., [17]. Indeed, these studies have agreed that fuzzy B-spline algorithm is an accurate tool for 3-D surface reconstruction from 2-D data.

5 Conclusions

This work has demonstrated method to reconstruct 3-D coastal front in RADARSAT-1 SAR F1 mode data. Three algorithms of velocity bunching, Volterra and fuzzy B-spline are involved to reconstruct 3-D coastal front. The velocity bunching algorithm modeled significant wave height, Volterra algorithm simulated coastal current

movement while fuzzy B-spline implemented the significant wave height to reconstruct 3-D coastal front. The study shows fuzzy B-spline reconstructed 3-D front with smooth graphic feature. Indeed, fuzzy B-spline algorithm can keep track of uncertainty with representing spatially clustered gradient of flow points across the front. In conclusion, the fuzzy B-spline algorithm can be used for 3-D front reconstruction with integration of velocity bunching and Volterra algorithm.

References

1. Adeyemo, J., Fred, O.: Optimizing planting areas using differential evolution (DE) and linear programming (LP). *International Journal of Physical Sciences* 4, 212–220 (2009)
2. Alpers, W.R., Ross, D.B., Rufenach, C.L.: On the detectability of ocean surface waves by real and synthetic aperture radar. *Journal Geophysical Research* 86, 6481–6498 (1981)
3. Anile, A.M.: Report on the activity of the fuzzy soft computing group. Technical Report of the Dept. of Mathematics, University of Catania, 10 (March 1997)
4. Anile, A.M., Deodato, S., Privitera, G.: Implementing fuzzy arithmetic. *Fuzzy Sets and Systems*, 72 (1995)
5. Bowden, K.F.: *Physical oceanography of coastal waters*. Ellis Horwood Ltd., England (1993)
6. Guillermo, D.L.T., Soto-Zarazúa, G.M., Guevara-González, R.G., Enrique, R.G.: Bayesian networks for defining relationships among climate factors. *International Journal of the Physical Sciences* 6, 4412–4418 (2011)
7. Inglada, J., Garello, R.: Depth estimation and 3D topography reconstruction from SAR images showing underwater bottom topography signatures. In: *Proceedings of IGARSS 1999* (1999)
8. Jiang, L., Yang, X.H., Klemas, V.: Remote sensing for the identification of coastal plumes: case studies of Delaware Bay. *International Journal of Remote Sensing* 30, 2033–2048 (2009)
9. Johannessen, J.A., Shuchman, R.A., Digranes, G., Lyzenga, D.R., Wackerman, C., Johannessen, O.M., Vachon, P.W.: Coastal ocean fronts and eddies imaged with ERS 1 synthetic aperture radar. *Journal of Geophysical Research* 101(C3), 6651–6667 (1996)
10. Hassasi, N., Saneifard, R.: A novel algorithm for solving fuzzy differential inclusions based on reachable set. *International Journal of the Physical Sciences* 6, 4712–4716 (2011)
11. Khadijeh, M., Motameni, H., Enayatifar, R.: New method for edge detection and de noising via fuzzy cellular automata. *International Journal of Physical Sciences* 6, 3175–3180 (2011)
12. Klemas, V.: *Remote Sensing of Coastal Plumes and Ocean Fronts: Overview and Case Study*. *Journal of Coastal Research* (2011) (in Press)
13. Krogstad, H.E., Schyberg, H.: On Hasselman's nonlinear ocean –SAR transformation. In: *Proc. of IGARSS 1991*, held at Espo, Finland, June 3-6, pp. 841–849 (1991)
14. Fuchs, H., Kedem, Z.M., Useton, S.P.: Optimal Surface Reconstruction from Planar Contours. *Communication of the ACM* 20, 693–702 (1997)
15. Majid, K., Gondal, M.A.: An efficient two step Laplace decomposition algorithm for singular Volterra integral equations. *International Journal of the Physical Sciences* 6, 4717–4720 (2011)
16. Marghany, M., Mazlan, H., Cracknell, A.P.: 3-D visualizations of coastal bathymetry by utilization of airborne TOPSAR polarized data. *International Journal of Digital Earth* 3, 187–206 (2010)

17. Marghany, M., Mazlan, H.: Simulation of sea surface current velocity from synthetic aperture radar (SAR) data. *International Journal of the Physical Sciences* 5, 1915–1925 (2010)
18. Robinson, I.S.: *Satellite Oceanography: An Introduction for Oceanographers and Remote – sensing Scientists*. Johan Wiley & Sons, New York (1994)
19. Vachon, P.W., Krogstad, H.E., Paterson, J.S.: Airborne and spaceborne synthetic aperture radar observations of ocean waves. *Atmosphere-Ocean* 32, 83–112 (1994)
20. Vachon, P.W., Liu, A.K., Jackson, F.C.: Near-shore wave evolution observed by airborne SAR during SWADE. *Atmosphere-Ocean* 2, 363–381 (1995)
21. Vachon, P.W., Campbell, J.W.M., Dobson, F.W.: Comparison of ERS and RADARSAT SRS for wind and wave measurement. Paper Presented at third ERS Symposium ESA, held at Florence, Italy, March 2-18 (1997)
22. Vogelzang, J., Wensink, G.J., Calkoen, C.J., van der Kooij, M.W.A.: Mapping submarine sand waves with multiband imaging radar, 2, Experimental results and model comparison. *Journal of Geophysical Research* 102, 1183–1192 (1997)
23. Zaki, M.S.: On asymptotic behaviour of a second order delay differential equation. *International Journal of Physical Sciences* 2, 185–187 (2007)

A New Self-Learning Algorithm for Dynamic Classification of Water Bodies

Bernd Fichtelmann and Erik Borg

German Aerospace Center, German Remote Sensing Data Center, Kalkhorstweg 53,
17235 Neustrelitz, Germany
{Bernd.Fichtelmann, Erik.Borg}@dlr.de

Abstract. In many applications of remote sensing data land-water masks play an important role. In this context they can be a helpful orientation to distinguish dark areas (e.g. cloud shadows, topographic shadows, burned areas, coniferous forests) and water areas. However, water bodies cannot always be classified exactly on basis of available remote sensing data. This fact can be caused by a variety of different physical and biological factors (e.g. chlorophyll, suspended particles, surface roughness, turbid and shallow water and dynamic of water bodies) as well as atmospheric factors (e.g. haze and clouds). On the other hand the best available static water masks also show deficiencies. These are essentially caused by the fact that land-water masks represent only a temporal snapshot of the water bodies distributed worldwide and therefore these masks cannot reflect their dynamic behavior. This paper presents a dynamic self-learning water masking approach for AATSR remote sensing data in the context of integrating high-quality water masks in processing chains for deriving value-added remote sensing data products. As an advantage to conventional water masking algorithms, the proposed approach operates on basis of a static water mask as data base for deriving an optimized dynamic water mask. Significant research effort was spent to develop and validate a dynamic self-learning algorithm and a processing scheme for operational derivation of actual land-water masks as basis for operational interpretation of remote sensing data. Based on this concept actual activities and perspectives for contributions to operational monitoring systems will be presented.

Keywords: self-learning algorithm, land-water mask, interpretation, remote sensing, cloud cover.

1 Introduction

Satellite Earth observation is a major data source for analyzing environmental subjects. The full-coverage description of status and dynamics of ecological systems is in many cases subject of environmental investigations which deal with sustainable use of natural resources. But in many cases, the actual data base is fragmentary in the required scale [9], [15], [11].

It is not disputed that land-water masks can be helpful additional information for the automated and operational interpretation of remote sensing data. Carroll et al. [3] gives a summary of developments of land-water masks since 1996 and they show

possibilities of additional improvements in global land-water masks. The data can be available in the raster or vector format. The spatial resolution of the data is between 90 m to 25 km. Due to the different requirements of thematic applications and the resulting data management, there are different data sets in different spatial resolution. For example the latest version of the GSHHS data (Global Self-consistent Hierarchical High-resolution Shorelines) of the National Geophysical Data Center [16], released in 2011, is available in a spatial resolution <100 m. At the moment the best available land-water mask is the SRTM Water Body Detection (SWBD) with an accuracy better than 30 m for included water bodies in the geographical region between 54° South and 60° North. Caused by the limited temporal duration (only 11 Days in February 2000) of the SRTM mission, the delivered mission coverage includes data gaps in the data set. Carroll et al. [3] describes according to personal information from the SWDB team in 2006, that the team has tried to infill “these gaps with help of Landsat Geocover data”. However, if the Geocover data were too cloudy, then the appropriate gaps could not be filled.

The preparation of an exact as possible land-water mask aims for example at minimization of inclusion of water pixels in thematic interpretation algorithms for land applications or vice versa. Numerous endeavors exist to improve the quality of global land-water masks, since there is an increasing interest to use such a database in evaluation of remote sensing data.

For this reason the corresponding land-water distribution is an additional information layer of e.g. AATSR and MERIS data delivered to the users. Borg and Fichtelmann [2] suggested a procedure for automatic derivation of data usability of remote sensing data which also includes a water mask in the production process of LANDSAT/ETM+ data.

Available land-water masks are temporal snapshots. Therefore, the most important deficit of these masks is the fact that they cannot reflect the dynamic behavior of water bodies. That means land-water masks are a static information layer.

To counteract possible misinterpretations and to support the land-water classification of remote sensing data, a self-learning procedure was developed that uses available static land-water mask. In a first processing step, the water pixels of the mask will be regarded only as candidates for water. In a second processing step several classification algorithms are used as decision support to classify water. Like before, the water pixels resulting from this second processing step are regarded as possible candidates only. The partial results of the static mask and the different classification mechanisms are fused to an overall result in a third processing step.

Generally, the method is adaptable to other optical sensors. First results of the method applied to AATSR data are shown and discussed.

2 Material and Methods

2.1 Remote Sensing Data

The demonstration of the algorithm is based on multispectral data sets of the Advanced Along-Track Scanning Radiometer (AATSR). This is one of the

instruments on board the ENVISAT satellite. The ground resolution of AATSR data at nadir is 1 km. The AATSR-sensor measures reflected and emitted radiation at the centre wavelengths of 0.55 μm , 0.66 μm , 0.87 μm , 1.6 μm , 3.7 μm , 11 μm , and 12 μm . For the investigations the reflectance ρ of bands at 0.55 μm , 0.66 μm , 0.87 μm , 1.6 μm , and for the land surface temperature BT11 the band at 11 μm were used. Additionally two layers with information on latitude and longitude are necessary. The used image sections of complete data sets include all available 512 columns. The number of lines varies between 1000 and 7500. The algorithm was tested on regions of different degrees of difficulties: the Alps with terrain shadow, Scandinavia with inaccuracies of geometry in static mask and the region around Caspian Sea with partly strong changes (desiccation) of water bodies.

2.2 Available Static Land-Water Masks

The generation of a consistent static land-water mask is based on use of different global land-water masks of different spatial resolution and feature accuracy as CIA World-Map or SRTM Water Body Data (SWBD). A short description of global data sets used in these studies is given in the following.

CIA World-Map: As add on of the development software IDL (Interactive Data Language) the 1993 CIA World map database [12], or World DataBank II, is available for operational processing based on USGS map accuracy standards [14].

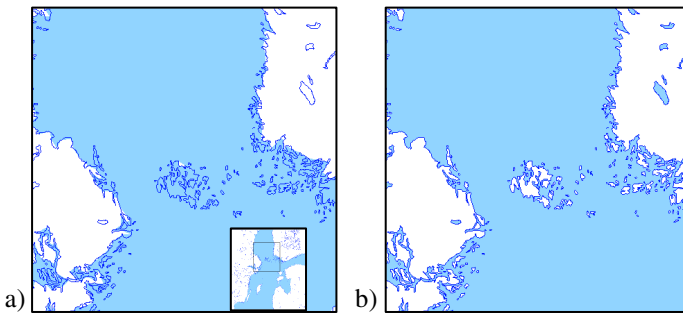


Fig. 1. CIA world map data include only raw land-water distribution (a) and the result after correction (b) for a part of the Baltic Sea around 60° North

But this add-on includes the disadvantage that the land-water distribution as a whole is not provided with the quality of the coastline information. In some cases continental lakes are not represented as water, in other cases islands are represented as water (Fig. 1a). Based on an object analysis it is possible to combine the information "water" or "land" with the corresponding objects which are embedded by coastlines (Fig. 1b).

SRTM Water Body Data (SWBD): Beside the documentation [13] this data set consists about 12,229 files, covering the Earth between 60° (54°) South and 60°

North. For land cells, which are not available in the data set, a data dummy had to be produced. Fig. 2 shows South-America with missing cells in the SWBD data base on the left side and with cells (dummies) infilling these gaps with the information land.

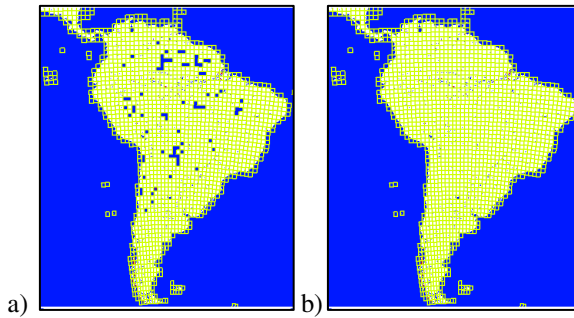


Fig. 2. Available SRTM cells for South America are presented in the left image. In the right image the missing cells were substituted by cells with land information.

3 Dynamically Self-Learning Evaluation Method (DySLEM)

The Dynamically Self-Learning Evaluation Method is based on the use of regional parameters. Therefore, the pre-processing module selects image frames of the complete input-data set. With respect to AATSR these frames have a defined size of 512 x 512 pixels. The data and auxiliary data are used to initiate the DySLEM processor. After finishing of DySLEM a post-processor produces the output product.

3.1 Structure of the Processor

The operational determination of dynamic land-water mask includes three processing steps. The processor structure is shown in Figure 3. In the following a short description of the processor, the used methods, and procedures is given.

Step 1: The work step WS1 generates a regional static land-water mask for selected remote sensing data. For this, the processor uses data of available global static land-water masks. The result is a mask of the percentage water content within image pixel. With regard to the dynamic of water bodies the identified water pixels are regarded only as candidates for water pixels.

Step 2: The work step WS2 includes two different sub-processors for identifying all “candidates” for water. The classification of water is based on spectral properties, relations between different spectral bands or the vegetation index of water bodies. Beside of water the results can include other dark regions.

Step 3: The aim of the work step WS3 is the data fusion of land-water masks processed in WS1 and WS2. On regional level, the data fusion is based at first on the

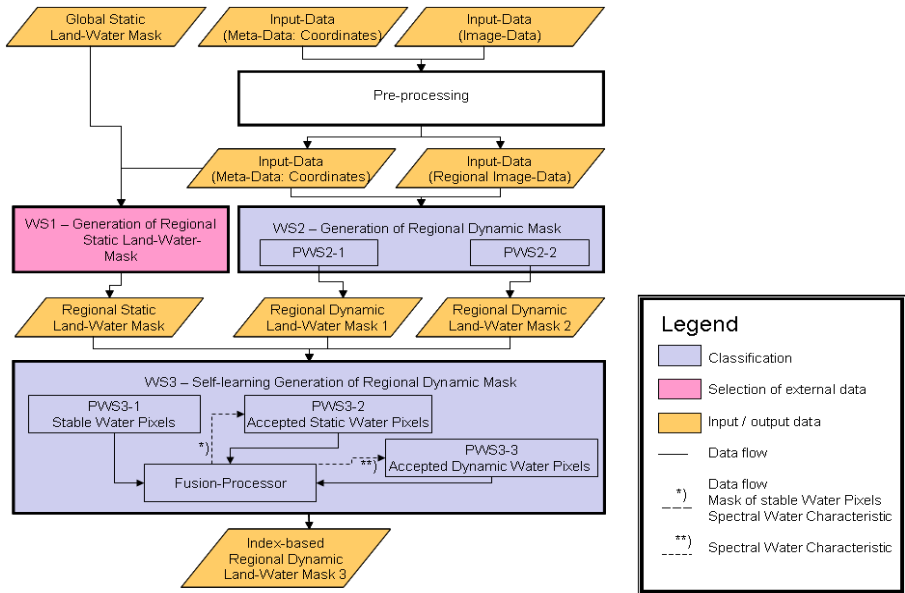


Fig. 3. Processing chain of the processor

water objects reliably identified by part 1 of WS3 (PWS3-1), which were regarded as candidate water pixels in static (WS1) as well as in dynamic masks (WS2). Corresponding water pixels are signed by the fusion processor in a new intermediate mask. Using the spectra of these identified water pixels a mean spectrum is determined. For all other water pixels in the static land-water mask the Fusion Processor initiates sub-processor PWS3-2. This processor tests the derived mean spectrum versus all remaining pixel spectra. Fulfilling this relation, candidate pixels of static mask will be accepted and labelled in the resulting mask and all non-accepted pixels are excluded from further processing. Next the Fusion Processor initiates a third sub-processor PWS3-3. Pixel candidates which are accepted as water in WS2 and are not accepted by sub-processor PWS3-1 will be tested a second time by the sub-processor PWS3-3. Pixels which are defined as water pixels by processor PWS3-3 according to spectral behaviour will be masked as accepted dynamic water and labelled by the Fusion Processor in the final regional dynamic land-water mask.

3.2 Generation of Regional Static Land Water Mask (WS1)

The objective of the first work step (WS1) is the generation of a regional static land water mask **lwms** (**l**and **w**ater **m**ask, **s**tatic, **s**ection). Therefore, the pre-processing separates AATSR frames of 512 x 512 pixels from the data stream. On basis of corresponding corner coordinates a first map template can be constructed for an area equivalent projection of the AATSR frame into this map [5]. The ground resolution of the static land-water mask (**lwms**) (<100 m) is higher than the ground resolution of

AATSR-data (1000 m). This fact allows a sub-pixel calculation of water within a pixel in percentage. A shoreline represented by digital pixels cannot present the real analogous land-water distribution. Therefore, a generated second map template of higher resolution (1 original pixel to 9 x 9 sub-pixels) is of advantage, for example. After this preparation the information of available global land-water masks is filled in the prepared template. Therefore, the land-water distribution represented by the 81 sub-pixel environment is more precise than that of the covering AATSR pixel of lower spatial resolution. This processing step allows more realistic information of water area in the AATSR pixel in percentage. Additionally to this consideration a higher degree of flexibility of the center coordinates of the AATSR pixel with respect to the real geographical situation is given.

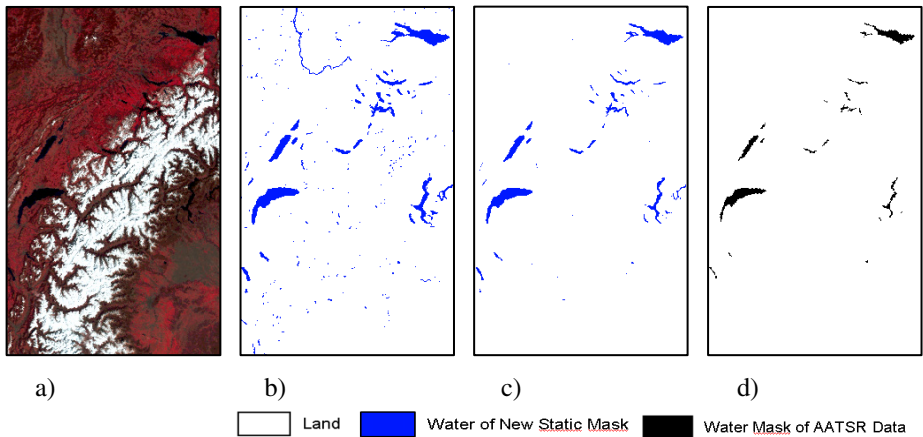


Fig. 4. Section from a 12th March 2007 AATSR scene of the western Alps (a), the corresponding static masks with different limits of water content per pixel area (b) $\geq 10\%$, (c) $\geq 60\%$ and the land mask (d), included in AATSR data set

After calculation of the water area in the 9 x 9 sub-pixel environment the water area information has to be transformed to the corresponding AATSR pixel. This mask *lwms* is basic information to estimate the land-water distribution on sub-pixel level of AATSR-data between 0 and 100 percentages. Fig. 4 gives an impression of the *lwms* mask quality for the western region of the Alps with different predefined minima thresholds of water content in comparison with the original AATSR data. Advantage consists in the fact, that many static water pixels can be identified based on only sub-pixel information.

3.3 Classification Algorithms of DySLEM

The classification of water is difficult because the spectral characteristic of water bodies can vary significantly. There are many well-known reasons: different physical and biological factors (e.g. chlorophyll, suspended particles, surface roughness, turbid and shallow water and dynamic of water bodies) as well as atmospheric factors

(e.g. haze and clouds). Additionally, the observed area of a pixel can consist of an unknown ratio between water and land as discussed before.

Generally the best classification includes always a probability of incompleteness and/or inaccurateness. The pixels classified as dynamic water are considered as candidates. For final decision the self-learning algorithm of the method in WS3 will be used. For these studies two different algorithms for dynamic land-water mask were used to identify all possible different types of water within the section (lwmds).

PWS2-1-Generation of lwmd1s: The bands ρ_{870} , ρ_{670} , ρ_{550} , and ρ_{1600} are the calibrated reflectance of input-data. The decrease of reflectance ρ with increase of wavelength, the reflectance ρ_{550} versus the threshold 0.22, and the surface temperature BT11 versus threshold 273 K are to be checked according [6]. The result is mapped into the first land-water mask (lwmd1-AATSR).

Rule (equation 1) for lwmd1s-AATSR:

$$(\rho_{550} > \rho_{670}) \wedge (\rho_{670} > \rho_{870}) \wedge (\rho_{870} > \rho_{1600}) \wedge (\rho_{550} < 0.22) \wedge (BT11 > 273) ? \cdot 1 : 0 \quad (1)$$

PWS2-2-Generation of lwmd2s: The second algorithm is based on an unpublished work at the Rutherford Appleton Laboratory by Stevens, A.D. (in [1]).

He developed a pixel-based classification scheme using NDVI (Equation 2) and additionally a NDVI-like index NDI2 (Equation 3) for pixel-by-pixel classification based on pre-defined classification criteria. The algorithm was developed to classify clouds. It is also applicable to classify land use and additional classes as water.

$$NDVI = (\rho_{870} - \rho_{670}) / (\rho_{870} + \rho_{670}) \quad (2)$$

$$NDI2 = (\rho_{670} - \rho_{550}) / (\rho_{670} + \rho_{550}) \quad (3)$$

Stevens uses both indices to define a two-dimensional classification space. Plotting NDVI versus NDI2 for each pixel, different land use types forming different clusters can be identified. When adapting the different surface types including water by the following algorithm, a result lwmd2s comparable with the algorithm before can be derived.

Rule (equation 4) for lwmd2s-AATSR:

$$((NDI2 < 0.1) \wedge (NDVI < -0.15)) \vee ((NDI2 < 0.0) \wedge (NDVI < 0.0) \wedge (NDI2 < ((NDVI + 0.025) / 1.25))) ? \cdot 1 : 0 \quad (4)$$

The results of the two masks lwmd1 and lwmd2 are presented in Fig. 5b and c for the same region as in Fig. 4a.

3.4 A Self-Learning Algorithm to Identify Temporal Dynamic of Water Bodies

The different independently operating sub-algorithms of the self-learning algorithm allow the generation of different land-water masks. In consequence, these results can be rule-based selected or merged.

The third work step (WS3) includes three sub-classification processors and a fusion processor. The three masks *lwms*, *lwmd1s* and *lwmd2s* of the same frame are available.

The aim of the first processing step (PWS3-1) is the determination of stable water pixels. As such all pixels are defined which are included in the static mask as well as identified as water in at least one of the two dynamic land-water masks. Based on static land-water mask *lwms* containing the land-water ratio between 0 und 100 percentage, all stable water pixels with more than 60 percentage water content can be determined following the assumption according equation 5.

$$(lwmd1s = 1 \vee lwmd2s = 1) \wedge lwms > 60\% : 1 : 0 . \tag{5}$$

All accepted static water pixels are stored by the Fusion Processor in the final matrix *lwmd3s*. Corresponding water bodies are shown in Figures 5d, 6f, and 7g. The probability of the existence of water is very high in this case.

But it may be that clouds, haze, or ice mask further water pixels in the image data. In case of haze and thin clouds an inclusion of water objects in the resulting mask is useful. In other cases water pixels of the static land-water mask detected by AATSR as dry have to be excluded from further processing. Such a decision is task of PWS3-2.

Based on all *n* pixels with *lwmd3s* = 1 the fusion processor calculates generally a regional mean spectrum (all spectral bands *i*) in preparation of PWS3-2:

$$\bar{\rho}_i = \left(\sum_1^n \rho_i(lwmd3s = 1) \right) / n . \tag{6}$$

For AATSR sensor only a regional mean temperature was calculated for BT11 on basis of this equation.

Further preparation includes the identification of all pixels of the static mask with a water content $\geq 10\%$ which are not included in *lwmd3s*. Thus all pixels already marked as stable water will be excluded. After that the Fusion Processor initiates the second step (PWS3-2) of the self-learning algorithm. In case of AATSR data the sub-processor uses the mean surface temperature for testing against the temperature of all water pixels identified by the Fusion Processor.

For improving the results an offset of 5 K is used for (Equation 7). The additionally identified water pixels are encoded with 2 in the resulting mask *lwmd3s*.

$$lwmd3s \neq 1 \wedge lwms \geq 10 \wedge BT11s > 273 \wedge BT11s \leq MEAN(BT11s(lwmd3s = 1) + 5) : 2 : 0 . \tag{7}$$

In regions with local shift between static mask and image data results of Equation 7 have shown that vegetation pixels are partly identified as water. Therefore, the inclusion of an additional relation with NDVI is necessary.

As first definite criterion $NDVI < -0.04$ is used for identifying water pixels which are before identified as candidates of static water. That means Equation 8 allows an

additional adjustment of the results before. The concerning pixels are marked in the resulting mask $lwmd3s$.

$$\begin{aligned}
 &lwmd3s \neq 1 \wedge lwms \geq 10 \wedge BT11s > 273 \wedge \\
 &BT11s \leq MEAN(BT11s(lwmd3s = 1) + 5) \wedge NDVI < -0.04 \quad ? : 2 : 0 \quad .
 \end{aligned} \tag{8}$$

In contrast, the second less definite criterion $NDVI < +0.15$ allows to accept static water pixels of lower probability.

$$\begin{aligned}
 &lwmd3s = 0 \wedge lwms \geq 10 \wedge BT11s > 273 \wedge \\
 &BT11s \leq MEAN(BT11s(lwmd3s = 1) + 5) \wedge NDVI < +0.15 \quad ? : 3 : 0 \quad .
 \end{aligned} \tag{9}$$

The results of Equation 8 and Equation 9 shown in Fig. 5d as part of final mask show that pixels of shorelines of different lakes can be identified. Parts of Lake Constance which are masked by haze can be identified as water, too. The principle of the algorithm adapted from Stevens (equation 4) has shown that the restrictions of NDVI are too strong for quality control of static water mask. The use of the relation with NDVI (equation 9) has demanded the additional use of corresponding relations with $NDI2$, given in equation 10 and equation 11.

$$\begin{aligned}
 &lwmd3s = 0 \wedge lwms \geq 10 \wedge BT11s > 273 \wedge \\
 &BT11s \leq MEAN(BT11s(lwmd3s = 1) + 7) \wedge NDI2 < -0.15 \wedge . \\
 &(\rho_{1600} - \rho_{550}) < 0.03 \wedge \rho_{870} < 0.17 \quad ? : 4 : 0 \quad .
 \end{aligned} \tag{10}$$

$$\begin{aligned}
 &lwmd3s = 0 \wedge lwms \geq 10 \wedge BT11s > 273 \wedge \\
 &BT11s \leq MEAN(BT11s(lwmd3s = 1) + 7) \wedge NDI2 < 0.00 \wedge . \\
 &(\rho_{1600} - \rho_{550}) < 0.03 \wedge \rho_{870} < 0.17 \quad ? : 5 : 0 \quad .
 \end{aligned} \tag{11}$$

Based on these relations, rivers or shallow water can be effectively identified as accepted static water. After this processing step all pixels of the static water mask with a water area $\geq 10\%$ pixel coverage have been examined. Pixels which satisfy the criteria in any form are included by the Fusion Processor in mask $lwmd3s$.

Subsequently all pixels of dynamic land-water masks $lwmd1s$ and $lwmd2s$ and which are not marked in the static land-water mask ($lwms \geq 10$) will be determined by the Fusion Processor on basis of equation 12. The resulting pixels are marked as “candidate” in the intermediate result mask $lwmdis$, initiating the next sub-processor.

$$lwms < 10 \wedge lwmd1s = 1 \wedge lwmd2s = 1 \quad ? : 1 : 0 \quad . \tag{12}$$

The sub-processor PWS3-3 calculates the second dynamical effect of the mask for dynamic water bodies which are not available in the static water mask. By reason that shadow pixels can likewise fulfil these conditions, an exclusion of these pixels from the final mask is necessary. For shadow pixels the difference of reflectance from band to band is smaller than for water pixels. Therefore for exclusion shadow pixels the

following parameters have to be modified in processing. The following procedure (Equation 13) can be applied [7]:

$$\begin{aligned} \rho_{550}(lwmdis = 1) - \rho_{670}(lwmdis = 1) > 1.0 \wedge \\ \rho_{670}(lwmdis = 1) - \rho_{870}(lwmdis = 1) > 0.8 \wedge \\ \rho_{870}(lwmdis = 1) - \rho_{1600}(lwmdis = 1) > 1.0 \cdot 6 : 0 \end{aligned} \quad (13)$$

Only pixels which fulfil this relation are encoded in mask *lwmd3s* with 6 as accepted dynamic water pixel. The work step WS3 will be finished with the DySLEM-output *lwmd3s*.

After *n*-runs of DySLEM the *n* subsets will be integrated into a complete final mask *lwmd3* by equation 14.

$$lwmd3 = \sum_1^n lwmd3s \quad (14)$$

The other masks (*lwms*, *lwmd1s*, *lwmd2s*) can be combined in the same way.

4 Results

Decisive advantages of the proposed procedure are the identification of difficult classifiable water pixels (see chapter 3.3) and the identification of both "wrong" water pixels caused by data quality problems (e.g. insufficiently accurate geo-correction of the mask) and of water pixels of the static land-water mask which changed their spectral properties after the preparation of the static mask data base (e.g. dried areas).

Based on the following exemplarily discussed results the efficiency and the stability of the proposed procedure will be demonstrated. The selected images include terrain shadows (Alps region), dry or shallow water bodies (region around the Caspian Sea) and qualitative limited static water mask (Scandinavia).

The RGB-image (Fig. 5a) uses the bands ρ_{1600} , ρ_{870} , and ρ_{550} for a better visibility of dark regions than in Fig. 4a. The visual comparison of RGB-image (Fig. 5a) and final mask (Fig. 5d) shows a good agreement of identified pixels. Fig. 5d demonstrates also that the largest proportion of the water area in the image is identified on basis of both the stable land-water information (compare Fig. 4c) and the dynamic water information. Fig. 5e and 5f show in more detail sections of river Rhine and Lake Constance. The identified pixels of River Rhine (brown colored) are based only on the results of Equation 13. It can also be seen that terrain shadows (Fig. 5c) in the centre of the image can be suppressed (see Fig. 5d).

The comparison of the classification results including terrain shadows (Fig. 5c) with the resulting mask in Fig. 5d shows that the shadow information is eliminated by processing step PWS3-3. In contrast to this, in most cases the cloud shadows in images are no problem for the correct classification because they are already eliminated in processing step WS2.

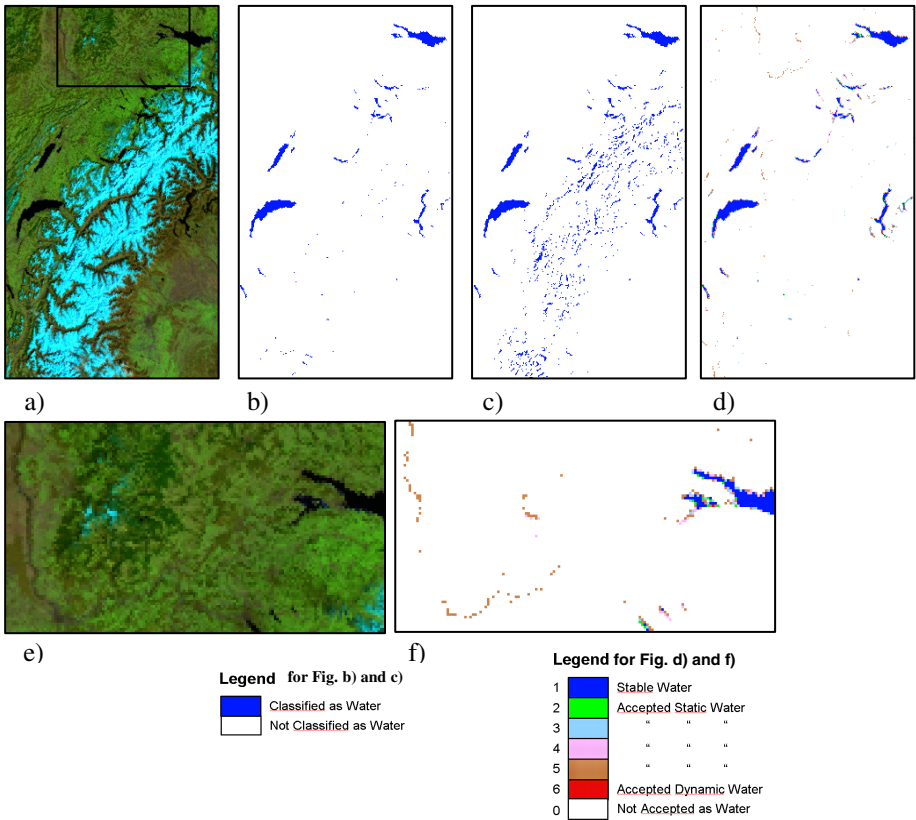


Fig. 5. RGB-image (ρ_{1600} , ρ_{870} , ρ_{550}) (a), the masks lwmd1 (b), lwmd2 (c), lwmd3 (d) and e) the subset of a) with parts of River Rhine and Lake Constance and f) the corresponding subset of d)

For understanding the behaviour of the classification algorithm based on static land-water mask it is necessary to look at Fig. 6. Fig. 6a shows on the right side, upper part of the image a large, elongated lake. In comparison to Fig. 6a this lake is correct marked as water in the static land-water mask ($lwms \geq 10$) (Fig. 6b), but the spatial dimension of the lake is too large.

The preview to the final mask (Fig. 6f) shows the mapped lake in its real spatial dimension. This fact exemplarily discussed for this lake is relevant for many other lakes of the static mask. These results are consistent to the RGB image.

In the same figure (right site, below) it can be seen that water pixels of Caspian Sea having different optical properties (probably caused by glint) can be identified nevertheless as water using the static land-water information (Fig. 6b, Fig. 6c) as well as using the dynamic classification steps of WS3-1 and WS3-2. Thus, based on the static land-water mask it is possible to identify water pixels in cases of changed optical or spectral characteristic (e.g. ice or haze).

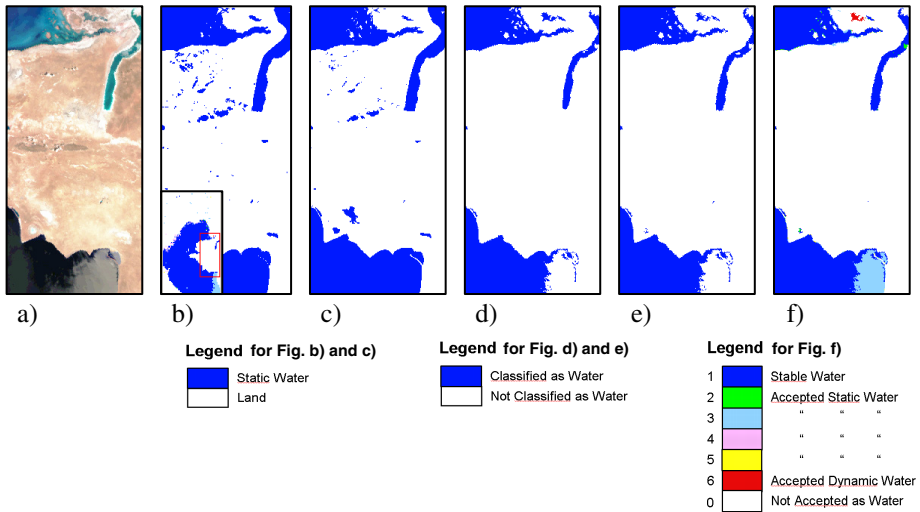


Fig. 6. a) Selected region of AATSR data (29th July 2002, RGB-image (p870, p670, p550), results of lwms based on SRTM data with ≥ 10 (b) and ≥ 60 percent water content (c), the masks based on the spectral algorithm (d) and the proposed algorithm in [1] (e), and mask lwmd3 of the method DySLEM (f)

A further interesting result is the red coded object in upper part of Fig. 6f. This object is only identified due to the dynamic masks in PWS3-3. Singular water pixels are also included in Fig. 7g. Only at the tongue of land in the south-west stable water was detected. In the RGB image this region would be visually interpreted as dry region. But the results in Fig. 6 are based on the static information as well as on the result of the adapted algorithm for lwmd2. The reason for such misinterpretation has to be examined.

A further interesting region is given with Scandinavia around 60° North. This is the latitude region of transition of SRTM to CIA WDB II data base, closely linked to a loss of detail in water information. To overcome this problem water objects detectable by both dynamic classification algorithms of WS2 will be generally accepted as dynamic water bodies or pixels in the final land-water mask (see red coded pixels in Fig. 7g).

In Scandinavia for water bodies a local shift between remote sensing data and static mask can be found. Fig. 7b shows a shift into south-west direction for Lake Pyhäjärvi in Finland. It is an image detail of an AATSR data set which demonstrates this problem. Some water pixels of the static mask (Fig. 7d) on the south-west lakefront cannot be identified by means of the dynamic masks and will not be included in the resulting mask. Some other pixels of the lake (north-eastern lakefront), outside the static land-water information, will be identified as accepted dynamic water pixels and can be seen in the resulting mask (Fig. 7g). The basic requirement for identification of such water pixels is given with pre-classification in WS2 to define the candidate status.

It seems that in some cases the offset in Equation 11 is a little bit too large. But therefore water can be identified below clouds or in some other cases salt lakes can be identified as water. In the context of the project "ESA-CCI Burned Area" funded by the ESA (European Space Agency) different masks for clouds, snow and ice and salt lakes are generated [4] in order to support the identification of critical water regions.

Additionally this is of advantage for a continuous transition of one mask to the next.

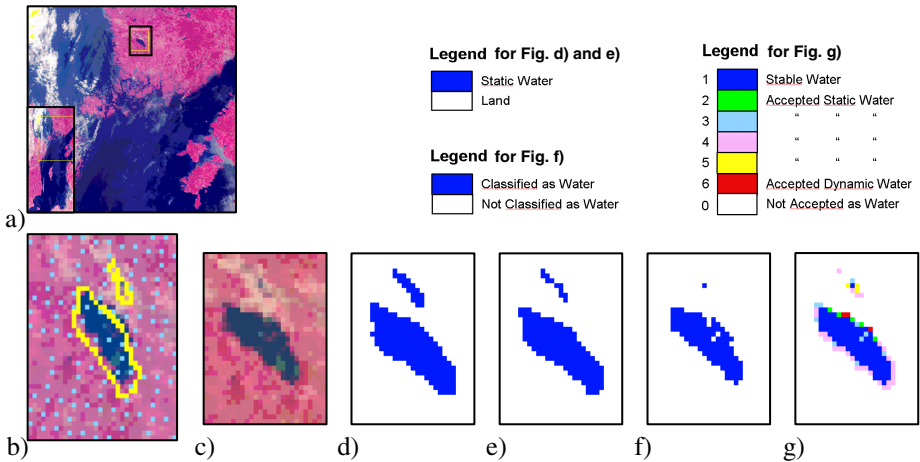


Fig. 7. Example of shift between static mask and image data. a) section of a data set of Baltic region with Lake Pyhäjärvi, b) data and coastline (yellow) within a map (without use of Nearest Neighbor), c) corresponding section of the data set, d) water mask with $\geq 10\%$, e) $\geq 60\%$ water within pixel area, f) classification lwm1, in this section equal to lwm2, g) result mask lwm3 (see legend).

5 Conclusions

Calculated frames from 14 AATSR passes in different regions show good classification results of water by fusion of static and dynamic masks based on a self-learning process.

It could be demonstrated that the proposed algorithm operates stable and produces good classification results. Interpretation mistakes can be minimized by using static land-water information and dynamic classification algorithms to derive independent land-water masks. These masks can be used to react to actual image content, so that a temporal snapshot of land-water information can be controlled in its actuality.

Experiences with AATSR data can be transferred to MERIS and VEGETATION data. Further progress can be expected by introducing additional classification rules for open water and water of lakes and rivers. Thus the inclusion of this additional information would be possible when splitting equation 7 into two different relations with the corresponding mean surface temperatures used in equations 8-11.

Acknowledgements. This study was made possible by ESA CCI ECV Fire Disturbance project (fire_cci), N°4000101779/10/I-LG. Furthermore we thank K. Guenther for the constructive discussions and his hints.

References

1. Birks, A.R.: Improvements to the AATSR IPF relating to Land Surface Temperature Retrieval and Cloud Clearing over Land, AATSR Technical Note, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, U.K (2007)
2. Borg, E., Fichtelmann, B.: Determination of the usability of remote sensing data- EP 1591961 B1 (2005)
3. Carroll, M.L., Townshend, J.R., DiMiceli, C.M., Noojipady, P., Sohlberg, R.A.: A new global raster water mask at 250 m resolution. *Int. J. of Digital Earth* 2, 291–308 (2009)
4. ESA CCI ECV Fire Disturbance (fire_cci), N°4000101779/10/I-LG
5. Fichtelmann, B., Borg, E., Kriegel, M.: Verfahren zur operationellen Bereitstellung von Zusatzdaten für die automatische Fernerkundungsdatenverarbeitung. In: 23rd AGIT Symposium Angewandte Geoinformatik 2011, Strobl, Blaschke, Griesebner, Salzburg, pp. 12–20 (2011)
6. Günther, K.P.: ESA-CCI-Burnt Area Pre-Processing, Kick-off Meeting. ESA-CCI Burnt Area Pre-Processing (2010a)
7. Günther, K.P.: Private communication (2010b)
8. Haas, E.M., Bartholomé, E., Combal, B.: Time series analysis of optical remote sensing data for the mapping of temporary surface water bodies in sub-Saharan western Africa. *J. Hydrology* 370, 52–63 (2009)
9. Justice, C., Giglio, L., Korontzi, S., Owens, J., Morisette, J., Roy, D., Descloitres, J., Alleaume, S., Petitcolin, F., Kaufman, Y.: The MODIS fire products. *Remote Sensing of Environment* 83(1&2), 244–262 (2002)
10. Landsat 7 User Handbook, http://landsathandbook.gsfc.nasa.gov/data_properties/
11. Lehner, B., Doll, P.: Development and validation of a global database of lakes, reservoirs, and wetlands. *Journal of Hydrology* 296, 1–22 (2004)
12. Pape, D.: CIA World DataBank II, <http://www.ev1.uic.edu/pape/data/WDB/> (last modified, 2004)
13. USGS: Documentation for the Shuttle Radar Topography Mission (SRTM) Water Body Data Files, http://dds.cr.usgs.gov/srtm/version2_1/SWBD/SWBD_Documentation/Readme_SRTM_Water_Body_Data.pdf
14. USGS (U.S. Geological Survey): Map Accuracy Standards, USGS Fact Sheet, pp. 171–199 (1999), <http://erg.usgs.gov/isb/pubs/factsheets/fs17199.pdf>
15. Wan, Z., Zhang, Y., Zhang, Q., Li, Z.: Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. *Remote Sensing of Environment* 83(1&2), 163–180 (2002)
16. Wessel, P., Smith, W.H.F.: A global, self-consistent, hierarchical, high-resolution shoreline database. *J. Geophys. Res.* 101(B4), 8741–8743 (1996)

DEM Accuracy of High Resolution Satellite Images

Mustafa Yanalak, Nebiye Musaoglu, Cengizhan Ipbuker, Elif Sertel,
and Sinasi Kaya

ITU Dept. of Geomatic Eng. 34469 Maslak, Istanbul-Turkey
buker@itu.edu.tr

Abstract. The aim of this research is to study the accuracy of Digital Elevation Models (DEMs) generated from two different satellite data namely OrbView-3 and IKONOS stereo images. 21 GCPs (Ground Control Points), 182 CPs (Check Points) and selected transects representing different land covers and topography were used for the accuracy analysis. Two DEMs were generated from OrbView-3 and one DEM from IKONOS stereo data. The accuracy of the used model for DEM generation was quantified using the RMSE (Root Mean Square Error) values of GCPs derived from Global Positioning System survey and error analysis over research area was performed using 182 CPs derived from traditional field survey. Several transects were formed over different parts of the study area and height values along these transects obtained from different DEMs were compared to determine and examine the accuracy. The results are analyzed with the empirical accuracy criterion for heights on analog maps.

Keywords: DEM, Accuracy, IKONOS, OrbView-3.

1 Introduction

Three dimensional evaluation of the earth surface is important since topographical data and its derivatives like slope and aspect can be used for variety of engineering applications in geology, geophysics, hydrologic modeling, and ecology and geospatial analysis [1] [2]. Traditionally, Digital Elevation Models (DEM) have been created using field survey whereas this approach has some limitations like being time consuming and not being appropriate for inaccessible and large areas. usage of stereo satellite images for DEM generation is another common approach by providing fast, reliable and accurate solutions.

First DEM generation from stereo images had started with the launch of Systeme P'our l'Observation de la Terre (SPOT) satellite series beginning in 1986 with SPOT-1. Afterwards, different satellite sensor systems were launched with stereo imaging capability. Several scientists have used high and medium resolution radar and stereo optical images to create DEM with different scale and accuracy [3] , [4] , [5]. Welch et al. [1] used ASTER stereo images and found that DEMs could be created with 15 to 25 m accuracy with a suitable distribution of GCPs. Erten et al. [6] investigated the accuracy of DEM generated from ASTER data by comparing ASTER derived DEM

with a DEM produced from 1/25000 scale topographic map. They found the elevation accuracy of ± 15 m to ± 25 m.

Hirano et al. [7] evaluated the vertical accuracy of ASTER stereo images resulting from stereocorrelation and indicated $RMSE_z$ values of ± 7 ve ± 15 m for their DEM. Cuartero et al. (2005) generated 146 DEMs, 91 from SPOT High Resolution Visible and 55 from ASTER stereo images and performed error control with 315 CPs determined by differential global positioning systems. Their results illustrated that RMSE values for elevations of Terra ASTER DEMs was 13 m, whereas RMSE value for SPOT HRV DEM was 7.3 m.

In addition to optical images, radar data can also be used to create DEMs. Toutin [8] generated DEMs using various RADARSAT stereopairs for three different topographic study sites. He analyzed the accuracy of stereo-derived DEMs as a function of slopes and aspects by comparing terrain slope and aspect computed from reference DEMs.

High resolution DEMs have been generated with the launch of very high resolution satellites. Wang et al. presented four different models to refine the rational function derived ground coordinates. They examined different configurations of GCPs to evaluate the impact on accuracy improvement. They found that GCP errors can be reduced from 5–6 to 1.5 m in horizontal and from 7 to 2 m in vertical directions, if an appropriate model and GCPs were used. Toutin [9] extracted DEMs from SPOT-5, EROS-A, IKONOS-II, and QuickBird using a three-dimensional multisensor physical model. He found elevation error values of 6.5, 20, 6.4, and 6.7 m for SPOT, EROS, IKONOS, and QuickBird, respectively with 68 % confidence level. Toutin [10] extracted Digital Terrain Models (DTMs) from SPOT-5 High Resolution Stereoscopic (HRS, 10m resolution) in-track stereo-images and High Resolution Geometric (HRG, 5m resolution) across-track stereo-images. He found errors of 6.4m, 6.8m and 5.1m in X, Y and Z axes, and 2.6m, 2.2m and 2.9m in X, Y and Z axes for HRS and HRG, respectively.

Although there have been several researches on investigation of DEM accuracy of very high resolution satellite images, there is only limited number of research about the generation and accuracy analysis of OrbView-3 data derived DEM. This research aims to analyze accuracy of 3 different DEMs generated from OrbView-3 and IKONOS images which two are generated from OrbView-3 and one generated from IKONOS. RMSE of GCPs and CPs were used to check the accuracy of DEMs. Transects were formed and height values obtained from DEMs were compared to determine and examine the accuracy.

2 Study Area and Data Used

2.1 Study Area

The study area used in this research is located in Istanbul Metropolitan area. Istanbul is in the northwest of Turkey and lies on the Bosphorus connecting Europe to Asia (Fig.1). It is among the most crowded cities of the World. Land surface characteristics of the cities have been changing significantly due to rapid economic development, industrialization and urbanization. A test area was selected within the study area to

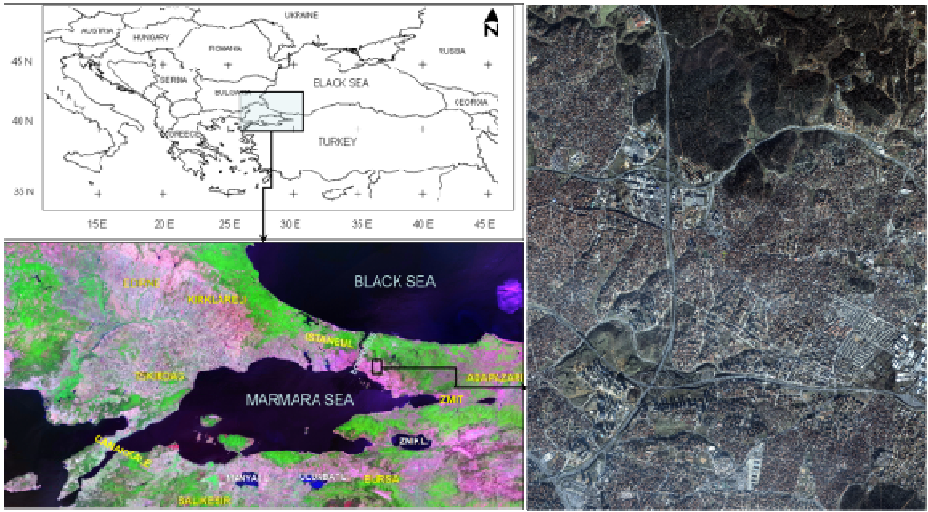


Fig. 1. Location of the study and test areas [11], [12]

conduct detailed DEM accuracy assessment. This area lies at the center of the study area including different land cover/use types like industrial buildings, residences, roads, sport complex and vegetation which have different geometry and heights. Ground height values within the study area ranges from 136 m to 206 m. Both regular and irregular buildings exist in urban areas like one or two floor irregular buildings with different geometry and six floor regular building groups with regular geometry [11], [12].

2.2 Satellite Data

OrbView-3 satellite was among the world's first commercial satellites providing high-resolution imagery from space with one-meter and four meter spatial resolutions for panchromatic and multispectral bands, respectively.. The IKONOS satellite (launched in September 1999) is the world's first commercial satellite with 1-m panchromatic, 4-m multispectral images in the very near infrared region. Two OrbView-3 stereo images received in 30 July 2006 and 28 October 2006 were used in the study. In addition to OrbView-3 stereo images, an IKONOS Geo image received in 30 December 2008 was also used for DEM generation [11].

2.3 Field Survey

A GPS network with 40 GCPs was formed including 8 permanent GPS stations of the Istanbul GPS Network as shown in Fig.2. The 3D coordinates of the GCPs were calculated from ties to the permanent GPS stations. 182 CPs are selected from building corners, road intersections etc. which are clearly identifiable on all images and used to generate DEMs and test the accuracy. GPS measurements were conducted on GCPs for 30 minutes with 5-second period.

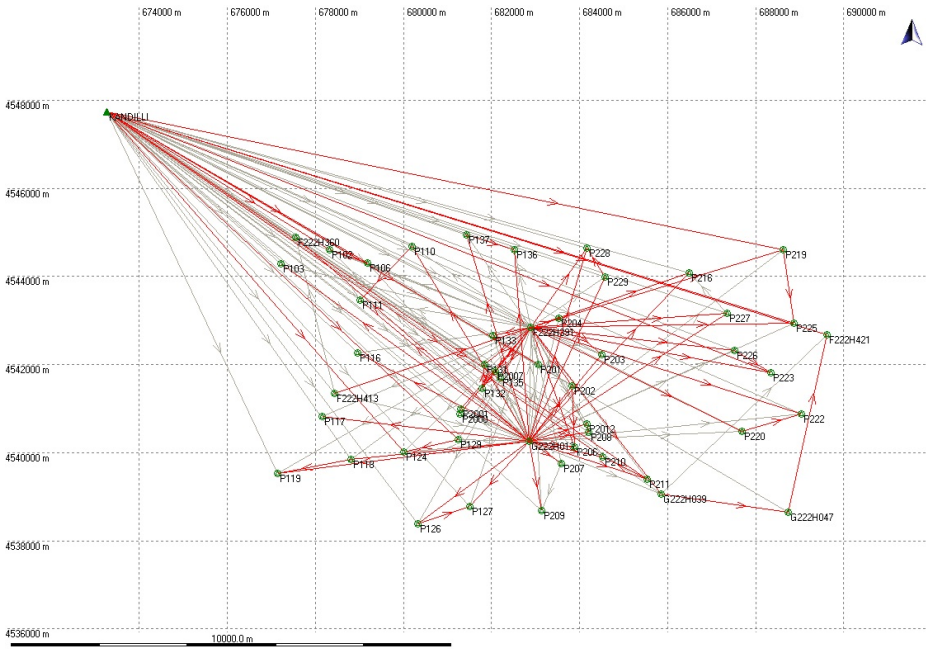


Fig. 2. GPS surveying canvas and measured bases

3 Methodology

Three DEMs were constructed using OrbView-3 and IKONOS data explained in the previous section. DEM1 was created from IKONOS stereo pairs whereas DEM2 and DEM3 were created using OrbView-3 stereo images. From now and on, these abbreviations will be used for DEMs created within this research. Static GPS-derived coordinates of GCPs and Tie Points (TPs) were used to generate DEM of stereo images. Numbers of GCPs and TPs used in the procedure are indicated in Table 1. Rational Polynomial Coefficients (RPC) were used to establish the relationship between image space and object space. Y-parallax value which is defined as the difference in perpendicular distances between two images of a point from the vertical plane containing air base was used to evaluate locations of TPs [13]. Maximum y parallax value was obtained as 1.8 m and this value assumed as reasonable since y parallax of 1 or 2 pixels can be acceptable.

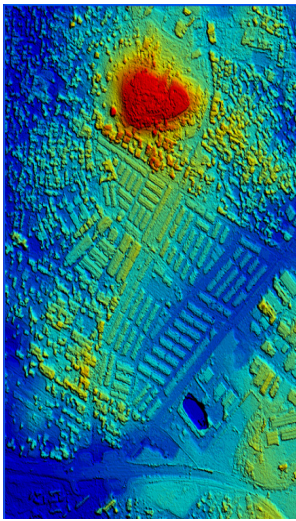
After the selection of GCPs and TPs, epipolar images defining the relationship between the pixels in the stereo pair were formed. Epipolar geometry represents that a ground point and the two optical centers lie on the same plane. Image matching finds the corresponding points on both the left and right images which represent the same ground feature. The quality of image matching impacts the quality of the output DEM.

Positions of TPs can be evaluated using y-parallax values. After evaluating TPs, left and right epipolar images were created. Epipolar images are oriented in such a way that ground feature points have the same y-coordinates on both images. Usage of epipolar

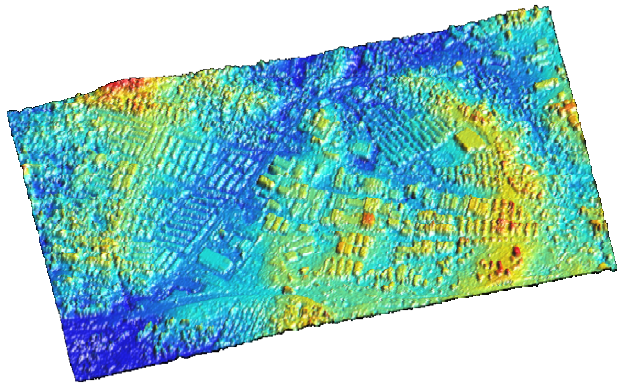
images increases the speed and reliability of image-matching procedure since it removes the one dimension of variability. At the next stage, a DEM is generated with 1x1 m pixel dimension considering the minimum correlation value of 0.70. Images were rectified to the UTM projection with WGS 84 datum and zone number 35.

Table 1. The number of GCPs and TPs

	DEM1 (Ikonos)	DEM2 (Orbview-3)	DEM3 (Orbview-3)
Date	30.12.08	30.07.06	28.09.06
Stereo pairs		(218-219)	(830-970)
Number of GCPs	20	10	10
Number of Tie Points	21	12	10
Mavimum parallax	0.98	0.80	1.61



a)



b)

Fig. 3. Digital Elevation Models (DEMs) derived from a) IKONOS and b) ORBVVIEW3 data

Homogeneously distributed and clearly identified points were selected during the generation of DEMs. Due to the cloud effect in OrbView-3 images, GCPs and TPs could not be selected from some parts of the image. Therefore, 10 GCPs could be identified over cloud-free areas while analyzing Orbview-3 images. The vertical accuracy of a DEM is calculated using the average vertical RMSEs of each grid. This accuracy depends on the number, accuracy and distribution of GCPs, used matching method and the relief of the area [14], [15].

The accuracy analysis of DEM1, DEM2 and DEM3 was conducted in two phases. In the first phase, height values of GCPs and CPs obtained from DEMs were compared with the height values of these points derived from GPS survey and traditional survey, respectively. For GCPs and CPs, height values derived from static GPS and traditional survey were assumed as actual values since they were highly accurate. Absolute error (ϵ_h) and Root Mean Square Error (RMSE) (m_h) of height values were calculated using Eq.(1) [16]. In Eq.1, $h(GPS)$ is height values obtained from GPS measurements whereas $h(DEM)$ is height values obtained from the related DEM and n is the number of measurements. These values were calculated for GCPs and CPs separately. The results were presented on Table 2 and 3.

$$\epsilon_h = h(GPS) - h(DEM) \quad m_h = \sqrt{\frac{\epsilon_h \cdot \epsilon_h}{n}} \tag{1}$$

After evaluating absolute errors, ϵ_h values, the differences which were greater than $3m_h$ were assumed as gross error and were not taken into consideration. Eq. 2 is used to calculate the differences of double measurements obtained from two different DEMs extracted from multi temporal OrbView-3 stereo images (DEM2 and DEM3). Double measurements represent the elevation values of the same point obtained from two different DEMs, DEM2 and DEM3 for OrbView-3 in this case.

$$d = h(DEM2) - h(DEM3) \quad M_h = \sqrt{\frac{d \cdot d}{2n}} \tag{2}$$

If the heights are being read from DEM2 and DEM3, then the RMSE of the difference of those two heights can be calculated using the equation below based on error propagation law,

$$d = h1 - h2 \quad m_d = \sqrt{m_{h1}^2 + m_{h2}^2} = \sqrt{0.9^2 + 1.1^2} = 1.4m \tag{3}$$

The M_h values were calculated for both GCP's and CP's and presented at the last rows of Table 2 and 3. M_h value can not be calculated for IKONOS, since there was only one stereo pair.

In the last phase of the study, transects were selected from three DEMs which represent different land cover types, geometric shape and heights and height values along these transects were evaluated to conduct accuracy assessment.

4 Results

The investigation of the accuracy obtained from the DEMs created from IKONOS and OrbView-3 stereo images has been realized in two stages: *i*) the RMSE values were calculated for GCPs and CPs, *ii*) then transects were analyzed.

RMSE values for heights were calculated for DEM1, DEM2 and DEM3. An additional RMSE value was also calculated using the differences of double measurements belonging to multitemporal OrbView-3 DEMs.

Table 2. DEM accuracy values calculated from CPs

	DEM1 (Ikonos)	DEM2 (OrbView-3)	DEM3 (OrbView-3)
Date		30.07.06	28.09.06
Stereo pairs		(218-219)	(830-970)
Max. Error (m)	2,6	2,0	2,7
m_h	0,9m	0,9m	1,1m
M_h (obtained from the differences of double measures)		1,0m	

According to statistical t table the difference of heights which has a RMSE value of 1.4m, has to be greater than $2m_d$ which has %95 statistical confidence or $2,6m_d$ which has %99 statistical confidence for being significant. These boundary values were calculated as 2.8m and 3.64 m, respectively.

Table 3. DEM accuracy values calculated from GCPs

	DEM1 (Ikonos)	DEM2 (OrbView-3)	DEM3 (OrbView-3)
Date		30.07.06	28.09.06
Stereo pairs		(218-219)	(830-970)
Max. Error (m)	2,4	1,8	2,7
m_h	0,8m	0,8m	1,1m
M_h (obtained from the differences of double measurements)		1,0m	

An approximate RMSE value for the double measurement of IKONOS image was calculated by taking the same RMSE (0.9 m) for calculation. Since there were not multitemporal IKONOS images, RMSE obtained from a second DEM was assumed as the same with DEM1. Equation 4 illustrates the calculation of RMSE for double measurements:

$$m_d = \sqrt{0.9^2 + 0.9^2} = 1.27m \quad (2m_d = 2.55m \text{ and } 2,6m_d = 3.31m) \quad (4)$$

The result for the height differences obtained from another combination of images, i.e. one IKONOS and one OrbView-3 image is as follows

$$m_d = \sqrt{0.9^2 + 1.0^2} = 1.35m \quad (2m_d = 2.70m \text{ and } 2,6m_d = 3.51m) \quad (5)$$

Here, the average RMSE of 1.0m is taken for OrbView-3.

The confidence value of DEM can be calculated based on the number of GCPs and the expression relating reliability to number of CPs is given in the Eq. 5 [4] , [17]

$$R(e) = \frac{1}{\sqrt{2(n-1)}} \times 100\% \tag{6}$$

where R(e) is the confidence value in percent and n is the number of CPs which is 182 for this research. R(e) value is calculated as 5.25 % for the study, 94.75 % statistical confidence.

Transects were selected representing the regions with different land use/cover types and different height values like transects lying over building series and roads. Elevation values along the transects were plotted as spatial profiles using OrbView-3 and IKONOS DEMs. Figure-4 shows transects derived from DEM2 and DEM3. Since DEM2 and DEM3 were derived from multitemporal OrbView-3 stereo images, any significant change within the transects created over these DEMs will point out abrupt changes in vertical dimension on the related region. The regions selected for Transect 1, 2 and 3 represent the unchanged areas. There are small discrepancies between spatial profile of height values obtained from DEM2 and DEM3. These discrepancies caused by the errors of DEMs as illustrated with RMSE. Differences of height values for DEM2 and DEM3 are around 2.5 and 3 m . These differences are within limit of 95% and 99% confidence levels which are represented with $2m_d = 2.8m$ and $= 3.64m$ given above.

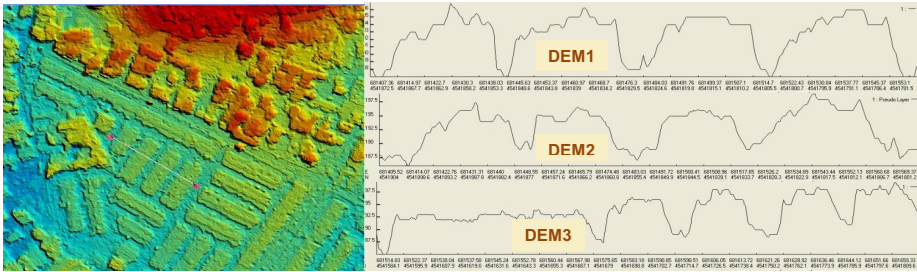


Fig. 4. Transect 1 and profiles from DEM1, DEM2,DEM3

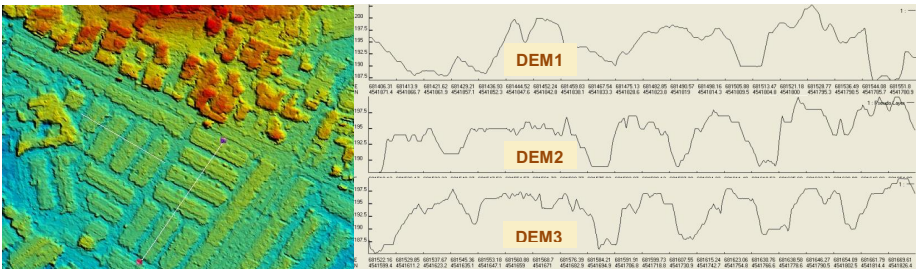


Fig. 5. Transect 2 and profiles from DEM1,DEM2,DEM3

Figure 4, and 5 show the transects derived from DEM1 (IKONOS), DEM2 and DEM3 (OrbView-3). The sections given by Figure 3 cross over 4 buildings with the same height. The buildings and the distances between them can be easily identified on DEM1 according to DEM2. DEM1 shows the elevations approximately the same but they are distorted on DEM2.

The sections given by Figure 5 defines a direction which lies over five buildings with one high and four low buildings. The buildings and the distances between them can be easily distinguished on DEM1 and DEM2. For example; the 193 m elevation of the first building can be read as 192.5 m on the first section and as 195 m on the second [11].

The RMSE value (m_h) ± 1.0 m for heights given on the Tables 2 and 3 cover the map scale 1:5000 for rough topography [11].

5 Conclusions

This research aims to analyze two DEMs generated from OrbView-3 and one generated from IKONOS. Two different dated OrbView-3 stereo data were analyzed to determine the consistency between different DEMs created from same data source and to compare Orbview-3 derived DEMs with another DEM source namely IKONOS stereo data.

RMSE were calculated as 0.9 m for IKONOS and 1.1 m for OrbView-3. The RMSE of height values for IKONOS and OrView-3 derived DEMs are corresponded to accuracy of heights that could be achieved from 1:5000 scale maps. Multi-date DEMs derived from Orbview-3 and IKONOS stereo images were used to analyze the monitoring capacity of stereo images for 3D changes over time. The results showed that, it is not significant to monitor height changes smaller than 2.8 m considering 95 % statistical confidence and 3.64 m considering 99 % statistical confidence for OrbView-3 DEMs. Similarly, it is not significant to monitor height changes smaller than 2.5 m considering 95 % statistical confidence and 3.3 m considering 99 % statistical confidence for IKONOS DEM.

Vertical displacements and deformations greater than 2.5-2.8 m can be determined using OrbView-3 and/or Ikonos DEM's with %95 statistical confidence. It is possible to determine the amount of cut and fill areas with uncontrolled solid waste between these practical accuracy limits. Without a doubt, for an accurate accuracy assessment the results have to be evaluated considering the sensor properties, atmospheric conditions, topographic feature and characteristics, distribution of the ground control and test points. Obviously, OrbView-3 offers a single channel data. As a future work, the study may be enhanced with multi-channel data from multi-spectral sensors in order to search the thematic accuracy. Increasing the quality of satellite images, the researches and applications will be increase about the subject of DEM.

Acknowledgements. The authors would like to thank the Scientific and Technological Research Council of Turkey (TUBITAK) for supporting the study through the project 105Y124.

References

1. Welch, R., Jordan, T., Lang, H., Murakami, H.: ASTER as a Source for Topographic Data in the Late 1990's. *IEEE Transactions on Geoscience and Remote Sensing* 36(4) (July 1998)
2. Zyl, V., Jakob, J.: The shuttle radar topography mission (srtm): a breakthrough in remote sensing of topography. *Acta Astronautica* 48(5-12), 559–565 (2001)

3. Chen, Y., Shi, P., Li, J., Deng, L., Hu, D., Fan, Y.: DEM accuracy comparison between different models from different stereo pairs. *International Journal of Remote Sensing* 28(19), 4217–4224 (2007)
4. Cuartero, A., Felicísimo, A.M., Ariza, F.J.: Accuracy, Reliability, and Depuration of SPOT HRV and Terra ASTER Digital Elevation Models. *IEEE Transactions on Geoscience and Remote Sensing* 43(2) (2005)
5. Toutin, T., Cheng, P.: QuickBird - A milestone to high resolution mapping. *Earth Observation Magazine* 11(4), 14–18 (2002)
6. Erten, E., Musaoglu, N., Erbay, Y.: Quality assesment of Digital Elevation Model Produced from ASTER images. In: *Proceedings of 6th Geomatic Week, Barcelona (2005)*
7. Hiranoa, A., Welch, R., Lang, H.: Mapping from ASTER stereo image data: DEM validation and accuracy assessment. *ISPRS Journal of Photogrammetry & Remote Sensing* 57, 356–370 (2003)
8. Toutin, T.: Impact of terrain slope and aspect on radargrammetric DEM accuracy. *ISPRS Journal of Photogrammetry & Remote Sensing* 57, 228–240 (2002)
9. Toutin, T.: Review Article: Geometric Processing of Remote Sensing Images: Models, Algorithms, and Methods. *International Journal of Remote Sensing* 25(10), 1893–1924 (2004)
10. Toutin, T.: Generation of DSMs from SPOT-5 in-track HRS and across-track HRG stereo data using spatiotriangulation and autocalibration. *ISPRS Journal of Photogrammetry & Remote Sensing* 60, 170–181 (2006)
11. Yanalak, M., Musaoglu, N., Ormeci, C., Kaya, S., Alkan, R.M., Tari, E., Ipbuker, C., Turkoglu, H., Saroglu, E., Yavasoglu, H., Erden, T., Karaman, H., Bilgi, S., Cetin, M.: Investigation of the Accuracy and Enginnering Applications of Orbview-3 Images, TUBITAK Project, No: 105Y124 (2008) (in Turkish)
12. Yanalak, M., Sertel, E., Musaoglu, N., Ipbuker, C., Kaya, S.: Comparison of Planimetric and Thematic Accuracy of Orbview3 and Ikonos Images. *Journal of Indian Society of Remote Sensing* (2010)
13. CCRS, Canada Centre for Remote (2009), <http://www.ccrs.nrcan.gc.ca> (accessed March 2009)
14. Atak, O.: Geometric Accuracy and Feature Compilation Assessment of High Resolution Satellite Imagery, ITU Institute of Science and Technology, PhD Thesis, Istanbul (2007) (in Turkish)
15. Toutin, T., Chénier, R., Carbonneau, Y.: 3D models for high resolution images: examples with Quickbird, IKONOS and EROS. In: *Archives of ISPRS Symposium, Comm. IV, Ottawa, Ontario, Canada, vol. XLIII, pt. 4, pp. 547–551 (2002)*
16. USGS, National Mapping Program, Standards for Digital Elevation Models, Part 2: Specifications (1998)
17. USGS, Digital Elevation Models: Data Users. U.S. Geol. Surv., Reston (1987)

Low Cost Pre-operative Fire Monitoring from Fire Danger to Severity Estimation Based on Satellite MODIS, Landsat and ASTER Data: The Experience of FIRE-SAT Project in the Basilicata Region (Italy)

Antonio Lanorte, Fortunato De Santis, Angelo Aromando, and Rosa Lasaponara

CNR-IMAA (Institute of Methodologies For Environmental Analysis), Potenza, Italy
lasaponara@imaa.cnr.it

Abstract. This paper presents the results we obtained in the context of the FIRE-SAT project focused on the use of satellite data for pre-operational monitoring of fire danger and fire effects in the Basilicata Region. The use of satellite data was manifold, to obtain: (i) fuel property (type and loading) maps, mainly obtained from satellite Landsat TM data, (ii) fuel moisture estimation (mainly from MODIS), (iii) fire danger/susceptibility indices as well as (iv) post fire effects including fire severity and vegetation recovery assessment. Results obtained during the first year of project (2008) suggested that the integrated model identified the main fire danger zones by means of the integration of fuel types with daily fuel moisture and Greenness maps. MODIS multitemporal data analyses enable us to dynamically estimate fire severity as well as to map fire affected areas and evaluate the vegetation recovery capability over time. The pre-operative use of the integrated model, carried out within the framework of the FIRE-SAT project funded by the Basilicata Region, pointed out that the system enables us to timely monitor spatial and temporal variations of fire susceptibility and promptly provide useful information on both fire severity and post fire regeneration capability.

Keywords: satellite, fire monitoring, NDVI, NDWI, greenness, moisture, burned areas mapping, fire resilience.

1 Introduction

In the Mediterranean regions, fires are considered the most important cause of land degradation according to UNCCD (1994) [1]. Every year, on average 45,000 forest-fires break out in the Mediterranean basin causing the destruction of about 2.6 million hectares according to FAO reports, see for example [2]. Several studies dealing with the effects of fires on the vegetation cover carried out within the Mediterranean basin found that fires induce significant alterations in short as well as long-term vegetation dynamics.

In particular, during the last decades, the repeated occurrence of severe wildfires affecting various parts of the world. This heightened the significant damage caused by

fire that has a powerful influence on ecosystems dynamics and function and can lead to permanent changes in the composition of vegetation community.

Fires lead to permanent changes in the composition of vegetation community, cause decrease in forests, loss of biodiversity, soil degradation, alteration of landscape patterns and ecosystem functioning thus speeding desertification processes up (see, for example, [1]). Moreover, recently it has been found that fires facilitate alien plant invasion, patch homogenization, and create positive feedbacks in future fire susceptibility, fuel loading, fire spreading and intensity.

Among the Mediterranean ecosystems, forests are considered to be very vulnerable to fires, and, therefore, they can trigger disasters, such as soil erosion, flooding, etc. Variations in the composition and distribution of vegetation represent one of the main sources of systematic change on local, regional, or global scale since the characteristics of vegetation cover, including cover type and phenology, affect processes such as water cycle, absorption and re-emission of solar radiation, momentum transfer, carbon cycle, and latent and sensible heat fluxes. Therefore, the use of reliable monitoring system is of primary importance to limit fire damage. After fire, the ability to characterize the reaction of vegetation to disturbance phenomena over large landscapes and with high temporal frequency is of primary importance for an adequate management of damaged resources and for limiting future fire danger.

To mitigate fire-related problems, forest and land management agencies require an early warning system to assist them in implementing fire prevention and management plans. In this context, remote sensing technologies can provide useful data for fire management spanning from danger estimation as see, for example, Lasaponara [3], fuel mapping see for example the approach developed by Lasaponara and Lanorte [4, 5, 6], fire detection [7], fire mapping and severity estimation [8, 9] to post fire monitoring [10,11].

In this paper we focus on investigations performed in the Basilicata Region (Southern Italy) which in the recent years has been characterized by an increasing incidence of fire disturbance. Even if fires are generally really small (generally ranging from 1 to 10 hectares) they also tend to affect protected (Regional and national parks) and natural vegetated areas. FIRE_SAT project has been funded by the Civil Protection of the Basilicata Region in order to set up a low cost methodology for fire danger/risk monitoring based on satellite Earth Observation techniques. To this aim, NASA Moderate Resolution Imaging Spectroradiometer (MODIS) data were used. The spectral capability and daily availability makes MODIS products especially suitable for estimating the variations of fuel characteristics.

In this work, we focus on significant results we obtained in the context of FIRE-SAT project, spanning from the danger monitoring to fire severity and vegetation recovery estimation

2 Methodology

Remote sensing technologies can provide useful data for fire management from risk estimation [3], fuel mapping [4, 5, 6], fire detection [7], to post fire monitoring [8, 9, 10,11] for assessing both burn-severity and vegetation fire resilience.

The methods generally used to estimate fire danger, burned areas and severity from satellite are based on fixed threshold values, which are not suitable for fragmented landscapes and vegetation types, or geographic regions different from those they were devised. In this paper, we present the new approaches, based on satellite and geo-statistical analysis, we devised for multifold applications ranging from fire danger monitoring, burn severity mapping to vegetation recovery, as described in section 2.1, 2.2, 2.3.

2.1 Fire Danger Estimation: Rationale and Approaches

Satellite technologies represent a cost-effective mean for obtaining useful data that can be easily and systematically updated for the whole globe. Nowadays medium resolution satellite images, such as Landsat TM or ASTER can be downloaded free of charge from the NASA web site. The use of satellite imagery along with reliable spatial analysis techniques can be used for fire monitoring at a detailed level.

In order to obtain a dynamical indicator of fire susceptibility based on multitemporal MODIS satellite data, up-datable in short-time periods (daily), we used the spatial/temporal variations of the following parameters:

- Relative Greenness Index
- fuel moisture content

2.1.1 Relative Greenness Index

Relative Greenness Index map are obtained from the NDVI which is widely used for vegetation monitoring and also as a surrogate to estimate vegetation water content even with strong limitations mainly encountered when vegetation coverage is dense and the index is close to the saturation level.

$$NDVI = \frac{\rho_{NIR} - \rho_R}{\rho_{NIR} + \rho_R} \tag{1}$$

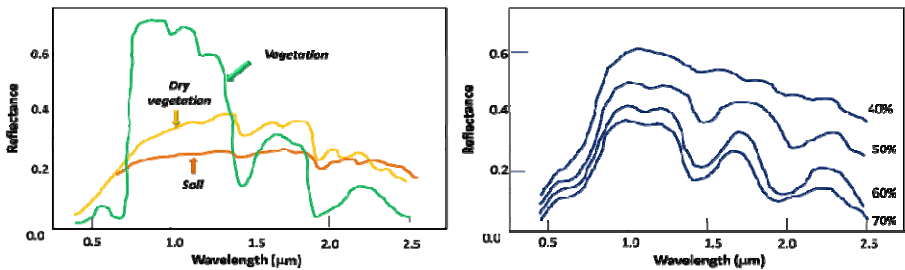


Fig. 1. Reflectance spectral behaviour for soil, vegetation(on the left); whereas on the right Reflectance spectral behavior according to soil moisture content variations.

The arithmetic combination of the red and Near infrared (NIR) enables us to exploit the different spectral behavior of vegetation cover in the different bands. NDVI provides a dimensionless numerical value. The formula is designed as a ratio, in order to normalize its variability field between -1 and +1. NDVI assumes values less than 0 for water, is slightly higher than 0 for soils and higher than 0.4 for vegetation, dense vegetation can exceed 0.8 or be close to saturation (1) for a rainforest.

MODIS data have been used to obtain both (i) Greenness and (ii) moisture index.

(i) Greenness is a quite popular fire danger index, developed by Burgan, et al. in 1998. The basis for calculating RG is historical NDVI data that defines the maximum and minimum NDVI values observed for each pixel. Thus RG indicates how green each pixel currently is in relation to the range of historical NDVI observations for it. RG values are scaled from 0 to 100, with low values indicating the vegetation is at or near its minimum greenness. Specifically the algorithm is:

$$RG = (ND0 - NDmn)/(NDmx - NDmn) * 100 \quad (2)$$

where

ND0 = highest observed NDVI value for the considered composite period which 8 days

NDmn = historical minimum NDVI value for a given pixel

NDmx = historical maximum NDVI value for a given pixel

The purpose of using relative greenness in the fire danger estimation is to partition the live fuel load between the live and dead vegetation fuel classes.

2.1.2 Moisture

Over the years, NDVI was recognized as a useful "surrogate" for the estimation of vegetation water content for grassland but, as a general rule, the relationship between this index and the vegetation moisture content is strongly linked with the amount of vegetation. NDVI provides information closer to the amount and greenness of vegetation rather than moisture content and it is generally limited by soil reflection. Moreover, its effectiveness is reduced due to its sensitivity to atmosphere. The advantage to using NDVI for water content estimation is that this index is simple and available routinely and globally, but a number of limitations can be summarized as follows:

- (i) NDVI saturates at intermediate values of leaf area index (LAI), therefore it is not responsive to the full range of the canopy.
- (ii) Each plant species has its own relationship of chlorophyll and moisture content;
- (iii) A decrease in chlorophyll content does not imply a decrease in moisture content;
- (iv) A decrease in moisture content does not imply a decrease in chlorophyll content.

The NDVI has been used as an indicator of vegetation moisture content mainly because it was the only information available before the launch of Terra satellite platform with onboard MODIS hyperspectral sensor.

More recently, the potentiality of using satellite SWIR spectral bands for moisture content estimation has been supported by both modeling and experimental studies based on the available multispectral satellite datasets.

Using multispectral satellite data, the estimation of moisture content into soil and vegetation may be improved using vegetation indices based upon NIR and SWIR and in general on the longer wavelength reflective infrared range (1240–3000 nm), for example, the short-wave infrared (SWIR) reflectance (1300–2500 nm).

Several spectral indices, such as Normalized Vegetation Moisture Index (NVMI) or Normalized Difference Water Index (NDWI), mainly based on SWIR bands, can be computed to estimate moisture content for both soil and vegetation. The mathematical formulation of these indices (see formula 3 and 4) is very similar to the NDVI, but based on specific bands of water absorption.

$$NVMI = \frac{\rho_{NIR} - \rho_{SWIR}}{\rho_{NIR} + \rho_{SWIR}} \tag{3}$$

$$-NDWI = \frac{\rho_{RED} - \rho_{SWIR}}{\rho_{RED} + \rho_{SWIR}} \tag{4}$$

Both of these two indices NVMI and NDWI are sensitive to water content in vegetation and soil, respectively; being that the absorption of water content of vegetation close to NIR band (and that of soil close to red band) is negligible, whereas a small absorption is present into the SWIR spectral range.

Moreover, in comparison with NDVI both NVMI and NDWI are less sensitive to the effects of the atmosphere, but, the effects of soil reflection are still present. To date, among the wide range of vegetation indices, specifically devised to estimate vegetation water content, we adopted the MSI. This choice was driven by the results from statistical analyses that we performed on a significant time series. Such results pointed out the all the available satellite-based moisture index exhibited high correlation values.

$$MSI = R_{1600} / R_{820} \tag{5}$$

where R_{1600} and R_{820} denote the MODIS Reflectance as acquired in the spectral bands 1600nm and 820 nm.

The danger classification on live fuel can be estimated by dividing the range of the MSI maps into different classes. Finally, The fire danger index (FDI), related to vegetation state, was obtained by combining the danger classes obtained from RG and those from MSI following an approach similar to that adopted in Lasaponara [3] for combining NDVI and Temperature. High fire danger, as classified by FDI, was deduced by a combination of high dryness and low RG values, and low fire danger was deduced by a combination of low dryness and high RG values.

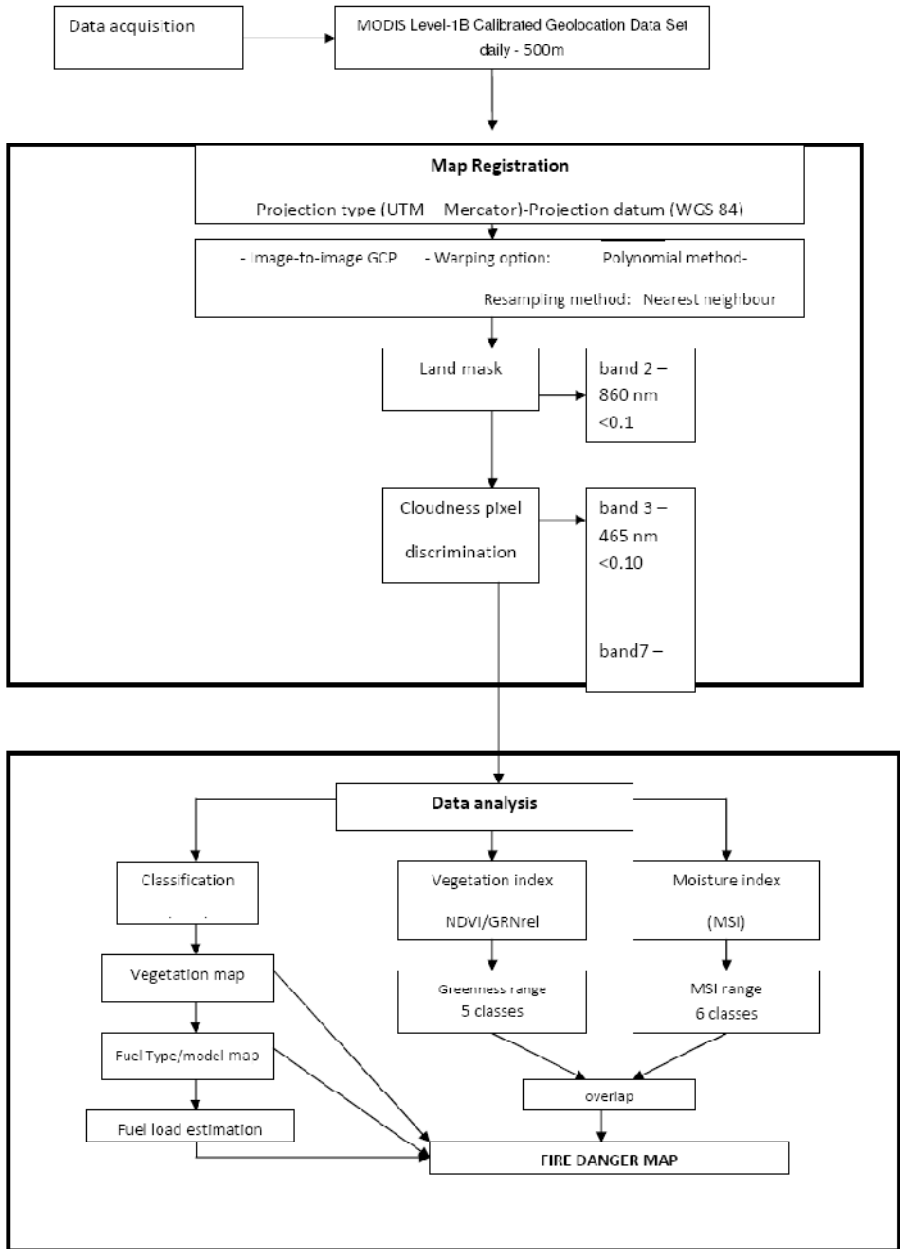


Fig. 2. Earth Observation Processing chain development

Figure 2 shows the flow chart of the data process used for creating a model for fire vegetation susceptibility assessment. Both MODIS and Landsat TM images were used. In particular, MODIS data were used to obtain variations in vegetation

Greenness and moisture content, while Landsat TM data to obtain fuel types and TM were processed using supervised classification techniques and spectral analysis methodologies performed at sub-pixel level to map: (i) Vegetation type (ii) Fuel type (Prometheus system), (iii) Fuel model (NFFL system), (iv) Fuel load (see, for example, Lasaponara & Lanorte and references therein quoted [4,5,6].

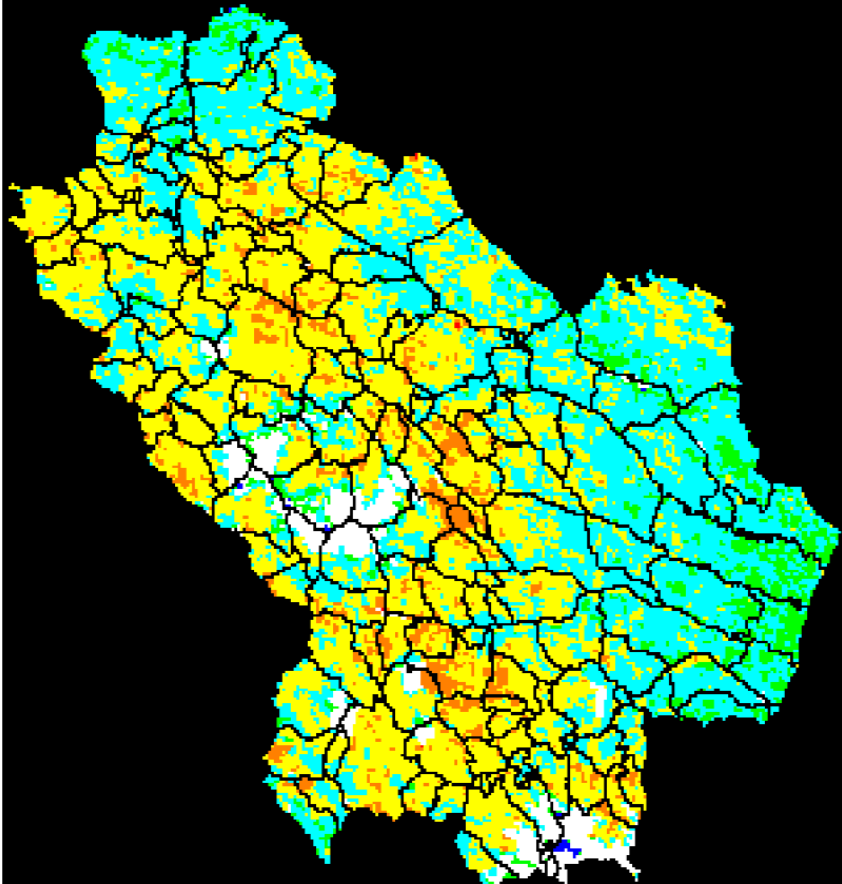


Fig. 3. Example of moisture map

In particular, the dead fuel moisture content is a key factor in fire ignition. Dead fuel moisture dynamics are significantly faster than those observed for live fuel. Dead fine vegetation exhibits moisture and density values dependent on rapid atmospheric changes and strictly linked to local meteorological conditions. For this reason, commonly, the estimation of dead fuel moisture content is based on meteorological variables. In this study we propose to use MODIS data to estimate meteorological data (specifically Relative Humidity) at an adequate spatial and temporal resolution.

Figures 3 and 4 show example of both moisture and Relative Greenness maps (different colours were used according to the moisture and the Greenness values). The assessment of dead fuel moisture content plays a decisive role in determining a fire dynamic danger index in combination with other factors.

This greatly improves the reliability of fire danger maps obtained on the basis of an integrated approach of the dynamic factors mentioned above and the static factors (fuel physical properties, morphological parameters and social-historical factors).

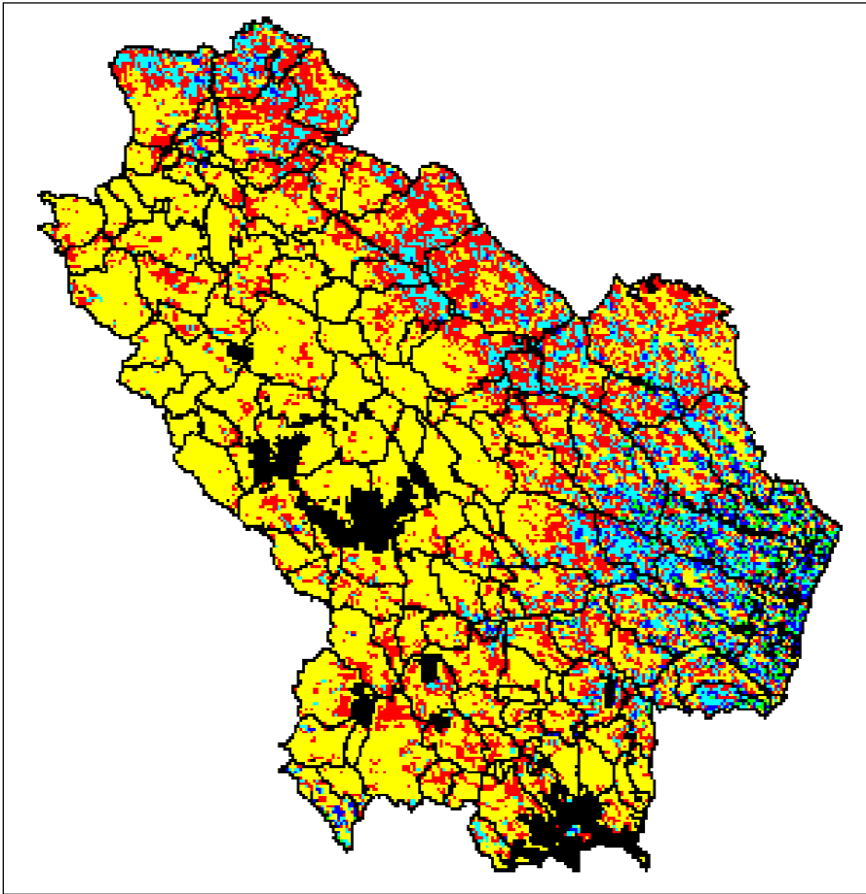


Fig. 4. Example of Relative Greenness map red to blue indicate

2.2 Fire Severity Estimation

Traditional methods of recording fire burned areas and fire severity involve expensive and time-consuming field survey. The available remote sensing technologies may allow us to develop standardized burn area and fire severity maps for evaluating fire

effects and addressing post fire management activities. This section is focused on results we obtained from ongoing research carried out to evaluate spatial variability of fire effects on vegetation. For the purposes of this study satellite ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) data have been used. Both single (post-fire) and multi-date (pre and post fire) ASTER images were processed for some test areas in Southern Italy. Spatial autocorrelation statistics, such as Moran’s I, Geary’s C, and Getis-Ord Local Gi index (see [12,13,14], were used to measure and analyze the degree of dependency among spectral features of burned areas.

Several vegetation indices were used with ASTER VNIR data in order to map the burnt areas. The SVI (Simple Vegetation Index), TVI (Transformed Difference Vegetation Index), SAVI (Soil Adjusted Vegetation Index) and NDVI (Normalized Difference Vegetation Index) have been computed as combination of spectral bands as in formula 1 to 5. Moreover, Normalized Difference Burn Ratio (NDBR) has been considered and computed from NIR and SWIR bands. These indices have been compared for assessing their ability in providing burnt area mapping.

Vegetation indices based on ASTER-VNIR are the result of the following formulas

$$SVI = \frac{NIR}{R} \tag{6}$$

$$NDVI = \frac{NIR - R}{NIR + R} \tag{7}$$

$$SAVI = \frac{NIR - R}{NIR + R + 0,5} \times 1,5 \tag{8}$$

$$TVI = \sqrt{NDVI + 0.5} = \sqrt{\frac{NIR - R}{NIR + R} + 0.5} \tag{9}$$

$$NDBR = (NIR - SWIR) / (NIR + SWIR) \tag{10}$$

where R, NIR, SWIR are the Red, Near-Infrared, and SWIR spectral bands, respectively.

All the vegetation indices were processed using spatial autocorrelation statistics to assess which of them performs better for fire burned area and fire severity mapping. Spatial autocorrelation statistics measure the degree of spatial dependency among the observations, the similarity of objects within an area, the level of interdependence between the variables, the nature and strength of the interdependence.

2.3 Estimation of Vegetation Recovery Capability

The method used in this work is the Detrended Fluctuations Analysis (DFA), which is suited for the study of long-range correlations. Traditional approaches, such as power

spectrum and Hurst analysis to quantify the correlations, are applicable only to stationary signals. A time series is stationary if its mean, standard deviation higher moments, and the correlation functions, are invariant under time translation. Signals that do not obey these conditions are nonstationary.

The DFA method has emerged as an important tool for the detection of long-range correlations in non-stationary time series and it works well for certain types of nonstationarym especially slowly varying trends, as in the case of vegetation behaviour.

DFA provides a quantitative parameter, the scaling exponent, which describe the properties of autocorrelation in long-range signals [10,11].

The main advantages of DFA compared to traditional methods are the following:

(i) it is capable to capture correlations in seemingly non-stationary time series and also prevent false detections (i.e. 'artifact of non-stationarity").

(ii) it can systematically eliminate both trends of various species due to different external effects

(ii) it reduces noise caused by imperfect measures.

The DFA method investigates the temporal evolution of the variance of integrated time series [10] by analyzing the scaling of a fluctuation function. It consists of the following steps:

1) The considered interval time series (of total length N) is integrated using formula 1

$$y(k) = \sum_{i=1}^k x(i) - \langle x \rangle, \tag{1}$$

where $\langle x \rangle$ is the mean value of x

.2) The integrated signal $y(k)$ is divided into boxes of equal length n.

3) For each n-size box, we fit $y(k)$, using a linear function, which represents the trend in that box. The y coordinate of the fitting line in each box is indicated by $y_n(k)$.

4) The integrated signal $y(k)$ is detrended by subtracting the local trend $y_n(k)$ in each box of length n.

5) For given n-size box, the root-mean-square fluctuation, $F(n)$, for this integrated and detrended signal is given by

$$F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^N [y(k) - y_n(k)]^2} \tag{2}$$

6) The above procedure is repeated for all the available scales (n-size box) to furnish a relationship between $F(n)$ and the box size n, which for long-range power-law correlated signals is a power-law

$$F(n) \sim n^\alpha. \tag{3}$$

The scaling exponent α quantifies the strength of the long-range power-law correlations of the signal: if $\alpha=0.5$, the signal is uncorrelated; if $\alpha>0.5$ the correlations of the signal are persistent, where persistence means that a large (small) value (compared to the average) is more likely to be followed by a large (small) value; if $\alpha<0.5$ the correlations of the signal are anti-persistent, which indicates that a large (small) value (compared to the average) is more likely to be followed by a small (large) value.

3 Results

The analysis was performed in the Basilicata (9,992 km²) Region for the 2008 year. Figures 5 show some fire danger maps obtained for the summer season. Currently, the fire susceptibility maps are provided daily during the fire season (summer season) and weekly for the rest of the year.

Figures 5 show some example of fire danger maps obtained for the 2008 summer season (on the left for 15 June 2008 and on the right for the 15 July 2008) September). In figure zoom of the same image as in Figure 5, finally Figure 7 show an example of the fire alert map daily provided to the Protezione Civile of the Basilicata Region.

Results obtained during the first year of the project (2008) shows that more than 85% of fires occurred in the areas classified as high and very high danger. The 15% percentage of fires which occurred in areas classified as moderate or low danger generally took place in forest areas and this was mainly due to the fact that the understory and dead fuel are masked by the canopy.

The satisfactory results obtained for the study area suggests that the MODIS-based model identified the main fire danger zone. In particular, the integration of the fuel type/model map, with fuel moisture daily and Greenness maps into a single, integrated indices allow us to properly capture the spatial and temporal variation of fire susceptibility.

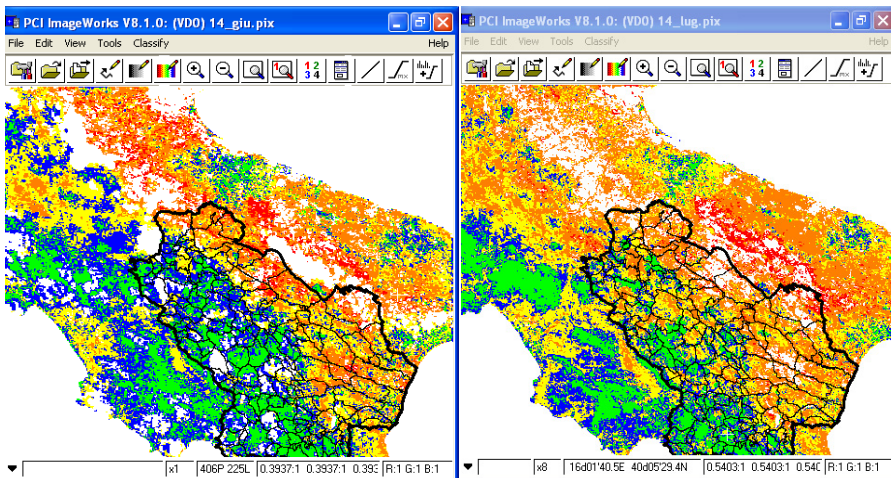


Fig. 5. Example of fire danger maps from blue to red pixels with increasing fire danger

The validation of the fire danger indices was carried out using statistics of occurred forest fires. The validation results show satisfactory agreement with the fire danger map taking into account that fire events are indirect indicator of fire danger; indeed, many factor influence fire ignition and spread such as human pressure, fire-fighting conditions, wind, etc...

For the performance evaluation step carried out in the framework of this project, we have defined and used several ad hoc fire statistic indices which were useful for the validation of the fire danger maps and for the creation of the basic elements for devise and define a validation protocol procedure.

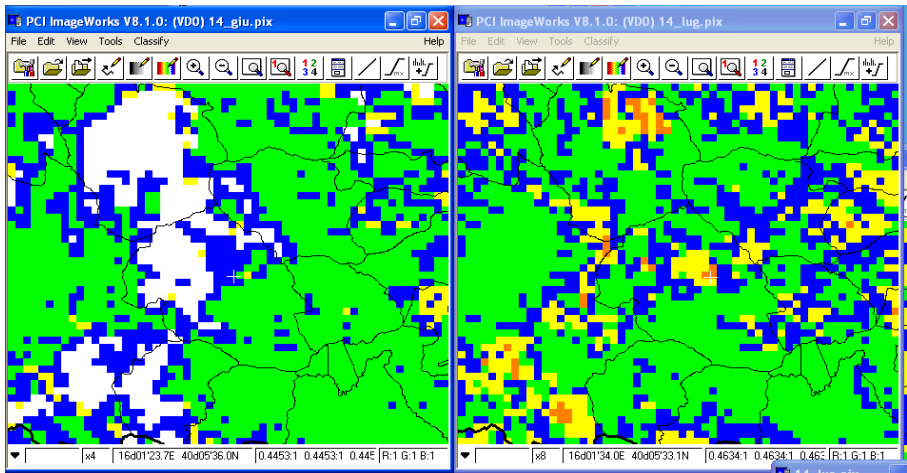


Fig. 6. Zoom extracted from the Example of fire danger maps

Relating the estimation of fire severity maps, results obtained from spatial statistics enable us to map the areas affected by fire and to estimate the degree of fire severity. Results from several analyses confirmed that the best indices are the NDBR single and multivariate. The use of multitemporal information is recommended, but, of course only for cloud free pixels. Figure 8

Our results pointed out that spatial autocorrelation statistics applied to ASTER data allow us to discriminate fire severity and to improve the monitoring of fire effects over time.

Such information are effective data source for performing a timely evaluation of erosion/runoff, biomass and carbon issues, and other issues which depend on fire severity characteristics.

As in the case of fire danger estimation also for fire severity the adopted approach is independent of (i) fixed threshold values, (ii) sensors used for the evaluation and also vegetation cover types affected by fire. Therefore, the model could be incorporated directly into the mapping process from local up to global scale, using TM and ASTER images or MODIS according to the detail requested.

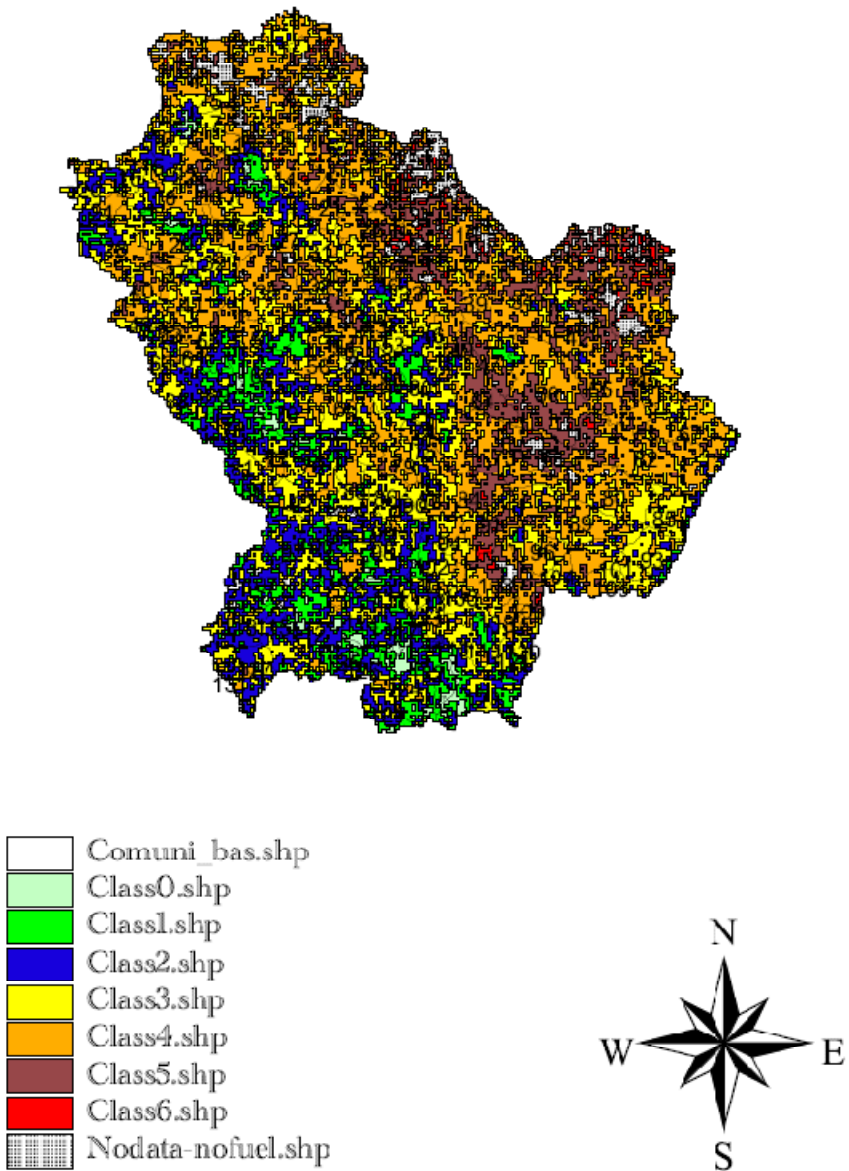


Fig. 7. Example of fire danger maps provided to the Balicata Protezione civile for daily fire alert

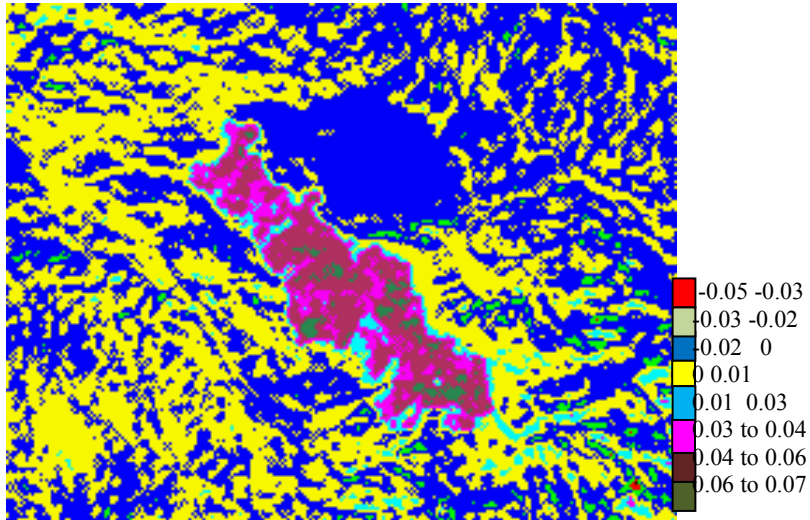


Fig. 8. Show an example of fire severity map obtained from ASTER

The characterization of vegetation dynamics before and after fire occurrence is performed by using detrended fluctuation analysis (DFA) applied to a MODIS time series. The DFA has already been applied to data from SPOT-VEGETATION (Telesca et al.) but, within the FIRE-SAT project DFA is generally applied to MODIS time series, which, as expected, provide significant improvements due the higher spectral (36 spectral bands ranging from 0.4 micrometers to 14.4 micrometers) and spatial resolutions (acquisition from 250 m to 1 km). As in the case of fire severity, also for fire resilience a number of diverse vegetation indices were processed using DFA to assess which one performs better. In particular, NDVI, MSI NDWI and the NBR time series were compared. The values obtained for the four investigated indexes show that the NDVI is the vegetation index which best enable us to better discriminate burned areas from un-burned.

All the values of scaling for fire affected and affected are larger than $1/2$ which means that the fluctuations of time series are persistent that the values of the indices are positively correlated. The estimated scaling exponents of both condition (with and without fire) suggest a persistent character of the vegetational dynamics. But, some indices exhibit a discriminable behavior before and after fire occurrence other did not show any significant difference. In particular, the two moisture vegetation indices herein examined, MSI and NDWI, have showed the same behavior for both the two sites (burnt and un-burnt). As a whole, results from the DFA showed that the NDVI (Normalized Difference Vegetation Index) seems to be more effective than other indices, as also found in previous studies (2011 see Lanorte et al. EGU 3690) and NBR (Normalized Burn Ratio) see Montesano et al. (2011 abstract EGU 3623).

This behavior suggested that, the fire scaling exponent of NDVI can be used to assess and quantify the fire resilience, namely the capability of natural vegetation to

recover after fire. In the case of agriculture areas, all the indices do not provide significant difference before and after fire and this due to the fact that the vegetation dynamics are those induced by agricultural practice and it is not possible to see any recovery being not natural i.e. unmanaged

4 Conclusion

The available satellite time series may allow us to develop standardized methodologies for evaluating fire danger and effects. The new approached we devised to estimate fire danger and fire severity are locally self-adaptive and, therefore, suitable to overcome the drawbacks linked to the traditional methods based on fixed threshold values. The model could be incorporated directly into the mapping process from local up to global scale, being that it is independent on the sensors used for the evaluation as well as on geographic regions and vegetation cover types affected by fire. In particular, the prompt availability of fire severity map may be very useful in addressing post fire management activities. Thus also overcoming limits linked to the traditional methods of recording the fire effect disturbance which involve expensive and time-consuming field survey. Our results reveal that fires contribute in increasing the persistence of time dynamics of vegetation. Our finding suggests a new approach to be fruitfully adopted in operational context for fire monitoring, with particular reference to the danger estimation and the fire alert which is reliable done at sub-municipality municipality, with a spatial detail, less than 250 m (pixel size of MODIS greenness) which must be further improved with a rescaling at 15 m according to the ASTER fuel map availability.

Acknowledgment. The authors thahsk the Basilicata Protezione Civile, particularly Guido Loperte and Giuseppe Basile.

References

1. UNCCD, United Nations Convention to Combat Desertification, report, Paris (1994)
2. FAO, Global forest fire assessment 1990-2000. Forest Resources Assessment Programme, working paper no. 55 (2001), http://www.fao.org:80/forestry/fo/fra/docs/Wp55_eng.pdf
3. Lasaponara, R.: Inter-comparison of AVHRR-based fire danger estimation methods. *International Journal of Remote Sensing* 26(5), 853–870 (2005)
4. Lasaponara, R., Lanorte, A.: VHR QuickBird data for fuel type characterization in fragmented landscape. *Ecological Modelling* in press (ECOMOD845R1) 204, 79–84 (2007)
5. Lasaponara, R., Lanorte, A.: Remotely sensed characterization of forest fuel types by using satellite ASTER data. *International Journal of Applied Earth Observations and Geoinformation* 9, 225 (2007)
6. Lasaponara, R., Lanorte, A.: Multispectral fuel type characterization based on remote sensing data and Prometheus model. *Forest Ecology and Management* 234, S226 (2006)

7. Lasaponara, R., Cuomo, V., Macchiato, M.F., Simoniello, T.: A self-adaptive algorithm based on AVHRR multitemporal data analysis for small active fire detection. *International Journal of Remote Sensing* 24(8), 1723–1749 (2003)
8. Lasaponara, R.: Estimating Spectral separability of satellite derived parameters for burned areas mapping in the Calabria Region by using SPOT-Vegetation data. *Ecological Modelling* 196, 265–270 (2006)
9. Lanorte, A., Danese, M., Lasaponara, R., Murgante, B.: Multiscale mapping of burn area and severity using multisensor satellite data and spatial autocorrelation analysis. *International Journal of Applied Earth Observation and Geoinformation* (2012), doi:10.1016/j.jag.2011.09.005
10. Telesca, L., Lasaponara, R., Lanorte, A.: 1/f fluctuations in the time dynamics of Mediterranean forest ecosystems by using normalized difference vegetation index satellite data. *Physica A* 361(2), 699–706 (2006)
11. Telesca, L., Lasaponara, R.: Fire-induced variability in satellite SPOT-VGT NDVI vegetational data. *International Journal of Remote Sensing* 27, 3087–3095 (2000)
12. Geary, R.: The contiguity ratio and statistical mapping. *The Incorporated Statistician* (5) (1954)
13. Getis, A., Ord, J.: The analysis of spatial association by distance statistics. *Geographical Analysis* 24, 189–206 (1992)
14. Anselin, L.: Local indicators of spatial association – LISA. *Geographical Analysis* 27, 93–115 (1995)

Investigating Satellite Landsat TM and ASTER Multitemporal Data Set to Discover Ancient Canals and Aqueduct Systems

Rosa Lasaponara and Nicola Masini

CNR-IMAA, C.da S. Loja 85050 Tito Scalo (PZ), Italy

lasaponara@imaa.cnr.it

CNR-IMAA, C.da S. Loja 85050 Tito Scalo (PZ), Italy

n.masini@ibam.cnr.it

Abstract. In this paper, we focus on the use of the Landsat and ASTER multitemporal data set for extracting information on ancient irrigation systems and artificial wet agro-ecosystems. The study area is the Nazca basin in Southern Peru selected mainly for its extreme drought. Despite these critical environment conditions, the area was populated since millennia ago thanks to adequate survival strategies developed by ancient Nazca populations. To cope with hostile environmental factors and water scarcity, efficient aqueduct systems, today called puquios, were devised and some of them are still in use today. The main purpose of our investigations was the identification of buried unknown puquios by using satellite multitemporal maps of vegetation indices and moisture content. Results from satellite data were also identified on the ground, checked and confirmed in situ. The successful results obtained in the Nazca Basin suggest that our methodological approach can be efficiently re-used in a number of areas, characterized by similar environmental conditions and long human frequentation.

Keywords: GIS, satellite based Analysis, ancient irrigation systems, Spatial variation, Moisture index, vegetation index, Nazca (Peru).

1 Introduction

Precious information to reconstruct ancient environmental features, still fossilized in the present landscape, may be captured from active (Radar) and passive (Imaging) remotely sensed data today available from medium to high spatial resolution. In particular, satellite derived moisture content may facilitate the identification of areas involved in early environmental manipulation mainly addressed to set up irrigation and artificial wet agro-ecosystems where the natural rainfall was insufficient for domestic use or to support agriculture, etc.. Up to now, only a few number of archaeological and local palaeo-environmental studies have been conducted on ancient irrigation systems, mainly on the basis of in situ data analysis. The current availability of satellite radar and optical data can enable us to set up systematic

investigations in different areas in the world using "quite low cost" technologies which can also provide useful tools for monitoring the investigated area and contribute to the preservation of the sites. Nevertheless, up to now, only a few number of studies based on satellite data (see, for example [1] [2] and references therein quoted) have been addressed to detect systematically ancient environmental changes, still fossilized in the present landscape.

In this paper, we exploited multitemporal and multisensor satellite data to extract information and detect remains of ancient irrigations in the Nazca basin (Peru). We consider this area particularly interesting mainly because it was populated since millennia ago despite its critical environment due to the extreme aridity [3,4,5,6]. To face these extreme conditions a very efficient system for retrieval water was devised by the ancient populations of the Nazca River valley. Underground aqueducts, called in Spanish *puquios*, were set up and systematic maintenance activities were carried out over centuries and millennia, so that some *puquios* are still in use today. This aqueduct system is today considered very advanced because it provided not only water for religious ritual and agricultural needs, but also for domestic uses thanks to a filtration system which provided drinkable water. Up to now, several in-situ analyses have been undertaken in the framework of a number of studies [7-12], but only recently satellite data have been used [1] to carry on systematic studies exploiting the synoptic view offered by remote sensing.

This paper focuses on the identification of buried unknown *puquios* through satellite multitemporal vegetation and moisture maps. They provide detailed information on the seasonal availability and variation of surface and subsurface water. The unknown *puquios* we detected were also confirmed by ground survey.

The methodological approach we applied in the Nazca Basin can be efficiently re-used in a number of areas in Meso-America, Middle East, North Africa, Asia, historically characterized by similar environmental conditions and long human frequentation; so that similar aqueduct systems, such as the Karst (Maya-MesoAmerica) and kanat (Middle east, Asia, Africa) were devised to face drought and retrieve water. These ancient systems may be re-used today to improve modern land use and land cover in desert or drought areas and to combat desertification processes.

2 Study Area: History and Environmental Setting

The area of our investigation, shown in Figure 1, is close to the Andean foothills, composed of Jurassic and Cretaceous formations (see for example, Montoya [6]) and spreads out on a Pre-Montane desert formation near the coastline, geologically located on Quaternary sedimentary rock formations.

The geomorphology of the coastal desert, has been shaped by the drainage basin of the Río Grande. It empties into the Pacific Ocean after passing through the coastal range and collecting water from nine tributaries.

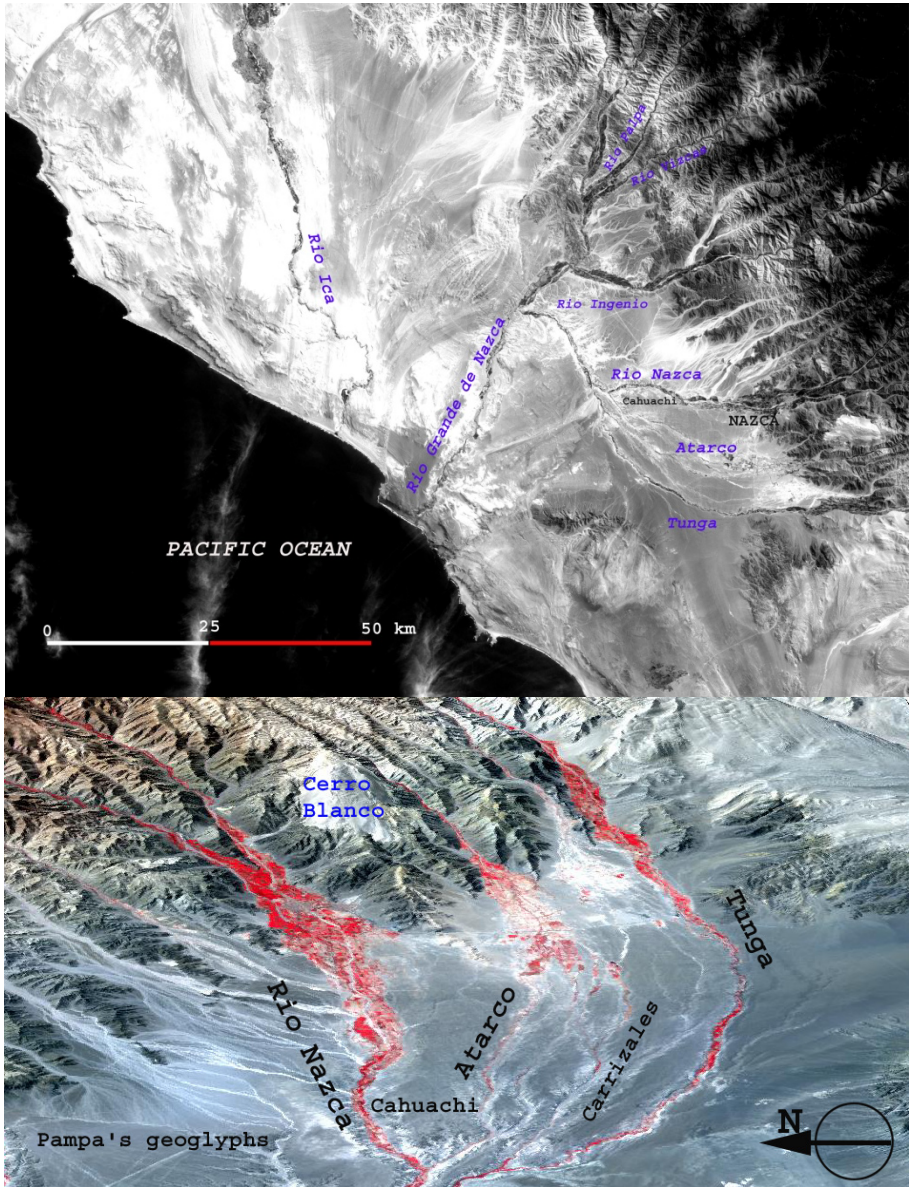


Fig. 1. The image shows the drainage basin of the Rio Nazca including the tributaries Atarco, Carrizales and Tunga. Up from below: the drainage basin of Rio Grande de Nazca and Rio Ica, from Landsat ETM+; band 1; and the 3D false colour composition from Aster data.

The area of concern is the drainage basin of the Rio Nazca, considered one of the most arid areas of the world due to a cold ocean current (the Humboldt Current) and other climatic factors[10]. Even if the environmental conditions were the same as today over last centuries, the area was, surprisingly, densely settled in pre-columbian times as clearly testified by archaeological evidence [4,5,11].

Over the last millennia (since about 5000 BC), the Nazca River and its main tributaries were essential for human occupation and the complex societies established in this area. The southern tributaries of the Nazca River (see Figure 1), fed by the seasonal precipitations coming from the Andes (at an altitude higher than 2,000 meters above sea level), are substantially smaller than those of the northern tributaries and those from other coastal valleys.



Fig. 2. Puquios in Cantalloc

The lack of water was (and still is) due to two main factors: (i) the scarce pluvial precipitations and the (ii) high infiltration capacity, which causes yearly a significant reduction of the surface water. Over the millennia, long periods of drought occurred and frequently the lack of water was persistent for several years.

As proved by archaeological findings [10], during the Early Intermediate Period (known also as Middle Nazca) Nazca population devised an efficient water retrieval system, based on underground aqueducts today called *puquios*. Subsurface water table

is, and likely was, very close to the surface level and, therefore, quite easily accessible.

Puquios indicate aqueduct systems based on horizontal water wells, trench and/or underground gallery that connects a place on the surface with the ground water source (figure 2). For additional details the reader is referred to Lasaponara and Masini [1].

Exploiting puquios technology for thousands of years, populations have successfully adapted to water shortage. For these reasons, throughout the years, several researchers have investigated puquios, see for example [1, 7-12]. Studies from Lancho Rojas and Schreiber [7] started up back since 1985 and continue as of today.

3 Satellite Based Analysis

Satellite derived parameters, such as temperature and moisture as well as single channel behavior (see figure 3) with their spatial and temporal pattern variations, can help us to extract precious information to reconstruct ancient environmental changes still fossilized in the present landscape, see, for example Lasaponara and Masini [1].

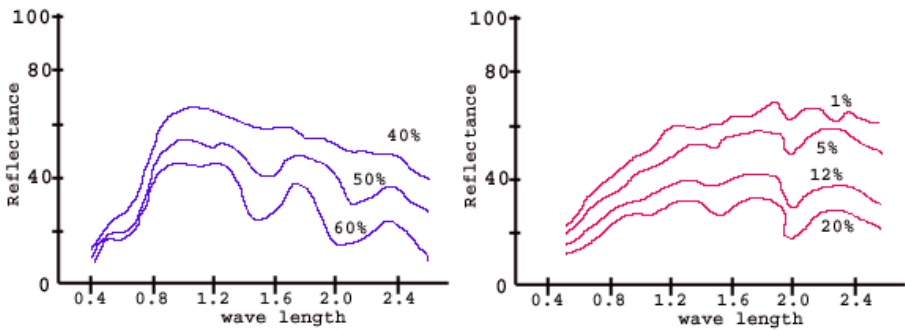


Fig. 3. Typical reflectance curves according to moisture content variations for vegetation (left) and soil (right)

Our investigations focus on the Nazca drainage basin for which we captured spatial and temporal variations of both vegetation and soil moisture using a multispectral, multi-temporal satellite data set made up of medium-resolution images. To cover a larger time window we analyzed both Landsat TM, ETM+ and ASTER which provide spectral channels useful for soil and vegetation moisture variation and monitoring.

Surface reflectance is sensitive to moisture content being that water tends to reduce the signal, thus resulting in a reduction of reflectance as the moisture increases, as evident from the spectral behavior of both soil and vegetation (Figure 3).

Along with the use of single channels, another useful approach can be based on the use of spectral index computation and their multitemporal spatial pattern. Our assumption is that changes in vegetation and moisture content are greater for shallow water table compared to arid soils or irrigated farming fields.

Using spectral channel combination, we computed two spectral indices: the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI). NDVI arithmetic combination of the red and Near infrared channels according the following formula

$$NDVI = \frac{(NIR-R)}{(NIR+R)} \quad (1)$$

NDVI enables us to exploit the different spectral behaviour of vegetation cover in the different bands. Higher is the NDVI values more healthy is the vegetation.

NDWI is computed to estimate moisture content for both soil and vegetation, by using also SWIR bands, according the formula (2)

$$NDWI = \frac{(RED-SWIR)}{(RED+SWIR)} \quad (2)$$

According to formula 2, low moisture content should be characterized by high NDWI value and vice versa.

The investigations based on remote sensing have been carried out also using Tasselet cup transformation (TCT), computed on the multispectral Landsat images to better discriminate the vegetation from desert areas and to emphasize the spatial variation of moisture content from the most drought areas. The TCT is a fixed coefficient linear transformation, based on empirical analysis of physical features space, proposed by Kauth and Thomson [14] mainly for agricultural monitoring. The computation of TCT provides a new reference system composed of six components, among which the first three are: 1) "soil brightness" defined by the signature of non vegetated areas, 2) "greenness" obtained from vegetation signatures, and 3) wetness, which measures the interrelationship of soil and canopy moisture, respectively.

Our analysis was conducted using a multitemporal data set made up of medium and high resolution (HR) satellite images: Landsat TM (1990), Landsat 7 ETM+ (2000) and ASTER (2003, 2004 and 2007).

The Landsat satellite program was designed mainly for vegetation monitoring. The ETM+ provides data in eight channels: some available at 30 m spatial resolution such as four in the visible/near infrared (VNIR) and two in the short wave infrared (SWIR), one in the thermal range at 90 m of resolution, and, finally, a panchromatic scene with at a resolution of 15 m.

The available HR satellite datasets have been processed by using the most adequate algorithms according to the available spectral bands. Herein, we will focus on Landsat/7 images, acquired at the end of April 2000 and on a multitemporal dataset of ASTER acquired on June of 2003, 2004 and 2007, during the dry season. The available data set should enable us to characterize the spatial and temporal moisture pattern, discriminating the differences in moisture content along the tributaries of the Nazca river characterized by an ephemeral hydraulic regime.

Data processing was carried out according to Lasaponara and Masini [1]. Additional data analyses herein performed were mainly based on the characterization of space /temporal behavior of the investigated parameters using the time average and standard deviation to capture the spatio/temporal variability of the investigated parameters.

4 Results

Results from data processing of NDVI and moisture maps enabled us to characterize the dynamics of surface moisture content and to extract information about ancient canals and aqueducts. Spatial and temporal patterns of surface moisture content is one of the most significant parameters for capture information on ancient landscapes because historically civilizations have predominately located their communities close to water ways for their basic survival, agriculture, ritual and domestic use.

Figure 4 (a,b) shows the RGB composition of Landsat ETM+bands 3-2-1 and the three components (brightness, greenness and wetness) of TCT applied to the TM data, respectively. From the measures of TCT components we identified the following ranges of values for four different types of soil surfaces, such as : a) Pampa (see a in Figure 4a); b) arid beds of quebradas (see b in Figure 4a); c) partially vegetated cover of quebradas; d) fluvial oasis of the Nazca river (see a, b, c, and d in Figure 4a, respectively; and the relative histograms in Figure 5):

- a) In the Pampa at North of Nazca river and in the arid areas crossed by the quebradas (ravines) Atarco, Tariga, Carrizales and Tunga, the TCT component values range from 187 to 207 for the brightness (with maximum value around to 197), -160 to - 123 for the greenness and -105 to -74 for the wetness (max value around to -95).
- b) On the arid bed of quebrada, brightness values are higher (190 to 265, with maximum value around to 205), as well as greenness (-146 to -118) , whereas wetness values (-121 to -89) are lower respect to those measured in the Pampa (see Figure 5). The chromatic result of RGB composition (see Figure 4b) allows us to extract not only the arid beds of the ravines but also the traces of ancient and more recent *huaycos* which are essentially flash floods caused by torrential rains occurring in the mountains that carry rocks with them down. In figure 4b it is possible to observe the *huaycos* coming from the Sierra at North, crossing the Pampa of geoglyphs. Such mudslides generally stop in the Nazca riverbed thus protecting the Ceremonial Centre of Cahuachi, sited at South of the river. However it cannot exclude that in the past mudslides and flooding beyond the left bank of the river as showed in Figure 4b. This would prove the vulnerability to flash flooding of Cahuachi, in agreement with the archaeological record according which the decadence and the abandonment of Ceremonial Center would be caused also by two very destructive flash floods.

- c) The four quebradas at South of Nazca river are characterized by a partially vegetation cover, probably due to the shallow water table. This is confirmed by the measure of TCT components which put in evidence higher values of greenness and wetness (-136 to -75, and -115 to -72, respectively) compared to the arid bed of the quebrada. The RGB visualization makes the identification of such patterns very easily. In Figure 4b the black rectangular boxes denote some areas characterized by the above mentioned TCT value intervals, which are typical of an ephemeral idraulic regime that allowed the irrigation for farming just for a few months during the wet season (typically from the end of February to the first half of April). In such cases, the puquios could provide additional water resources for irrigation for more months over the year.
- d) Finally, the measure of TCT components on the Nazca fluvial oasis put clearly in evidence higher values of greenness and wetness (-75 to -38, and -95 to -70, respectively) respect the partially vegetation covers of some segments of quebradas, likely characterized by a shallow water table.

In Figure 4b, black boxes indicate those segments of the *quebradas* characterized by greenness and wetness values higher respect to the vegetated areas and lower respect to the arid surface of the Pampa, due to the presence of shallow water table. White and black arrows denote the traces of past mudslides of the Nazca river and *huaycos* coming from the Pampa of geoglyphs at North of the river, respectively.

It is worth to note that the investigated area from the pampa to the beds of the ravines are mostly arid and, therefore, visualized in RGB with black colour. However, we can identify some areas characterized by changes of NDVI over time in particular in the fluvial oasis of the Nazca river and in the upper part of the tributaries Taruga, Carrizales, Tunga and Atarco, where the presence of vegetation is due to shallow table water and puquios in use.

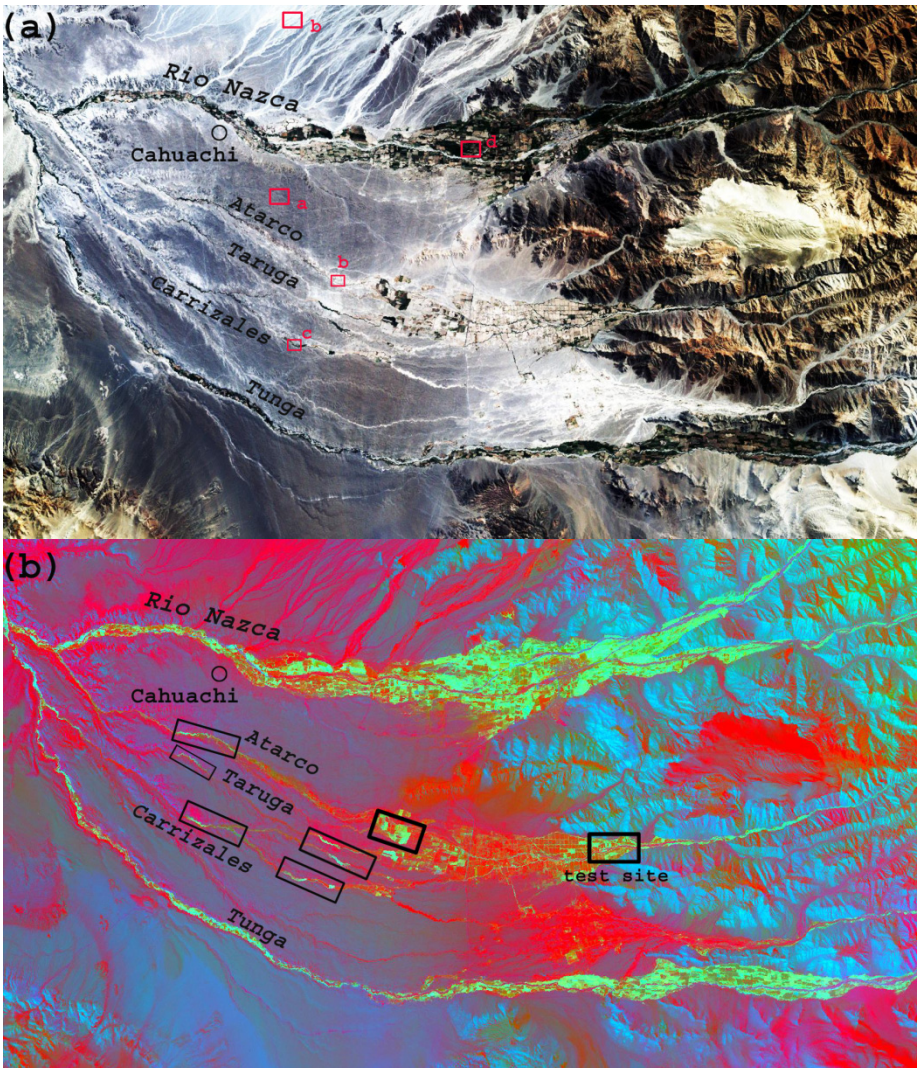


Fig. 4. (a) RGB composition of Landsat ETM+bands 3-2-1 (b) RGB composition of TCT components. Red boxes in Figure 3a, denote the masks, related to four different land cover types, histograms of TCT components have been in evidence higher values of greenness and wetness (-75 to -38, and -95 to -70, respectively) respect the partially vegetation covers of some segments of quebradas, likely characterized by a shallow water table.

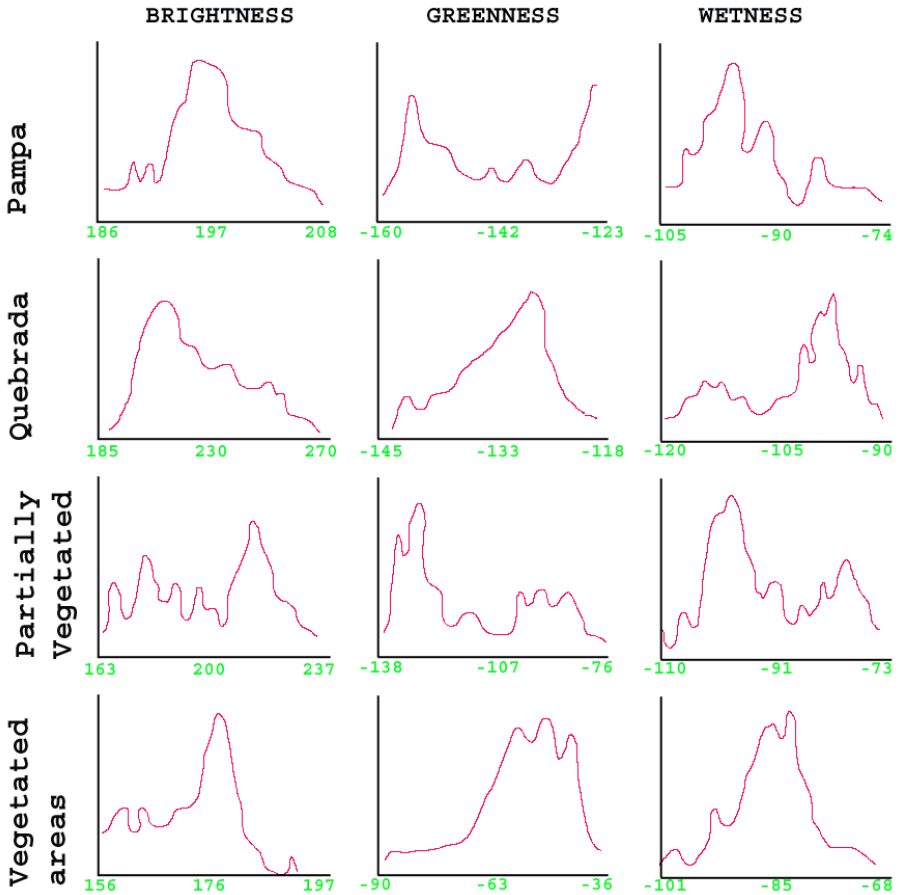


Fig. 5. Histograms of the three TCT component values measured on masks located on the four land cover types present in the study area: (i) Pampa, (ii) *quebradas*, (iii) *quebradas* with partially vegetated cover, and (iv) vegetation cover of the Nazca fluvial oasis.

Figure 6 shows NDVI maps which allow us to extract information about flow characteristics of the rivers at South of the Nazca basin. In particular, Figure 6a show the RGB composition of multitemporal NDVI maps obtained from the first to the last considered year (R:2003, G:2004 and B:2007). Figure 6b shows the standard deviation map computed from the multitemporal NDVI which puts in evidence the spatial variability along with the temporal patterns: constant low values (compared to the observed behavior) over time are characterized by darker tone to black, high variability (compared to the observed behavior) are visualized by lighter tone. Finally, Figure 6c enhances a classification of the NDVI variability.

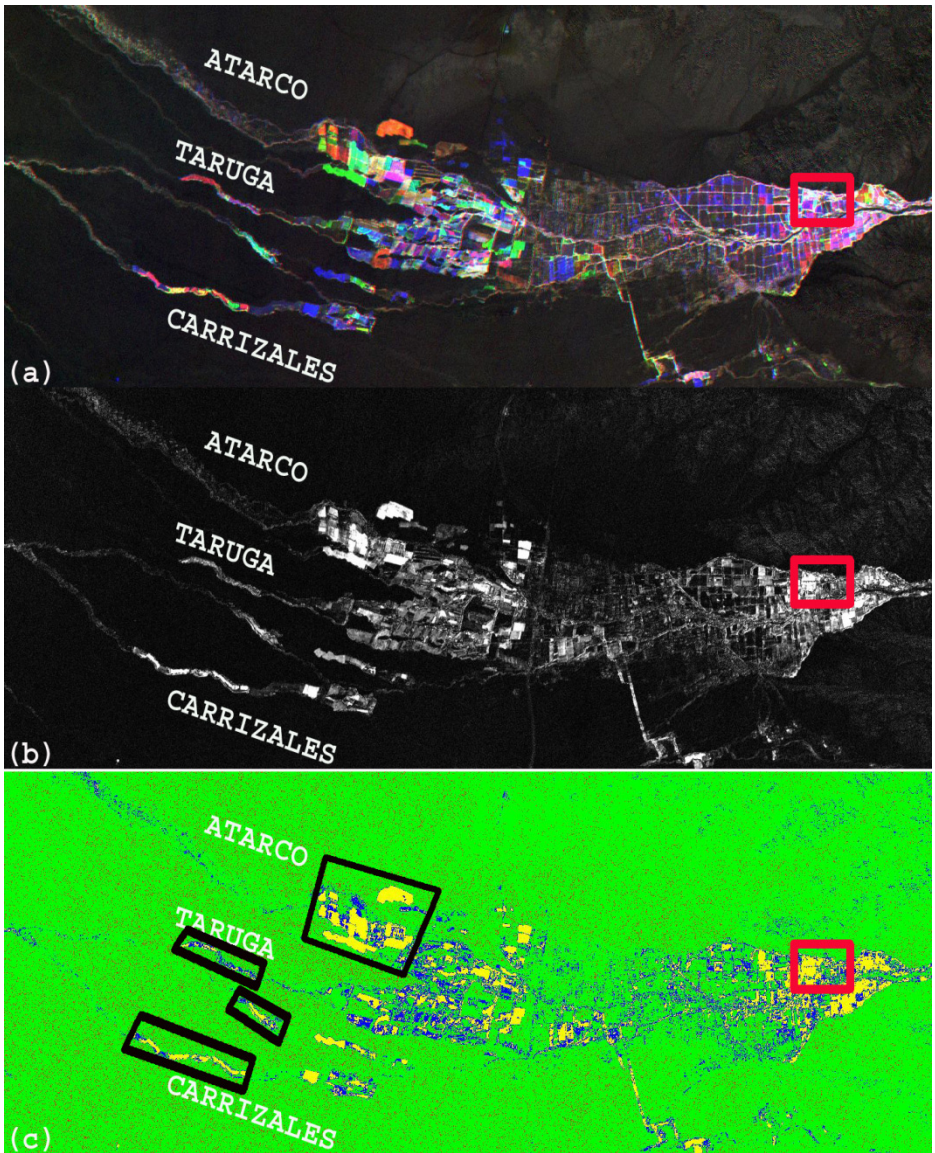


Fig. 6. (a) RGB (2003, 2004 and 2007) composition of NDVI Maps; (b) detail of standard deviation of NDVI maps: constant low values (compared to the observed behavior) over time are characterized by darker tone to black, high variability (compared to the average) are visualized by lighter tone, (c) classification of the NDVI temporal variability: green, blue yellow indicate increasing variability, respectively.

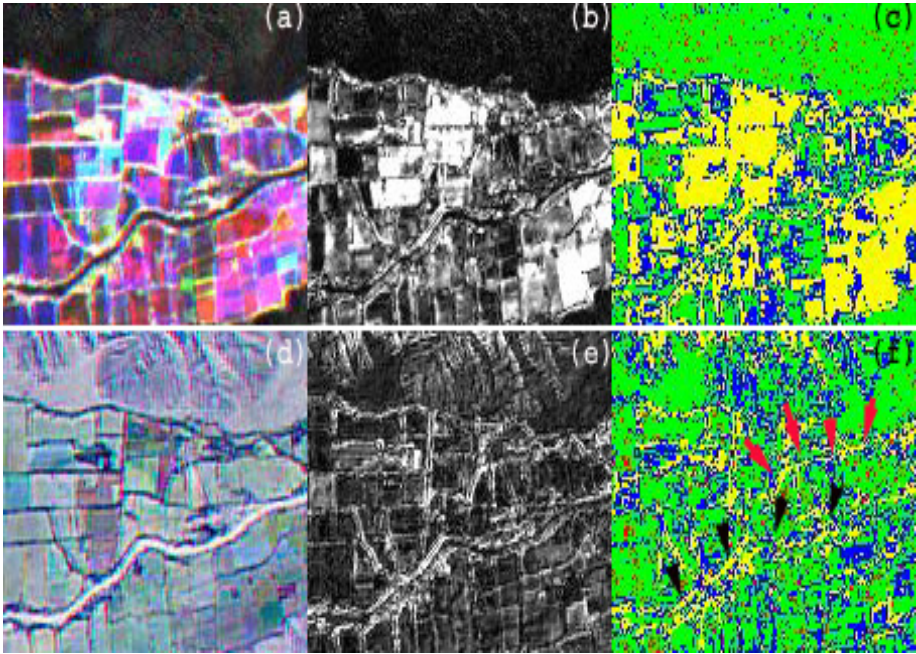


Fig. 7. shows RGB composition of both NDVI(a) and NDWI (d) maps, along with their standard deviations and classification of standard deviation maps.

Figures 7 show RGB compositions for a subset of both NDVI(a) and NDWI (d) maps which also allow us to extract information on flow characteristics of the rivers at South of the Nazca basin. Both parameters (NDVI(a) and NDWI (d) maps) exhibit a coherent behavior. In particular, from the RGB composition (Figure 7d) of NDWI (R=2003, G=2004, B=2007) we can distinguish three different spectral behaviors:

- (i) One is related to pixels characterized by constantly low NDWI values (from -0.1 to 0, over time associated with perennial rivers).
- (ii) The second spatial/temporal pattern, visualized as grey, is related to pixels with the highest NDWI values (0.25-0.30) and, therefore, we associate them with arid soil which can be found in the Pampa or in the beds of ravines characterized by deep water table.
- (iii) Finally, the third feature pattern, characterized by a spatial and temporal “dispersion” in NDWI values, is highlighted by the multitemporal behaviour. In other words, in Figure 7b, from one year to another the NDWI values change within a buffer which denotes that water table is not very deep (visualized by a colour with different intensity value of red, green and blue).

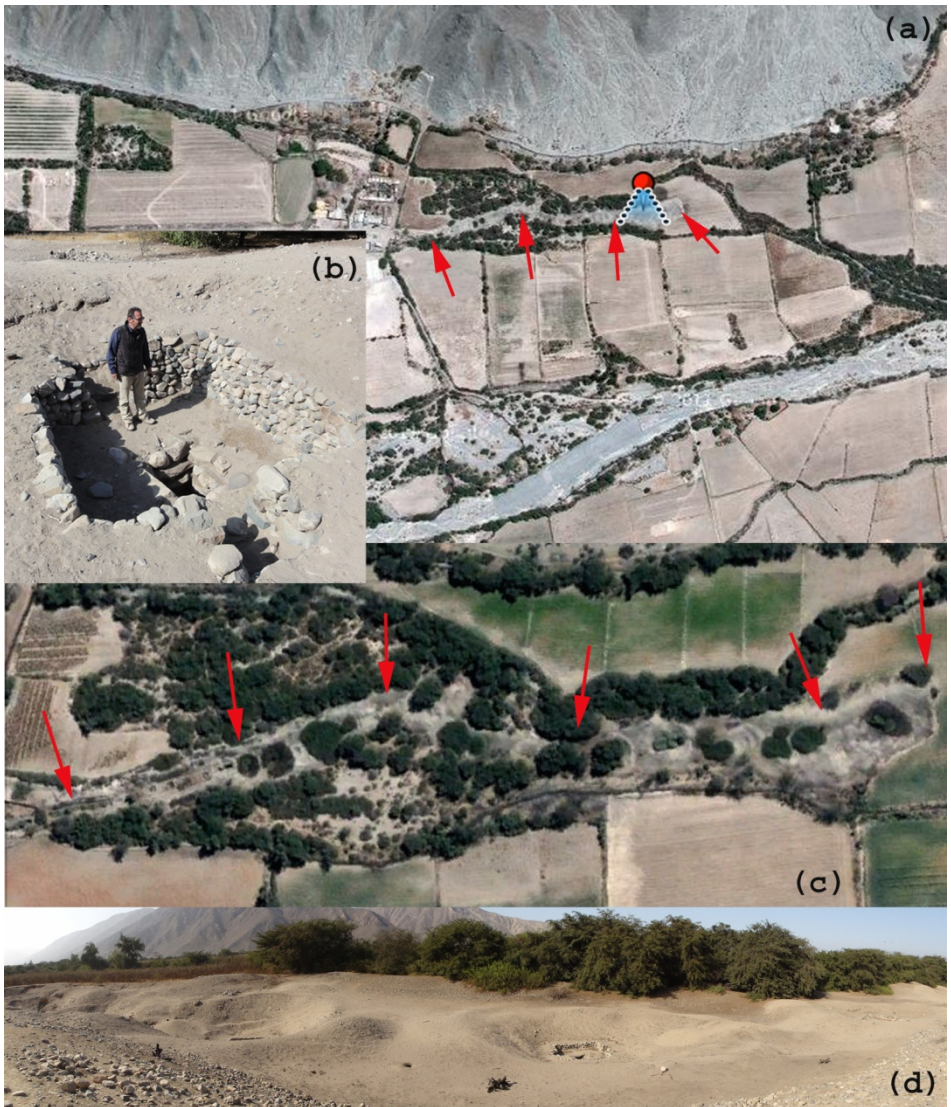


Fig. 8. Underground aqueduct (puquio) of Santa Maria near Rio Taruga. (a) image drawn by Google Earth; (b) detail of a chimney (known as ojo in Spanish) which allows access to the puquio for cleaning as well as the entrance of air and sunlight; (c) detail of the same puquio from a GeoEye image acquired on 2011 February 28; (d) photo of a puquio: in evidence the “ojos”.

The analysis of NDWI and NDVI variations over time allowed us to identify the hydraulic regime (perennial, ephemeral, dry) along each tributary of Rio Nazca. In particular, the survey of parts of the ravines or streams characterized by ephemeral

regime, that is fluctuation of water table, is crucial not only for the study of palaeo-hydrography but also for the detection of ancient hydraulic systems, such as the Nazca puquios.

The crossed observation of NDWI and NDVI maps put in evidence some areas where it is very likely to find puquios. One of them is located in Santa Maria near Rio Taruga well visible from very high resolution satellite data. Figures 8 shows (a) an image drawn by Google Earth; (b) detail of a chimney which allow the access to the puquio for their cleaning as well as the entrance of air and sunlight; (c) detail of the same puquio from a GeoEye image acquired on 2011 February 28; (d) photo of a puquio: in evidence the “ojos”.

5 Conclusions

In the framework of investigations aimed at supporting archaeological investigation in the Nazca Basin, vegetation indices and moisture content maps obtained from both TM and ASTER data, along with their spatial patterns and variations, were used to improve the present knowledge about the ancient Nazca technologies used for water retrieval.

This is a crucial point not only to detect unknown ancient settlement remains but also to understand the ability of the ancient Nazca populations in developing effective and complex agro-ecosystems to meet the demands of a large population. Archaeologists found evidence that the area could not have been inhabited without the existence of these aqueducts. Results from the satellite based analysis were verified and confirmed by a field survey in Santa Maria near Rio Taruga.

Our methodological approach, successfully adopted in the Nazca river basin, can be also easily re-applied to a number of diverse geographical areas, characterized by similar environmental conditions, such as –Meso America, Middle East, Asia, Africa, where ancient populations devised and set up similar systems for water retrieval to face drought. These systems may be re-used today to combat desertification in desert or drought areas.

The study of ancient water retrieval and management systems will be beyond their hydrogeological and physical characteristics, being that in different areas of worlds all these systems allowed human settlements, leaded, and influenced social rules, cultural activities and economical exploitation of them.

References

1. Lasaponara, R., Masini, N.: Following the ancient Nazca puquios from space. In: Lasaponara, R., Masini, N. (eds.) *Satellite Remote Sensing a New Tool for Archaeology*, pp. 269–290. Springer (2012)
2. Stargardt, J., Amable, G., Devereux, B.: Irrigation Is Forever: A Study of the Post-destruction Movement of Water Across the Ancient Site of Sri Ksetra, Central Burma. In: Lasaponara, R., Masini, N. (eds.) *Satellite Remote Sensing a New Tool for Archaeology*, pp. 247–267. Springer (2012)

3. Masini, N., Lasaponara, R., Rizzo, E., Orefici, G.: Integrated Earth observation methods in Cahuachi (Peru): studies and results of the ITACA Mission (2007-10). In: Lasaponara, R., Masini, N. (eds.) *Satellite Remote Sensing: a New Tool for Archaeology*, pp. 307–344. Springer, New York (2012)
4. Orefici, G.: Cahuachi, the largest adobe Ceremonial centre in the world. In: Nazca. *El desierto del los Dioses de Cahuachi*, Graph, Lima, pp. 36–59 (2009)
5. Orefici, G.: El Proyecto Azca. In: Nazca. *El desierto del los Dioses de Cahuachi*, Graph, Lima, pp. 18–35 (2009)
6. Montoya, M., Gracia, W., Caidas, J.: *Geología de los Cuadrangulos de Lomitas, Palpa, Nazca y Puquio, INGGEMET (Istituto Geologico Minero y Metallurgico)*, Lima (1942)
7. Mejia, X.T.: *Acueductos y Caminos Antiguos de la Hoya del Río Grande de Azca*. In: *Actas y Trabajos científicos del XXVII Congreso Internacional de Americanistas*. Librería e Imprenta GIL, vol. 1, pp. 559–569 (1942)
8. Schreiber, K.H.: *Irrigation and Society in the Peruvian Desert: The Puquios of Azca*. Lexington Books, Lanham (2003)
9. Schreiber, K.H., Lancho, R.J.: Los puquios de Nazca: un sistema de galerías filtrantes. In: *Boletín de Lima*, vol. 59, pp. 51–62. Editorial Los Pinos, Lima (1988)
10. Schreiber, K.H., Lancho, R.J.: El control del agua y los puquios de Azca. Nazca. *El desierto del los Dioses de Cahuachi*, Graph, Lima, pp. 132–151 (2009)
11. Silverman, H.: *Cahuachi in the Ancient Nazca World*. University of Iowa Press, Iowa City (1993)
12. Solar La Cruz, F.: *Nazca Filtering Galleries; galerías filtrantes*. Universidad Abraham Valdelomar, Lima (1997)
13. Lasaponara, R., Masini, N.: Image enhancement, feature extraction, and geospatial analysis in an archaeological perspective. In: Lasaponara, R., Masini, N. (eds.) *Satellite Remote Sensing: a New Tool for Archaeology*, pp. 17–63. Springer, New York (2012)
14. Kauth, R.J., Thomas, G.S.: The Tasseled Cap – a graphical description of the spectral-temporal development of agricultural crops as seen by Landsat. In: *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, Purdue University, West Lafayette, Indiana, pp. 4B41–4B51 (1976)

Using Spatial Autocorrelation Techniques and Multi-temporal Satellite Data for Analyzing Urban Sprawl

Gabriele Nolè^{1,3}, Maria Danese², Beniamino Murgante³, Rosa Lasaponara^{1,3},
and Antonio Lanorte¹

¹ Istituto di Metodologie per l'Analisi Ambientale (IMAA), CNR,
C.da S.Loja, 85050 Tito (PZ), Italy

² Istituto per i Beni Archeologici e Monumentali (IBAM), CNR,
C.da S.Loja, 85050 Tito (PZ), Italy

³ Università degli Studi della Basilicata, Viale dell'Ateneo Lucano, 10, 85100,
Potenza, Italy

{gabriele.nole, lasaponara, lanorte}@imaa.cnr.it,
maria.danese@ibam.cnr.it, beniamino.murgante@unibas.it

Abstract. Satellite time series offer great potential for a quantitative assessment of urban expansion, urban sprawl and for monitoring of land use changes and soil consumption. This study deals with the spatial characterization of expansion of urban areas by using spatial autocorrelation techniques applied to multi-date Thematic Mapper (TM) satellite images. The investigation focused on several very small towns close to Bari. Urban areas were extracted from NASA Landsat images acquired in 1976, 1999 and 2009, respectively. To cope with the fact that small changes have to be captured and extracted from TM multi-temporal data sets, we adopted the use of spectral indices to emphasize occurring changes, and spatial autocorrelation techniques to reveal spatial patterns. Urban areas were analyzed using both global and local autocorrelation indexes. This approach enables the characterization of pattern features of urban area expansion and it improves land use change estimation. The obtained results showed a significant urban expansion coupled with an increase of irregularity degree of border modifications from 1976 to 2009.

Keywords: Urban morphology, Remote sensing, Autocorrelation, Change Detection.

1 Introduction

Urbanization and industrialization are the key factors for social and economical development and represent specific response to economic, demographic and environmental conditions. In many European regions abandonment of agricultural land has induced a high concentration of people in densely populated urban areas during the last few decades. This phenomenon has been observed throughout the world. In 1950,

only 30% of the world's population lived in urban areas. By 2000 that proportion rose up to 47%, and by 2030 the estimated number will be around 60% [25].

Such a rapid industrialization and expansion of urban areas has caused strong and sharp land cover changes and significant landscape transformations, which significantly impact local and regional environmental conditions. Nowadays, the increase of concentration of people in densely populated urban areas is considered as a pressing issue in developing countries. For example, following land reform initiated in 1987, vast areas of China have been involved in a rapid urban expansion and new urban settlements [4], so that in a few years, several cities rapidly have become big centres or regional nodes.

The analysis of city size distribution deals with different disciplines such as geography, economy, demography, ecology, physics, statistics, etc., because the evolution of a city is a dynamic process involving a number of different factors. An issue of great importance in modelling urban growth includes spatial and temporal dynamics, scale dynamics, man-induced land use changes. Although urban growth is perceived as necessary for a sustainable economy, uncontrolled or sprawling urban growth can cause various problems, such as loss of open space, landscape alteration, environmental pollution, traffic congestion, infrastructure pressure, and other social and economical issues. To face such drawbacks, a continuous monitoring of urban growth evolution in terms of type and extent of changes over time is essential for supporting planners and decision makers in future urban planning.

Many recent researches have also explored ways of measuring dynamics of urban morphology. Shen [22], among others, compared the morphology of 20 urban areas in USA obtaining a wide range of results, due to the different size and character of each case study. Also, Frankhauser [6] used fractal dimension in the examination of outskirt areas in European cities, trying to obtain a typology of urban agglomerations. Finally, Benguigui et al. [2] examined the built-up settlement of Tel Aviv and concluded that fractal dimension tends to increase through time.

A critical point for understanding and monitoring urban expansion processes is the availability of both (i) time-series data set and (ii) updated information relating to current urban spatial structure to define and to locate evolution trends. In such a context, an effective contribution can be offered by satellite remote sensing technologies, which are able to provide both an historical data archive and up-to-date imagery. Satellite technologies represent a cost-effective mean for obtaining useful data that can be easily and systematically updated worldwide. Nowadays, medium resolution satellite images, such as Landsat TM or ASTER can be downloaded free of charge from NASA web site.

The use of satellite imagery along with spatial analysis techniques can be used for monitoring and planning purposes as these enable the reporting of ongoing trends of urban growth at a detailed level. Nevertheless, exploitation of satellite Earth Observation in the field of urban growth monitoring is a relatively new tool, although during the last three decades great efforts have been addressed to the application of remote sensing in detecting land use and land cover changes. A number of investigations were carried out using different sets of remotely sensed data [23] [14]

[18] [12] [21] and diverse methodological approaches to extract information on land cover and land use changes.

This study analyzes urban expansion over time in several towns of southern Italy, using satellite images. Sample towns are located south of Bari, one of the most important cities in southern Italy. Analyses were carried out using Landsat images acquired in 1976, 1999 and 2009. The obtained results showed a significant urban expansion and an increase of irregularity degree in the fabric of the city. Such a variation is related to economic factors, industrial expansion and population growth.

2 Materials and Methods

2.1 Change Detection

Over the years, different techniques and algorithms were developed for change detection from the simplest approach based on (i) a visual interpretation and/or manual digitization of change [11] [15] to the computation and filtering such as (ii) image algebra change detection [9], image regression, image rationing [10] and vegetation index differencing [20] [11] [15].

The effectiveness of change detection algorithms is strongly dependent on surface characteristics of the study area, on spectral and spatial resolution of available historical data sets, and on decision makers needs. All these critical aspects make it difficult to develop a general methods effective and reliable for all applications in different regions.

This study deals with the spatial characterization of expansion of urban areas in southern Italy, by using geospatial analysis applied to multi-date Thematic Mapper (TM) satellite images.

Over the years, satellite time series data sets, such as Landsat MSS and TM images have been used to assess urban growth, mainly for big cities [16] [26] [27]. The investigation herein presented focused on assessment of the expansion of several very small towns very close to Bari (one of the biggest cities in southern Italy). To cope with the fact that small changes have to be captured and extracted from TM multi-temporal data sets, we adopted the use of spectral indices to emphasize occurring changes, and geospatial data analysis for revealing spatial patterns.

Analyses have been carried out using global and local spatial autocorrelation applied to multi-date NASA Landsat images acquired in 1976, 1999 and 2009. The results we obtained show a significant urban expansion coupled with an increase of the irregularity degree of urban pattern in 1976, 1999 and 2009. This variation is also connected with urban expansion and population growth.

Since 1972, the Landsat satellites have provided repetitive, synoptic, global coverage of high-resolution multispectral imagery. The characteristics of TM bands were selected to maximize each band's capabilities for detecting and monitoring different types of land surface cover characteristics.

LANDSAT TM multispectral data have been acquired from a nominal altitude of 705 kilometers (438 miles) in a near-circular, sun-synchronous orbit at an inclination of 98.2 degrees, imaging the same 185-km (115-mile) swath of Earth's surface, every 16 days. All TM spectral bands (1 to 5 and 7) are listed in Table 1. All of remote sensed data have been georeferenced according to UTM projection.

Table 1. TM spectral bands

Thematic Mapper (TM)		
Landsat 4-5	Wavelength (micrometers)	Resolution (meters)
Band 1	0.45-0.52	30
Band 2	0.52-0.60	30
Band 3	0.63-0.69	30
Band 4	0.76-0.90	30
Band 5	1.55-1.75	30
Band 6	10.40-12.50	120
Band 7	2.08-2.35	30

The availability of a long time series of TM data systematically acquired, stored and now free available from NASA website for the whole globe makes the TM time series an invaluable data source for change detection. Moreover, geometric stability and high positional accuracy of TM data enable a reliable co-registration of multiple images, whereas radiometric consistency allows us to adjust scenes to spectrally match. Such characteristics make TM data valuable and reliable low cost technologies useful not only for assessing large-scale changes, such as land-use and land-cover, but also for assessing variations occurring at smaller scales, such as urban expansion with new houses and roads.

Satellite images acquired in different years (1976, 1999 e 2009) have been used in this work. Table 2 shows a comparison between Landsat Terra Aster sensors.

Table 2. Comparison among several Landsat and Terra Aster sensors

	Landsat MSS	Landsat TM	Landsat ETM+	Terra ASTER
Resolution	80 m	30 m	30 m	15 m (VNIR)
Green	1 (0,5-0,6 μm)	2 (0,52-0,6 μm)	2 (0,52-0,6 μm)	1 (0,52-0,6 μm)
Red	2 (0,6-0,7 μm)	3 (0,63-0,69 μm)	3 (0,63-0,69 μm)	2 (0,63-0,69 μm)
Near infrared	3 + 4 (0,7-1,1 μm)	4 (0,76-0,9 μm)	4 (0,76-0,9 μm)	4 (0,78-0,86 μm)

2.1.1 Spectral Band Analysis

Observing satellite spectral bands in RGB (Red, Green, Blue), the growth of a city is characterized by the transition from natural vegetation colours to lighter and brighter colours, generally due to high reflection of buildings and soils where vegetation has been removed.

In order to obtain an image in RGB, it is necessary to combine 3, 2, 1 bands. Using GRASS open source GIS software (www.grass.itc.it), it is possible to adequately combine the bands, through the module called `i.landsat.rgb`. This module performs a self-balancing action and increases the colour channels of a Landsat RGB image, obtaining a mixture of more natural colours. Original data remain intact and only

colour table of each band is changed. The module operates computing a histogram for each colour channel and removing a controlled amount of outliers before colour scale recalibration with an appropriate module (r.colors). The i.landsat.rgb tool works with any set of RGB images and the script can be easily modified to work with other datasets and bands. In the present paper three TM images, acquired in 1976, 1999 and 2009, have been used. Two of TM images are shown in Figure 1 (1976 and 2009, respectively) using RGB composition to emphasize areas of concern; light spots are related to urban areas.

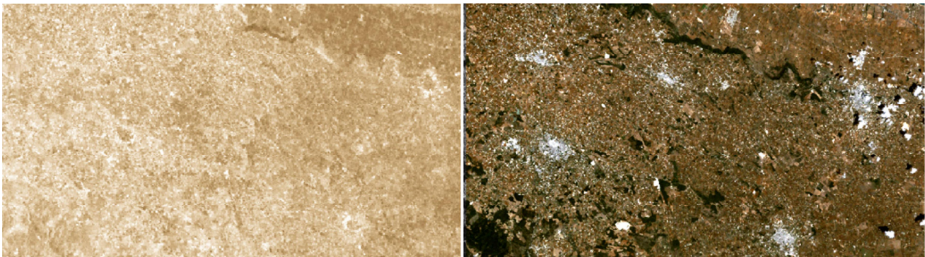


Fig. 1. Comparison of RGB Landsat images at 1976 e 2009 south of Bari municipality

The phenomenon of urban growth can be analyzed at different scales depending on the context of settlements to be examined. These slow dynamics can be analyzed with images of the same area at different times, with a not particularly high spatial resolution. In any case, urban areas can be identified by investigating spectral responses both in visible and infrared bands. The figure below highlights that buildings have higher values in the infrared than in the visible band.

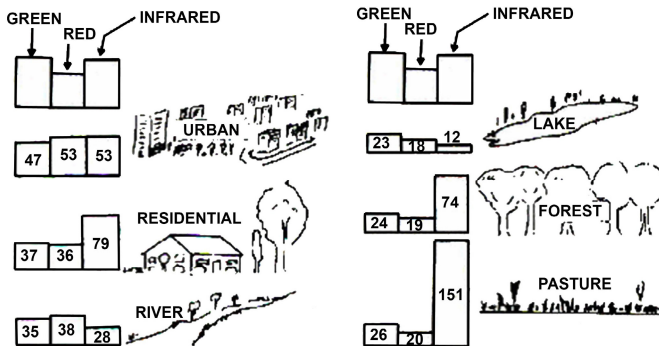


Fig. 2. Signal intensity in visible and infrared

A possible way to highlight changes, considering only three images at a time, is the composition in false colour. Another possibility is to combine bands in a different way. As for multispectral images, scenes with different acquisition dates, rather than bands with different spectral range, are combined. A colour image should be obtained

where unchanged areas appear in gray scale, while zones where changes have occurred show brighter colours. In the following image (Fig. 3) the three 2 bands at 1976, 1999, 2009 have been combined.

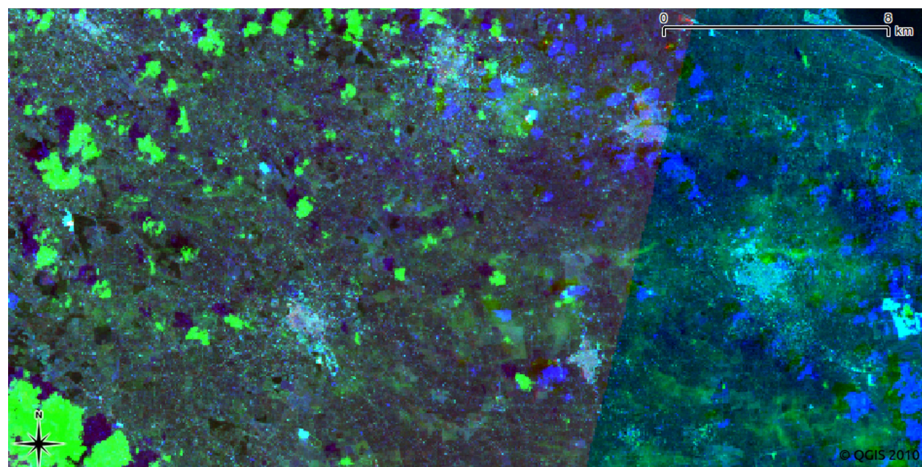


Fig. 3. False Colour composition

The most evident colours show variations occurred from 1976 to nowadays. Unfortunately, in the image there are clouds indicating pixel erroneous changes; but in other cases (indicated by light colour), the increase of urbanized area over the years is evident.

2.2 Spatial Autocorrelation Techniques

In addition to RGB images composition and band classification at different periods, in this study Landsat images acquired in 1999 and 2009 have been examined using spatial autocorrelation techniques.

The concept of spatial autocorrelation is rooted on Waldo Tobler [24] first law of geography: “*everything is related to everything else, but near things are more related than distant things*”. Spatial autocorrelation can be considered positive if similar values of a variable tend to produce clusters; in the same way spatial autocorrelation can be classified as negative when similar values of a variable tend to be scattered throughout the space (Boots and Getis, 1988).

Spatial autocorrelation takes into account the spatial attributes of geographical objects under investigation, evaluates and describes their relationship and spatial patterns, also including the possibility to infer such patterns at different times for the study area. The spatial patterns are defined by the arrangement of individual entities in space and by spatial relationships among them. Spatial autocorrelations measure the extent to which the occurrence of one object/feature is influenced by similar objects/features in adjacent areas. As such, statistics of spatial autocorrelation provide

(i) indicators of spatial patterns and (ii) key information for understanding spatial processes underlying the distribution of an object/feature and/or a given phenomenon under observation. Geographical observations can be arranged in spatial and temporal order, by latitude and longitude, and over given time periods. In this context time series data, such as aerial and satellite images, can provide useful data sets to examine changes in homogeneity over time, as well as to measure the strength of the relationship between values of the same variables over a given time window. Spatial autocorrelation statistics are considered very useful tools in analysing satellite images, since they consider not only pixel value (reflectance, temperature, spectral index) under investigation, but also the relationship between that same pixel and its surrounding pixels in a given window size.

In absence of spatial autocorrelation the complete spatial randomness hypothesis is valid: the probability to have an event in one point with defined (x, y) coordinates is independent of the probability to have another event belonging to the same variable. The presence of spatial autocorrelation modifies that probability. Fixed a neighbourhood for each event, it is possible to understand how much it is modified by the presence of other elements inside that neighbourhood. The presence of autocorrelation in a spatial distribution is caused by two effects, that could be clearly defined, but not separately studied in the practice:

(i) first order effects: they depend on region of study properties and measure how the expected value (mean of the quantitative value associated to each spatial event) varies in the space by equation 1:

$$\hat{\lambda}_T(s) = \lim_{ds \rightarrow 0} \left\{ \frac{E(Y(ds))}{ds} \right\} \tag{1}$$

where ds is the neighbourhood around s, E() is the expected mean and Y(ds) is events number in the neighbourhood;

(ii) second order effects: they express local interactions between events in a fixed neighbourhood, that tends to the distance between events i and j. These effects are measured with covariance variations expressed by the limit in formula 2:

$$\gamma^{(s_i, s_j)} = \lim_{ds_i, ds_j \rightarrow 0} \left\{ \frac{E(Y(ds_i)Y(ds_j))}{ds_i ds_j} \right\} \tag{2}$$

The characterization of spatial autocorrelation requires detailed knowledge on:

(a) the quantitative nature of dataset, also called intensity of the spatial process, that is how strong a variable happens in the space [5] [19], with the aim to understand if events are similar or dissimilar;

(b) the geometric nature of dataset: this needs the conceptualization of spatial relationships, usually done with the use of matrixes: (i) a distance matrix is defined to consider at which distance events influence each other (distance band); (ii) a contiguity matrix is useful to know if events influence each other; (iii) a matrix of spatial weights expresses how strong this influence is.

Concerning distance matrix, a method should be established to calculate distances in module and direction. For this concern the module, namely Euclidean distance (3), is the most adopted.

$$d_E(s_i, s_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (3)$$

2.2.1 Global Indicators of Spatial Association

Several indexes have been developed in order to measure spatial autocorrelation discovering the presence and intensity of clusters in the distribution. The two main indicators are *Moran I* [17] and *Geary C Ratio* [7] indexes.

Moran I index is defined by the following equation:

$$I = \frac{N \sum_i \sum_j w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j w_{ij}) \sum_i (X_i - \bar{X})^2} \quad (4)$$

where:

- N is the number of events;
- X_i and X_j are intensity values at points i and j (with $i \neq j$), respectively;
- \bar{X} is the average of variables;
- $\sum_i \sum_j w_{ij} (X_i - \bar{X})(X_j - \bar{X})$ is the covariance multiplied by an element of weight matrix. If X_i and X_j are both upper or lower than the mean, this term will be positive, if the two terms are in opposite positions compared to the mean the product will be negative;
- w_{ij} is an element of weight matrix which depends on contiguity of events. This matrix is strictly connected to adjacency matrix.

Moran index shows a trend similar to the correlation coefficient, consequently it can have values included between -1 and 1.

Geary C Ratio is quite similar to *Moran I index* and it is defined by the following equation:

$$c = \frac{(N-1)(\sum_i \sum_j w_{ij} (X_i - X_j)^2)}{2(\sum_i \sum_j w_{ij}) \sum_i (X_i - \bar{X})^2} \quad (5)$$

Parameters are very similar to equation 4: the main difference is represented by the cross-product term in the numerator, which in *Moran* is calculated using deviations from the mean, while in *Geary* is directly computed. The square root provides to remove all negative values of the formula, consequently *Geary C Ratio* ranges between 0 and 2. Values between 0 and 1 define positive autocorrelation, while values greater than 1 and smaller than 2 indicate negative autocorrelation. Value 0 represents a perfect positive autocorrelation, the same of neighbouring values with cross-product equal to 0. Value 2 defines a perfect negative spatial autocorrelation.

2.2.2 Local Indicators of Spatial Association

Luc Anselin [1] defines as a local indicator of spatial association, any statistic that satisfies the following two requirements:

- the index for a single observation produces a spatial result of the extent of clustering of similar values around that observation;
- the sum of all observations indexes is proportional to the global indicator of spatial association.

Local versions of spatial autocorrelation are used to measure the magnitude of spatial autocorrelation within the immediate neighbourhood. Values indicating the magnitude of spatial association can be derived for each areal unit and they can be located. The local version of statistics employs distance information to identify local clusters and relies on the distance information captured in Distance matrix.

The *Local Indicator of Spatial Association (LISA)* [1] represents the local version of *Moran index I* and it is defined by the relation:

$$I_i = \frac{(X_i - \bar{X})}{S_X^2} \sum_{j=1}^N (w_{ij} (X_j - \bar{X})) \tag{6}$$

where \bar{X} is the intensity mean of all events, X_i is the intensity of event “i”, X_j is the intensity of event “j” (with $j \neq i$), S_X^2 is the variance of all events and w_{ij} is the weight matrix. Considering z-score:

$$z_i = \frac{(X_i - \bar{X})}{S_X}$$

LISA index can be expressed in the following synthetic form:

$$I_i = z_i \sum_{j=1}^N w_{ij} z_j \tag{7}$$

The function by *Getis & Ord* [8] is represented by the following equation:

$$G_i(d) = \frac{\sum_{i=1}^n w_i(d) x_i - x_i \sum_{i=1}^n w_i(d)}{S(i) \sqrt{\left[(N-1) \sum_{i=1}^n w_i(d) - \left(\sum_{i=1}^n w_i(d) \right)^2 \right] / N - 2}} \tag{8}$$

which is very similar to Moran index, except for $w_{ij}(d)$ which, in this case, represents a weight which varies according to distance. These statistics allow us to locate clustered pixels, by measuring how much features inside a fixed neighbourhood are homogeneous. Nevertheless, the interpretation of Getis and Ord’s G_i meaning is not immediate, but it needs a preliminary classification that should be done comparing G_i with intensity values.

The local version of *Geary Ratio C* is defined as:

$$C_i = \sum_{j=1}^N w_{ij} (z_i - z_j)^2 \tag{9}$$

Local indicators of spatial association can be considered as local functions of statistical analysis and can be represented through georeferenced maps, constituting

very important tools for exploratory analysis of spatial structures especially with large databases.

3 The Case Study

This study deals with satellite based investigations on urban area expansion in some test areas of southern Italy, using change detection techniques and spatial statistics to capture and characterize the spatial characterization of feature variations.

The investigation herein presented was focused on the assessment of the expansion of several very small towns very close to Bari (in southern Italy), the second largest city of Southern Italy, located in Apulia (or Puglia) Region. It faces the Adriatic Sea and has one of the major seaports in Italy. Bari is the fifth largest province (more than 5,000 square kilometres) in Italy and also the most populated with around 1,600,000 inhabitants in 2007. The city has around 400,000 inhabitants. The area of concern is characterized by an active and dynamic local economy, mainly based on small and medium enterprises operative in commerce, industry and services.

Bari has become one of the top commercial and industrial leaders in Italy, so it is known as 'California of South', to indicate the significant growth and leadership much higher than other southern areas. Industrial activities are quite numerous and dynamic (chemicals, machinery, printed materials, petroleum and textiles production), but also agriculture is quite notable in Bari surroundings, with intensive production of cherries, tomatoes, artichokes, grapes and table wine.

Bari has also a long history since the Middle-Ages, when it was one of the main ports from which pilgrims sailed to the Holy Land.

3.1 Change Detection

The main aim of our investigation was to evaluate the possibility to enhance spatial patterns of urban development of years 1999 and 2009 in the area of concern. The expansion of urban areas has been assessed by using change detection techniques along with both global and local geospatial statistical analysis.

Change detection is the assessment of variations between multirate, or time series data sets, or, in the case of remotely sensed data, between two or more scenes covering the same geographic area and acquired in different periods.

To cope with the fact that small changes have to be captured and extracted from TM multitemporal data sets, it is important that an adequate processing chain must be implemented. Indeed, multirate imagery data analysis requires a more accurate pre-processing than single date analysis. This includes calibration to radiance or at-satellite reflectance, inter-calibration among multirate images, atmospheric correction or normalization, image registration, geometric correction, and masking (e.g., for clouds, water, irrelevant features). These procedures improve the capability in discriminating real changes from artefacts introduced by differences in sensor calibration, atmosphere, and/or sun angle. Some radiometric rectification techniques are based on the use of areas of the scene under investigation whose reflectance is nearly constant over time.

The images under investigations were pre-processed, co-registered and inter-calibrated to reduce sources of false changes, such as those caused by clouds, cloud shadows, and atmospheric differences.

Relating to change detection, we should consider that up to now, a number of change detection techniques have been devised and applied for capturing variations of surface characteristics, atmospheric components, water quality and coastal zones. Some methods focused on the monitoring of urbanization, agricultural development, forest land management, and environmental management.

These procedures generally are coupled with data transformation to vegetation indices, whose principal advantage over single-band radiometrics is their ability to strongly reduce data volume for processing and analysis, and also to reduce residual of atmospheric contamination. In our analyses, we adopted Normalized Difference of Index Vegetation (NDVI), which is the most widely used index for a number of different applications, ranging from vegetation monitoring to urban sprawl. The NDVI is computed using the following formula:

$$\text{NDVI} = \frac{R_{\text{NIR}} - R_{\text{RED}}}{R_{\text{NIR}} + R_{\text{RED}}} \quad (10)$$

This index was computed for both 1999 and 2009, to emphasize occurring changes and improve change detection analysis carried out as classification comparison from NDVI processed using geospatial data analysis.

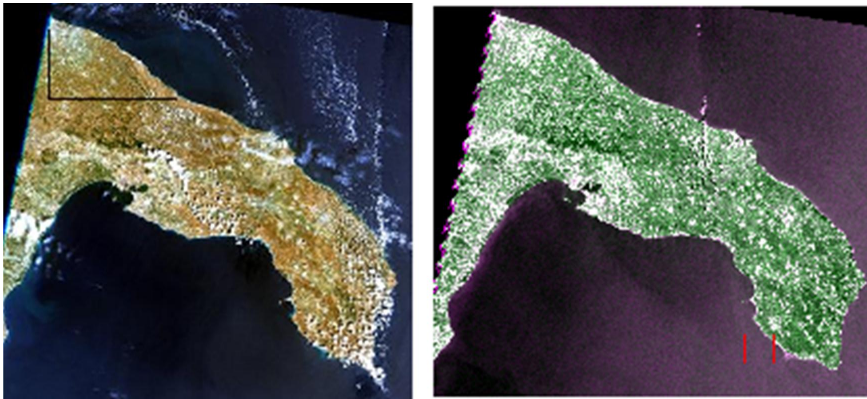


Fig. 4. RGB of TM images acquired in 1999 and 2009, note that light spots are urban areas. The black rectangle indicates the location of the study area.

Figure 4 show NDVI maps computed from TM images acquired in 1999 and 2009, respectively. A visual comparison between the figures clearly points out that the use of spectral combinations of red and NIR bands highlights light spots related to urban areas. In particular, the comparison between multidecade (1999 and 2009) NDVI maps emphasises the expansion of urban areas, which are easily recognizable by a visual inspection.

The following image shows the numerical difference between 1999 and 2009 maps. The increase in the extension of urban area was connected to economic and demographic factors.

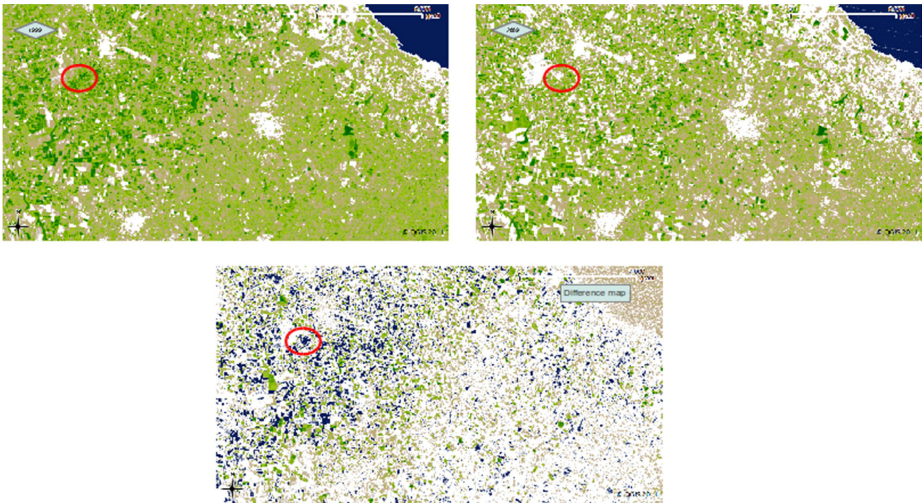


Fig. 5. NDVI map from the TM images acquired in 1999 and 2009, note that light spots are urban areas. NDVI difference map from the TM images acquired in 1999 and 2009, note that white pixels are urban areas. The red circle indicates a strong change in NDVI index where vegetation has been replaced by the urbanized area.

3.2 Spatial Autocorrelation

To study spatial autocorrelation in satellite data, it is important to define which are the spatial events, their quantitative nature (intensity) and the conceptualization of geometric relationships. A spatial event is clearly the pixel. Spatial autocorrelation statistics are usually calculated considering geographical coordinates of its centroid. Concerning the intensity, it should be chosen strictly considering the empirical nature of the case study.

In image processing, Global measures of spatial autocorrelation provide a single value that indicates the level of spatial autocorrelation within the variable distribution, namely the homogeneity of a given value within the image under investigation.

Local measures of spatial autocorrelation provide a value for each location within the variable distribution and, therefore, are able to identify discrete spatial patterns that may not otherwise be apparent [23]. Statistics output is an image for each calculated index, which contains a measure of autocorrelation around that pixel.

Both global and local statistics can be calculated using spectral channels, spectral combinations and/or multi-temporal combinations as intensity.

In order to identify areas of urban expansion, we looked for a change in spatial structure between two image dates.

Spatial dependency may be captured using spatial autocorrelation statistics such as join-counts, Moran's I and Geary's c . Therefore, we consider that the temporal change of geospatial statistic between image dates provides information on change in spatial structure of some unspecified nature between two image dates.

For a given pixel, the change from one date to another will be on account of changes in the spatial structure within the range of spatial windows of that pixel. Spatial differences, which are equal between the two dates for a given co-registered pixel window, will not induce a change.

However, results from these analyses may be unrepresentative if the nature and extent of spatial autocorrelation varies significantly over the area of interest. To cope with this issue, we considered: (i) local indicators of spatial association and (ii) the hypothesis that a region with urban settlements will exhibit spatial homogeneity in spectral response, due to the lowly variable spatial and spectral structure of concrete and building materials.

4 Results

In the current study, both global and local geospatial statistics were applied to 1999 and 2009 TM images, using spectral combinations of single bands to enhance variations occurring during the time window under investigation. The comparison was made using single date NDVI maps computed for both 1999 and 2009 along with the map obtained as difference between NDVI 1999 and 2009. Later on, the multivariate data set was analyzed using a pixel-by-pixel comparison followed by change region analysis and verification of results from the two successive temporal scenes (1999 and 2009). Figure 6 (from left to right) shows local autocorrelation indexes presented as RGB Getis G_i^* , Local Geary c and LISA applied with lag 2. All panels clearly reveal the increase in urbanized area; RGB Getis, Geary's c and Moran of the map well show variations linked to concrete and building materials.

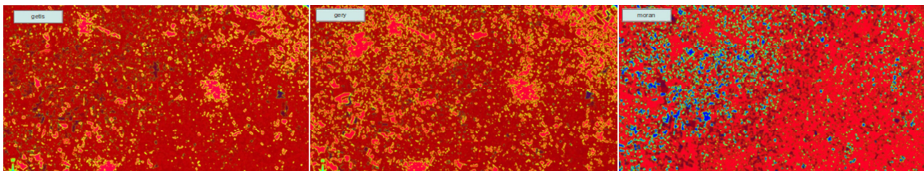


Fig. 6. (from left to right) show the local autocorrelation indexes presented as Getis G_i^* , Local Geary c and LISA applied with lag 2

Taking into account the obtained results, we can observe that the distribution of the built-up area was more homogeneous and less fragmented in 1999, without the presence of different urban centres. During the period up to 2009 changes led to an increase of density leading to an increase in urban areas expansion.

Another procedure in analyzing the evolution of urbanized areas is given by the combination of bands. In this way, more appropriate indices of remote sensing for the study of a given phenomenon can be built. One of the indices used for the study of urban phenomena is BAI Built-up Areas Index = $(blue - ir)/(ir + blue)$. The BAI is a very useful index for identifying impermeable surfaces like asphalt and concrete. Values generated using BAI index range from -1 to 1 and this also emerges from basic statistics (module *r.stats* GRASS) performed on raster data.

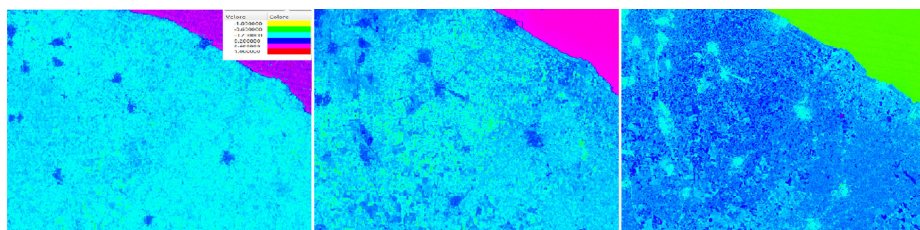


Fig. 7. Built-up Areas Index using Landsat images (1976-1999-2009)

5 Final Remarks

In the present paper, each step of the process has been carried out using free tools and data. Operating system (Linux Ubuntu) and GIS software (GRASS GIS and Quantum GIS) are open source type, while Landsat data are downloadable and ready to use. This aspect is very important since it puts no limit and allows everybody to carry spatial analyses on remote sensing data.

As regards autocorrelation analysis, it was considered as a method for examining transformations taking place in urbanized areas located in southern Italy. The main objectives of the study were: (i) to assess if the variation of urban structure over time can be quantitatively determined using TM images, (ii) to investigate and describe the modification of urban shape and morphology over time.

Analysing and comparing different years, the process of urban intensification has been observed, and the increase of urbanized area was revealed. This change shows the transformation that took place in the area under investigation and the transformation from quite regular to more fragmented peripheral settlements. The relevance of the technique herein used is that it provides a reliable way of analysing the urban structure and its transformation through time.

However, this study is preliminary and quite suggestive and its main objective was to present a way of applying autocorrelation analysis to the monitoring of urban area evolution. The need of analysing more time periods and a comparative analysis among many urban areas would be fruitful, and the application of the geostatistical analysis applied to satellite time series constitutes a major challenge for further investigation.

References

1. Anselin, L.: Local indicators of spatial association – LISA. *Geographical Analysis* 27, 93–115 (1995)
2. Benguigui, B., Chamanski, D., Marinov, M.: When and where is a city fractal? *Environ. Planning B* 27, 507–519 (2000)
3. Boots, B.N., Getis, A.: *Point Pattern Analysis*. Sage Publications, Newbury Park (1988)
4. Cheng, J., Masser, I.: *Modelling urban growth patterns: a multiscale perspective* (2002)

5. Danese, M., Lazzari, M., Murgante, B.: Kernel Density Estimation Methods for a Geostatistical Approach in Seismic Risk Analysis: The Case Study of Potenza Hilltop Town (Southern Italy). In: Gervasi, O., Murgante, B., Laganà, A., Taniar, D., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2008, Part I. LNCS, vol. 5072, pp. 415–429. Springer, Heidelberg (2008), doi:10.1007/978-3-540-69839-5_31.
6. Frankhauser, P.: The Fractal Approach, a new tool for the spatial analysis of urban agglomerations, *Population: An English Selection. New Methodological Approaches in the Social Sciences* 10(1), 205–240 (1998)
7. Geary, R.: The contiguity ratio and statistical mapping. *The Incorporated Statistician* (5) (1954)
8. Getis, A., Ord, J.: The analysis of spatial association by distance statistics. *Geographical Analysis* 24, 189–206 (1992)
9. Green, K., Kempka, D., Lackey, L.: Using remote sensing to detect and monitor land cover and land use. *Photogrammetric Engineering and Remote Sensing* 60, 331–337 (1994)
10. Howarth, J.P., Wickware, G.M.: Procedure for change detection using Landsat digital data. *International Journal of Remote Sensing* 2, 277–291 (1981)
11. Lacy, R.: South Carolina finds economical way to update digital road data. *GIS World* 5(10), 58–60 (1992)
12. Lambin, E.F.: Change detection at multiple scales seasonal and annual variations in landscape variables. *Photogrammetric Engineering and Remote Sensing* 62, 931–938 (1996)
13. Lanorte, A., Danese, M., Lasaponara, R., Murgante, B.: Multiscale mapping of burn area and severity using multisensor satellite data and spatial autocorrelation analysis. *International Journal of Applied Earth Observation and Geoinformation* (2012), doi:10.1016/j.jag.2011.09.005
14. Lichtenegger, J.: ERS-I: land use mapping and crop monitoring: a first close look to SAR data. *Earth Observation Quarterly*, 37–38 (May-June 1992)
15. Light, D.: The national aerial photography program as a geographic information system resource. *Photogrammetric Engineering and Remote Sensing* 59, 61–65 (1993)
16. Masek, J.G., Lindsay, F.E., Goward, S.N.: Dynamics of urban growth in the Washington DC metropolitan area, 1973-1996, from Landsat observations. *Int. J. Rem. Sensing* 21, 3472–3486 (2000)
17. Moran, P.: The interpretation of statistical maps. *Journal of the Royal Statistical Society* (10) (1948)
18. Muchoney, D.M., Haack, B.N.: Change detection for monitoring forest defoliation. *Photogrammetric Engineering and Remote Sensing* 60, 1243–1314 (1994)
19. Murgante, B., Danese, M.: “Urban versus Rural: the decrease of agricultural areas and the development of urban zones analyzed with spatial statistics” Special Issue on Environmental and agricultural data processing for water and territory management. *International Journal of Agricultural and Environmental Information Systems (JAEIS)* 2(2), 16–28 (2011) ISSN 1947-3192, doi:10.4018/jaeis.2011070102
20. Nelson, R.F.: Detecting forest canopy change due to insect activity using land sat MSS. *Photogrammetric Engineering and Remote Sensing* 49, 1303–1314 (1983)
21. Sailer, C.T., Eason, E.L.E., Brickey, J.L.: Operational multispectral information extraction: the DLPO image interpretation program. *Photogrammetric Engineering and Remote Sensing* 63, 129–136 (1997)
22. Shen, G.: Fractal dimension and fractal growth of urbanized areas. *Int. J. Geogr. Inf. Sci.* 16, 419–437 (2002)

23. Tateishi, R., Kajiwara, K.: Global Lands Cover Monitoring by NOAA NDVI Data. In: Proceeding of International Workshop of Environmental Monitoring from Space, Taejon, Korea, pp. 37–48 (1991)
24. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2), 234–240 (1970)
25. United Nations Population Division (2001) World Population Monitoring 2001: Population, environment and development (2001), <http://www.un.org/esa/population/publications/wpm/wpm2001.pdf>
26. Yang, X., Lo, C.P.: Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. *Int. J. Rem. Sensing* 23, 1775–1798 (2002)
27. Yuan, F., Sawaya, K., Loeffelholz, B.C., Bauer, M.E.: Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing. *Rem. Sensing Environ.* 98, 317–328 (2005)

A Framework for QoS Based Dynamic Web Services Composition

Jigyasu Nema¹, Rajdeep Niyogi¹, and Alfredo Milani²

¹ Department of Electronics & Computer Engineering,
Indian Institute of Technology, Roorkee-247667, India
jigyasunema@gmail.com, rajdpfec@iitr.ernet.in

² Department of Mathematics and Computer Science
University of Perugia, 062123 Perugia, Italy
milani@unipg.it

Abstract. Web services composition based on quality of services have been recently studied by several researchers. Service composition with minimum communication cost has been well studied for service overlay networks. However there are very few optimization algorithms available which considers more than one objective. Finding a solution that optimizes all the objectives is nontrivial. In multi-objective optimization problem, the solution can be given in such a way that it fulfills the requirements of the user without making much sacrifice from the optimal solution. In this paper we would like to address the problem of dynamic services composition in an existing framework by considering two QoS parameters that are communication cost and service cost. For this we suggest an algorithm. Some experimental results are also reported that validate the algorithm.

Keywords: Dynamic web services composition, quality of services, communication cost, service cost.

1 Introduction

Web services are a new type of Web applications. These applications can be published, located, and invoked across the Web. Web services perform functions that can be anything from simple requests for information to creating and executing complicated business processes. Once a Web service is deployed, it can be discovered and invoked by other applications (or other Web services).

A web services framework consists of three parts — communication protocols, service descriptions, and service discovery [4]. These are: (i) the simple object access protocol (SOAP) which enables communication among Web services, (ii) the web services description language (WSDL), which provides a formal, computer-readable description of Web services and (iii) the Universal Description, Discovery, and Integration (UDDI) directory, which is a registry of Web services descriptions.

In service computing, *service composition* is the process of assembling independent and reusable service components across different domains in the Internet to construct richer application functionality according to a *service composition request*. Generally, a service composition request is associated with a set of required

services and invocation orders between them. Such a relationship among services can be characterized by a directed acyclic graph (DAG) [1].

In the Internet, the websites that provide services are called service providers. The service composition solution of a service composition request is to assign each required service in the service composition request graph to an appropriate service provider [1]. Quality of the composition can be measured on some parameters like communication cost, service cost, reliability and availability [2].

Web services composition can be done in two ways on the basis of when the services are selected for composition. In static composition the services are selected at compile time. In dynamic composition service selection is done at run time. Dynamic service selection is much complex than static service selection. In dynamic web service selection the service request is broken in smaller sub-problem, then for those sub-problems services are selected then those services are composed together in such a manner that the complex task can be fulfilled. Benefits of the dynamic web services are on-demand service, unlimited number of new services [5]. In Static Service composition when any of the service required for composition is not available then the task will not be performed. But, in case of Dynamic web service composition if such a situation arises, the situation would be handled by selection another service providing the same functionality.

If there is only one service provider available for every service there is no need of any optimization technique. Typically a service is provided by more than one service provider. The objective is to select a set of providers such that a combination of one or more costs associated with a service is optimized. Such costs are characterized by QoS parameters. Some commonly used parameters include communication cost, service cost, reliability, and availability [2], [7-8],[14]. However we need to consider situations where the providers contributing to an optimal solution may not be available at some time. The non-availability of the providers may arise due to either failure of the provider or breakdown of links connecting the providers. Thus we would now look for a solution that is as close as possible to the optimal solution. In this paper we suggest a method for obtaining such suboptimal solutions when only two QoS parameters (namely, communication cost and service cost) are considered.

The remaining paper is structured as follows. In Section 2 we discuss some related work. In Section 3 we propose a new algorithm for two QoS parameters. In Section 4 we present our experimental results and provide an analysis of the results. Section 5 concludes the paper.

2 Related Work

Several approaches have been proposed for dynamic web service composition [5-6],[9]. A classification of dynamic web services composition techniques is presented in [5]. Some approaches include: (i) ontology based approach where every service description is marked with ontology; the composition is done on the basis of semantic similarity, (ii) runtime component adaptation, (iii) runtime reconfiguration using wrappers, and (iv) declarative composition. A workflow approach is proposed in [6]. A semantic graph-based composition algorithm is suggested in [9] where the composition is performed by matching the input-output parameters of web services.

Some works describe frameworks for dynamic web service composition. In [3] a web service composition system is presented which specifies the workflow model for

the composition of the services. In [11] model driven approach is proposed that enables the system of dynamic selection of services. A software system called FUSION has been developed in [10] for service portals---a collection of services made available to the user. The authors [10] feel that the architecture would be commercially useful for portal designers. In [15] for a given request a dependency graph is generated dynamically. Now on this graph an A* based search algorithm is applied to obtain the composed service. However all these frameworks are designed for only one QoS parameter.

In dynamic web service composition, QoS parameters are also taken into consideration during composition. In [12-14] QoS aware dynamic web service composition is discussed. In [7], [13] it is suggested that solving the composition problem of multi-objective function (involving QoS parameters) can be converted into a single objective problem. The difficulty in these approaches lies in the normalization of the QoS parameters. Genetic algorithm based multi-objective evolutionary algorithms have been suggested in [8] for web service composition. The algorithms used are well known Genetic Algorithms. These algorithms are SPEA2 and NSGA2. This paper considers only sequences of services. It is not capable of handling complex types of request such as those represented using DAGs.

3 A Randomized Algorithm for Multi-objective Optimization for Service Composition

We suggest an algorithm **Rand-Multiobj-Service-Comp** that tries to optimize a linear combination of two objectives, namely communication cost (CC) and service cost (SC). The inputs to this algorithm are (i) a service overlay network that consists of M service providers, N services, service cost matrix, communication cost matrix, and (ii) a service request (SR) (represented in the form of a DAG). The minimum communication cost (CC_{min}) and the minimum service cost (SC_{min}) for the given SR are computed using the algorithms **ServiceCost** and **CCmin** given in Section 3.1 and 3.2 respectively.

Our algorithm works as follows. It randomly picks a solution, say s, for the given service request. Then it computes the costs for s using the respective cost matrices. Let these costs be X and Y. Next it compares the sum of these costs (X+Y) with the sum of the optimal costs (X_{min} + Y_{min}). For some chosen threshold values it decides whether or not to accept the solution s. This process is repeated until a satisfactory solution is obtained.

Pseudo-code for Rand-Multiobj-Service-Comp:

1. $X_{min} := \text{ServiceCost}(\text{ServiceRequest}[n])$
2. $Y_{min} := \text{CCmin-algo}(\text{SR});$ // given in section nnn [1]
3. For each service j in SR, given as input in the form of a DAG,
4. Pick a random provider for service j from $\text{ProvidersOf}[n][M]$, say k.
5. Append (j, k) to the front of a list L
6. $X := X + \text{costofservice}(j,k)$ // obtained from the given matrix
7. endfor
8. Calculate the communication cost from the list L, say Y, using the communication cost matrix.
9. Choose β and β' suitably.

10. **If** $X - X_{\min} \leq \beta$ and $Y - Y_{\min} \leq \beta'$
 (β is the percentage of deviation from SC_{\min} , β' is the percentage of deviation from CC_{\min} .)
11. **then** $Z := X+Y$;
12. Solution is found, exit.
13. GOTO step 3.

3.1 Web Service Composition with Minimum Communication Cost [1]

We give a brief outline of the method suggested in [1] to find a service composition solution with minimum communication cost for a given service composition request. The idea is to reduce the problem into smaller sub problems. The Auxiliary graph is designed from User request and Service Overlay Network. Then it tries to find out the basic structure of three nodes from the User request. When a basic structure is found, it updates the user request graph and also auxiliary graph.

Cost of the basic block is calculated using any minimum cost algorithm. The auxiliary graph is updated accordingly. This process is continued until the user request is reduced into one of the three basic structure [1]. Now the auxiliary graph contains only three nodes. For them the solution can be found in polynomial time. If it is not possible to reduce the user request graph in any of the basic structures, the program stops. The overall method can be represented as shown in the flowchart given in figure 1.

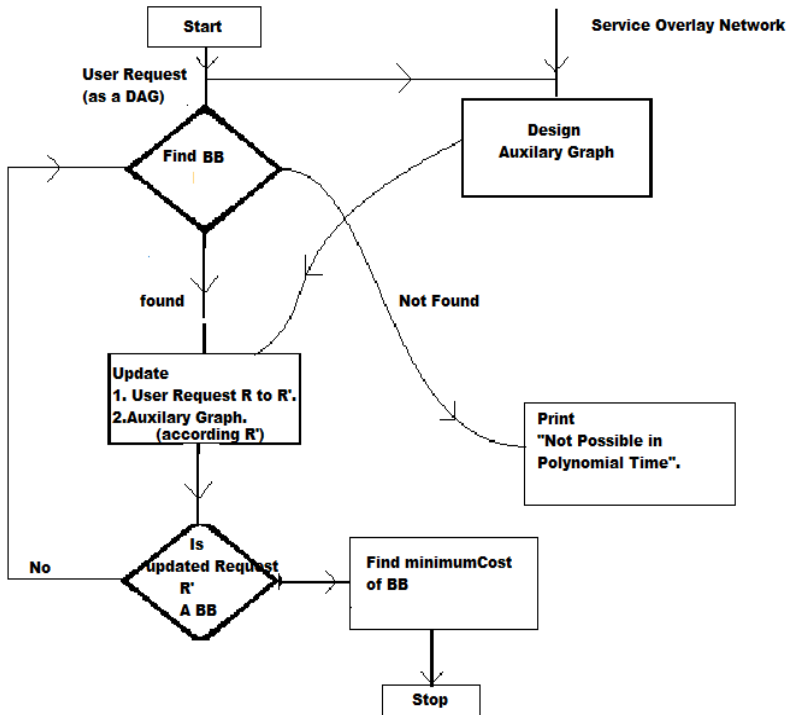


Fig. 1. General framework for web service composition with minimum cost

3.2 Web Service Composition with Minimum Service Cost

We suggest an algorithm that tries to optimize service cost for a given request(SC). The inputs to this algorithm are (i) a service overlay network that consists of M service providers, N services, service cost matrix, communication cost matrix, and (ii) a service request (SR) (represented in the form of a DAG).

This algorithm works as follows. The greedy approach is applied for finding the solution. Because choosing a provider for any service does not affect the selection of provider for another service. So, for getting the solution we take services one by one. Then for that service we look for the provider which is taking the least cost for that service. Total sum of the service costs of the provider chosen for that service is the optimum value for the given service request (SR).

Pseudo-code for the algorithm ServiceCost(ServiceRequest[n])

```

1. OverallCost := 0;
2. for i = 1 to n
3.     minCost := infinity; k := ServiceRequest[i]
4.     for providers j = 1 to M
5.         if (Provider[j][k] is ≠ 0)
6.             ProvidersOf[k][j] := 1
7.
8.             if (minCost > Provider[j][k])
9.                 { minCost := Provider[j][k];
10.                  SR-mincost-provider[i] := j; }
11.
12.             end if
13.         end if
14.     end for
15.     ProvidersOf[k][ SR-mincost-provider[i] ] := 0;
16.     OverallCost := OverallCost + minCost;
17. end for

```

4 Experiment Results

4.1 System Specification

We have performed the experiments shown below on a system with 2.26 GHz Intel Pentium, 4 GB of RAM and windows operating system.

4.2 Data Set

We have used the following data set for our experiments performed on a service overlay network. The overlay network consists of 20 service provider and 15 services. The information regarding the services provided by the service providers is stored in a matrix named **ProvidersOf**. The communication cost between two providers is stored in another matrix. The service cost associated with a service provider is stored in a matrix named **Provider**.

Table 2. Communication Cost matrix

	P 0	P 1	P 2	P 3	P 4	P 5	P 6	P 7	P 8	P 9	P 10	P 11	P 12	P 13	P 14	P 15	P 16	P 17	P 18	P 19
P0	0	4	4	3	3	3	3	4	4	2	2	4	4	3	3	3	3	4	4	1
P1	4	0	3	5	4	4	5	3	2	4	4	2	3	5	4	4	5	3	1	4
P2	4	3	0	5	4	4	5	2	3	4	4	3	2	5	4	4	5	1	3	4
P3	3	5	5	0	5	5	2	5	5	2	3	3	2	2	5	5	1	5	5	3
P4	3	4	4	5	0	2	5	4	4	3	3	4	4	5	2	1	5	4	4	3
P5	3	4	4	5	2	0	5	4	4	4	4	4	4	5	1	2	5	4	4	3
P6	3	5	5	2	5	5	0	3	5	3	3	5	5	1	5	5	2	5	5	3
P7	4	3	2	5	4	4	5	0	3	4	4	3	1	5	4	4	5	2	3	4
P8	4	2	3	5	4	4	5	3	0	3	3	1	3	5	4	4	5	3	2	4
P9	2	4	4	5	3	4	3	4	3	0	1	3	4	3	4	3	5	4	4	2
P10	2	4	4	5	3	4	3	4	3	1	0	3	4	3	4	3	5	4	4	2
P11	4	2	3	5	4	4	5	3	1	3	3	0	3	5	4	4	5	3	2	4
P12	4	3	2	5	4	4	5	1	3	4	4	3	0	5	4	4	5	2	3	4
P13	3	5	5	2	5	5	1	5	5	3	3	5	5	0	5	5	2	5	5	3
P14	3	4	4	5	2	1	5	4	4	4	4	4	4	5	0	2	5	4	4	3
P15	3	4	4	5	1	2	5	4	4	3	3	4	4	5	2	0	5	4	4	3
P16	3	5	5	1	5	5	2	5	5	5	5	5	5	2	5	5	0	5	5	3
P17	4	3	1	5	4	4	5	2	3	4	4	3	2	5	4	4	5	0	3	4
P18	4	1	3	5	4	4	5	3	2	4	4	2	3	5	4	4	5	3	0	4
P19	1	4	4	3	3	3	3	4	4	2	2	4	4	3	3	3	3	4	4	0

4.3 Results

We have implemented the above three algorithms. The service composition request is given in the form of DAGs. Some of the most common examples have been taken as input to the algorithms and the corresponding results are shown in Table 3. In the service request column different types of service composition is given. Corresponding to a service composition request, Solution (Service to Provider) column contains a solution obtained by the Randomized algorithm given in Section 3. The columns SC and CC correspond to the service cost and the communication cost respectively for the random solution. The column Z contains the sum of SC and CC. The column Z' contains the best possible value for that service composition request (SR). Solution for SCmin contains the service- provider pair for algorithm which provides the minimum service cost for the service composition request. Solution for CCmin contains the service-provider pair for algorithm which provides the minimum communication cost for the service composition request. From the experimental results obtained (shown in Table 3) we can conclude that for the same service composition request when individual objectives (CC and SC) optimization algorithms are applied the obtained results are conflicting.

Table 3. Experimental Results for Different Service Request

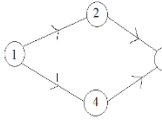
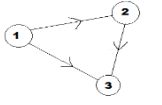
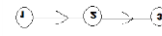
Service Request	Solution (Service to Provider)	P	SC	CC	Z	Z' (SCmin + CCmin)	Solution for SCmin	Solution for CCMIn
	S 1(3) 2(4) 3(5) 4(6)	P 3 3 5 5	20	10	30	24 (18+6)	19 3 2 5	19 19 5 5
	S 1(3) 2(6) 3(9)	P 3 6 9	22	7	29	24 (20+4)	19 5 9	3 6 6
	S 1(2) 2(4) 3(6)	P 2 3 3	20	5	25	20 (17+3)	1 3 5	19 3 3

Table 3. (continued)


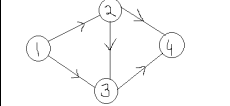
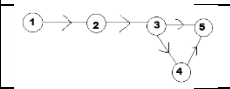
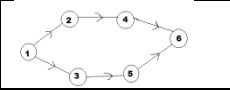
	S 1(2) 2(4) 3(6) 4(8)	P 2 4 3 7	24	23	47	33 (19+14)	1 3 5 5	1 1 5 8
	S 1(1) 2(2) 3(3) 4(4) 5(5)	P 14 1 2 4 4	24	15	39	28 (19+9)	14 1 19 3 2	15 19 19 4 4
	S 1(1) 2(3) 3(5) 4(7) 5(9) 6(11)	P 14 2 4 6 9 10	32	18	50	35 (21+14)	14 19 2 18 9 8	2 2 4 4 6 6

$(Z-Z')*100/Z'$ column contains the percentage deviation of sum of the costs from the best cost possible for the corresponding service request. The next two columns contain the percentage of deviation of individual objective value for the random solution from their best value possible. Time column contains the time taken for getting the appropriate random solution which is acceptable by the algorithm (this means that the result should not deviate more than a limit from the best solution).

Table 3. (continued)

Service Request	$(Z-Z')*100/Z'$	$(SC-SCmin)*100/SCmin$	$(CC-CCmin)*100/CCmin$	Time (seconds)
	25	11.11	66.67	0.640
	20.83	10	75	1.575

Table 3. (continued)

	25	17	66.67	1.342
	42.42	26.31	64.28	0.546
	39.28	26.31	66.67	0.936
	38.38	28	63.63	0.733

4.4 Analysis of the Result

Let us consider the Table 3 given above. For the service request given in row 1, we find that the service cost SC is 20 and the communication cost CC is 10. Thus the total cost Z is equal to 30. Now the corresponding minimum costs (SC_{min} and CC_{min}) are 18 and 6 respectively. Thus Z' is equal to 24. So the deviation is 25%. For the services 3,4,5,6 the corresponding providers, obtained by our Randomized Algorithm (given in Section 3), are 3,3,5,5. Now, from the Service cost matrix (given in Table 1) the service costs are 7,4,2,7 (that totals to 20 which is the SC value). From the Communication cost matrix (given in Table 2) the communication costs of the links (3,3),(3,5),(3,5),(5,5) are 0,5,5,0 (that totals to 10 which is the CC value). The running time of the algorithm for this request is 0.640 seconds.

From the experiment results we find that the maximum percentage deviation from the optimal value is around 43%. We have taken different types of service request graphs (DAGs). The running time is within 2 seconds.

5 Conclusion

In this paper we addressed the problem of dynamic web services composition considering two QoS parameters (communication cost and service cost). Most available techniques for dynamic services composition deal with a single QoS parameter, and they provide optimal solutions. However there are few works that consider multiple parameters for composition. These works use some types of evolutionary algorithms to obtain results that are close to optimal. We have suggested a randomized algorithm for obtaining suboptimal solution. Our results are compared with possible optimal solutions. The experiments performed demonstrate that the deviation from optimal solution is satisfactory.

However as part of our future and ongoing work we need to make our prototype system capable of handling different types of overlay network and more varied class of user requests. We also would like to study nonlinear combination of the QoS parameters. An interesting problem would be to design an algorithm for dynamic web services composition considering more than two parameters.

References

1. Wang, J., Wang, J., Chen, B., Gu, N.: Minimum Cost Service Composition in Service Overlay Network. *World Wide Web Journal* 14, 75–103 (2011)
2. Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., Sheng, Q.Z.: Quality Driven Web Services Composition. In: *International Conference on WWW 2003*, May 20–24, pp. 411–421 (2003)
3. Karakoc, E., Kardas, K., Senkul, P.: Workflow-Based Web Service Composition System. In: *Web Intelligence and Intelligent Agent Technology Workshops (WI-IAT 2006 Workshops)*, pp. 113–116 (2006)
4. Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., Weerawarana, S.: Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing* 6(2), 86–93 (2002)
5. Alamri, A., Eid, M., Saddik, A.E.: Classification of the state-of-the-art dynamic web services composition techniques. *Int. J. Web and Grid Services* 2(2), 148–166 (2006)
6. Piccinelli, G., Williams, S.L.: Workflow: A Language for Composing Web Services. In: van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. (eds.) *BPM 2003*. LNCS, vol. 2678, pp. 13–24. Springer, Heidelberg (2003)
7. Huang, A.F.M., Lan, C., Yang, S.J.H.: An optimal QoS-based Web service selection scheme. *Information Sciences* 179(19), 3309–3322 (2009)
8. Batouche, B., Naudet, Y., Guinand, F.: Semantic Web Services Composition Optimized by Multi-objective Evolutionary Algorithms. In: *Fifth International Conference on Internet and Web Applications and Services (ICIW)*, May 9–15, pp. 180–185 (2010)
9. Lécué, F., Silva, E., Pires, L.F.: A framework for dynamic web services composition. In: *2nd ECOWS Workshop on Emerging Web Services Technology (WEWST)* (November 2007)
10. Vander Meer, D., Datta, A., Dutta, K., Thomas, H., Ramamritham, K., Navathe, S.B.: FUSION: a system allowing dynamic Web service composition and automatic execution. In: *IEEE International Conference on E-Commerce (CEC)*, June 24–27, pp. 399–404 (2003)
11. Zhao, C., Duan, Z., Zhang, M.: A Model-Driven Approach for Dynamic Web Service Composition. In: *World Congress on Software Engineering (WCSE 2009)*, May 19–21, vol. 4, pp. 273–277 (2009)
12. Gu, X., Nahrstedt, K., Chang, R.N., Ward, C.: QoS-assured service composition in managed service overlay networks. In: *23rd International Conference on Distributed Computing Systems*, May 19–22, pp. 194–201 (2003)
13. Strunk, A.: QoS-Aware Service Composition: A Survey. In: *IEEE 8th European Conference on Web Services (ECOWS)*, December 1–3, pp. 67–74 (2010)
14. Khan, F.H., Javed, M.Y., Bashir, S., Khan, A., Khiyal, M.S.H.: QoS based dynamic web services composition and execution. *International Journal Computer Science and Information Security* 7(2), 147–152 (2010)
15. Rodriguez-Mier, P., Mucientes, M., Lama, M.: Automatic Web Service Composition with a Heuristic-Based Search Algorithm. In: *IEEE International Conference on Web Services (ICWS)*, July 4–9, pp. 81–88 (2011)

Data Summarization Model for User Action Log Files

Eleonora Gentili, Alfredo Milani, and Valentina Poggioni

Dipartimento di Matematica e Informatica,
Università degli Studi di Perugia
via Vanvitelli 1, Perugia, Italy
{eleonora.gentili,milani,poggioni}@dmi.unipg.it
<http://www.dmi.unipg.it>

Abstract. During last years we have seen an impressive growth and diffusion of applications shared and used by a huge amount of users around the world, like for example social networks, web portals or elearning platforms. Such systems produce in general a large amount of data, normally stored in its raw format in log file systems and databases. To prevent an unmanageable growing of the necessary space to store data and the breakdown of data usability, such data can be condensed and summarized to improve reporting performance and reduce the system load. This data summarization reduces the amount of space that is required to store software data but produces, as a side effect, a decrease of their informative capability due to an information loss. In this work the problem of summarizing data obtained by the log systems of applications with a lot of users is studied. In particular a model to represent these raw data as temporal events collected in time sequences is proposed, methods to reduce the data size, collapsing the descriptions of more events in a unique descriptor or in a smaller set of descriptors, are provided and the optimal summarization problem is posed.

1 Introduction

Chronology-dependent applications (CDA) are applications that produce log files in which system states and activities over the time are recorded. For example, network traffic data log systems, alarms in telecommunication networks and web portals which records the user activities are CDA producing in general large amount of data in the form of log sequences, stored in log file systems and databases. Mining historical data about events is a useful way to understand and optimize system behaviors.

Event sequences capture the information of system and user activities over time, and event log analysis poses significant challenges: by examining event patterns, system administrators can set up event and incident management rules to eliminate or mitigate IT risks. It has become a standard way to manage large scale distributed systems in IT companies [7]. But log files are usually big and noisy, and the difficulty of finding patterns is very high as well as the number of patterns discovered can be very large [19].

In the case of applications with thousands of users, a simple log file can contain millions of records representing instantaneous events. The volume necessary to hold and manage them could be huge. Data in such log files can be condensed and summarized to improve reporting performance and reduce the system load, because it could be more practical and meaningful for analysts to navigate in the chronology of summarized data which gather several events about a same topic, rather than to navigate the total amount of data. For these reasons data summarization could be very useful both in order to reduce space and to aggregate information.

Most existing research efforts focus on episode mining or frequency pattern discovery: these methods simply report a large number of frequent episodes or patterns, but they fail to provide a brief and comprehensible event summary revealing the big picture that the dataset embodies [2]. Instead of discovering frequent patterns, recent works on event mining have been focused on event summarization, compacting time sequences of events into shorter ones [8], [9], [19].

In this work, events are defined as entities characterized by a set of properties that are common to many real applications: in particular, each event is described by a set of users which made actions over objects either at a particular instant or during an interval. The time model we assume is discrete, events are instantaneous and are clearly representable by a tuple like (u, a, o, t) , with which we describe that a user u performed the action a over the object o at time t .

We present a new method to produce a concise summary of sequences of events related to time, based on the data size reduction obtained merging time intervals and collapsing the descriptions of more events in a unique descriptor or in a smaller set of descriptors. Moreover, in order to obtain a data representation as compact as possible, an abstraction operation allowing an event generalization process (as in [13]) is defined. Time sequence summarization takes as input a time sequence in which events are chronologically ordered, and aims to produce a summarized time sequence that can substitute the original time sequence in the data analysis.

The reduction of the amount of data produces also as a side effect, a decrease of their informative capability due to an information loss. For this reason, we formally define the summarization problem as an optimization problem that balances between shortness of the summary and accuracy of the data description.

1.1 Related Works

The summarization problem is presented also in [5], [7], [8], [9], [10], [13], [19], which addressing the problem from different points of view.

In [13], Pham et al. have proposed an algorithm that achieves time sequence summarization based on a generalization, grouping and concept formation process. Generalization expresses event descriptors at higher levels of abstraction using taxonomies while grouping gathers similar events. Concept formation is responsible for reducing the size of the input time sequence of events by representing each group created by one concept. The process is performed in a way

such that the overall chronology of events is preserved. The algorithm computes the summary incrementally and has reduced algorithmic complexity.

A different approach to the problem is in [8], [9], where Kiernan and Terzi rely to the Maximum Description Length principle to produce a comprehensive summary of an event sequence, where events are taken from a set of different event types. They have formally defined the summarization problem as an optimization problem that balances between shortness of the summary and accuracy of the data description. They have shown that this problem can be solved optimally in polynomial time by using a combination of two dynamic-programming algorithms. It has been also explored more efficient greedy alternatives and demonstrated that they work well on large datasets. Kiernan and Terzi also present in [10] a tool for summarizing large event sequences based on work proposed in [8], [9]. This method reveals the local patterns of the sequence but fails to provide the inter-segment relationships.

A more sophisticated method to summarize time sequence of events is the one proposed by Wang et al. in [19]. They are able to describe the patterns and also learn a Hidden Markov Model to characterize the global relationships by partitioning the time sequence into segments. This method can produce a comprehensive summary for the input sequence, but it only focuses on the frequency changes of event types across adjacent segments and ignore the temporal information among event types within a segment, generates the same number of event patterns with the same boundaries for all event types and the generated summary is difficult for system administrators to understand and take actions.

A novel framework called “natural event summarization” is proposed by Jiang et al. in [7]. This method summarizes an event sequence using inter-arrival histograms to capture the temporal relationships among events using the minimum description length principle to guide the process in order to balance between accuracy and brevity. Moreover, the authors use multi-resolution analysis for pruning the problem space. The summary created by this procedure is usable, robust to noise and scalable.

In [5], Chandola et al. formulates the problem of summarization of transactions that contain categorical data, introducing two measures, Compaction Gain and Information Loss, to assess the quality of the summary. The optimal summarization procedure consists of an optimization problem involving these two measures. They investigate two approaches to address the problem: the first one consists of an adaptation of clustering techniques and the second technique regards the use of frequent sets of items from the association analysis domain.

1.2 Roadmap

The rest of the paper is organized as follows: in Section 2 we present the summarization model, giving some basic definitions, properties and notational conventions. In Section 3, the optimal summarization problem is formally described, taking into account the metrics defined to assess the quality of a summarized time sequence with respect to the initial volume of data. Conclusions and future works are exposed in Section 4.

2 The Model

In this work we define events as entities that are characterized by a set of properties that are common to many real applications: each event is described by a set of users which made actions over objects either at a particular instant or during an interval.

The log files of most applications stores data containing this information. For instance, Moodle, one of the most popular e-learning software platform, produces a log file containing each action (login, read, upload, etc...) performed by a user with the user role, his/her IP address, and the object involved, as shown in Fig. 1; instead Facebook produces log files containing different information, according to the actions performed by users: for example, when an user writes on the wall of a friend, log file records information about user name, type of action, update time and eventual “likes” and “comments” of other users. Another example of platform that produces log messages is the Hadoop system [1], in which there is a sequence of terms that indicate the date and the time, the handler name, the port number and the action happened [16].

Time	IP Address	Full name	Action	Information
Tue 5 September 2006, 01:33 AM	70.109.156.137	Teacher Demo	course report live	Moodle Features Demo
Tue 5 September 2006, 01:33 AM	128.173.54.50	Student Demo	resource view	How to install the Features
Tue 5 September 2006, 01:33 AM	128.173.54.50	Student Demo	course view	Moodle Features Demo
Tue 5 September 2006, 01:32 AM	128.173.54.50	Student Demo	hotpot view	3
Tue 5 September 2006, 01:32 AM	128.173.54.50	Student Demo	hotpot view	4
Tue 5 September 2006, 01:32 AM	128.173.54.50	Student Demo	hotpot view all	
Tue 5 September 2006, 01:32 AM	128.173.54.50	Student Demo	course view	Moodle Features Demo
Tue 5 September 2006, 01:30 AM	72.147.138.34	Admin User	calendar add	Test Results on Tuesady
Tue 5 September 2006, 01:30 AM	72.147.138.34	Admin User	forum view discussion	Linear Equations
Tue 5 September 2006, 01:29 AM	72.147.138.34	Admin User	course view	Moodle Features Demo
Tue 5 September 2006, 01:29 AM	72.147.138.34	Admin User	forum add discussion	Linear Equations
Tue 5 September 2006, 01:28 AM	72.147.138.34	Admin User	course view	Moodle Features Demo
Tue 5 September 2006, 01:25 AM	70.109.156.137	Teacher Demo	user view	Teacher Demo
Tue 5 September 2006, 01:25 AM	70.109.156.137	Teacher Demo	course report particip	4
Tue 5 September 2006, 01:18 AM	70.109.156.137	Teacher Demo	course report live	Moodle Features Demo

Fig. 1. An example of log file produced by Moodle

The time model we assume is discrete, events are instantaneous and are clearly representable by a tuple like (u, a, o, t) , with which we describe that a user u performed the action a over the object o at time t .

In the case of applications with thousands of users, such a simple log file can contain millions of records representing instantaneous events. The volume necessary to hold and manage them could be huge. Moreover, this information is very often more useful when it is aggregated, because it can answer to specific requirements or can be handled with data mining tools. For these reasons data summarization could be very useful both in order to reduce space and to aggregate information.

In order to represent more general situations with respect the above definition of instantaneous events and potentially aggregate information of similar events, we define events as tuples involving sets of users, actions and objects possibly occurred during an interval. The underlying meaning is that all the users in a set U performed all the actions in a set A over all the objects in a set O at a time instant t or during an interval I .

For instance, consider a user of an application that reads some email at different and not consecutive time instants. The log file contains several different inputs, one of each is related to a user action. Our aim is to specify a model that can easily allow to summarize all these information in a more compact structure.

It is clear that in this summarization context taxonomies and hierarchies play a significant role when defined over the sets involved in the definition of events. Moreover, it is very common that taxonomies are used to represent hierarchies of users or different kinds of actions. For example, in the Moodle platform users are organized in hierarchies depending on their roles.

Definition 1. *An event descriptor is a t -uple*

$$X = (U, A, O, I, \delta)$$

representing a set of actions A made by a set of users U over a set of objects O , during a given time interval I according to the covering index δ that is defined as the ratio by the number of points in which the actions in A are actually executed and all the points of I .

It is clear that each instantaneous event is representable by a degenerate interval $I = [t_0, t_0]$ associated to a *covering index* $\delta = 1.0$.

Example 1. The event descriptor $X = (\{admin\}, \{login, logout\}, \{IT\}, [10, 50], 0.30)$ represents that the user *admin* made the action *login* and *logout* in the IT e-learning course in the 30% of the points of time interval $I = [10, 50]$. \square

Moreover, we assume that the labels in the sets U , A and O can be organized in taxonomies. In this work we allow taxonomies organized in hierarchies with multiple inheritance and we consider that each taxonomy is associated to an *abstraction* operator, represented by the symbol \uparrow , allowing to climb the hierarchy. Labels of the sets U , A , O are arranged in hierarchical taxonomy graphs, i.e. acyclic directed graphs, instead of common trees, because of their multiple inheritance property. In a natural way, we consider that the abstraction operator applied to a node of the taxonomy returns all the fathers of the node, i.e. $\uparrow(n) = fathers(n)$. While, when the abstraction operator is applied to a set of nodes $S = \{s_1, \dots, s_n\}$, the result is a set $\uparrow(S) = S'$ such that at least a s_i is substituted with $\uparrow(s_i)$. Let two different sets S_1 and S_2 , we define the *minimal abstracting set* as the first not null set S of common ancestor of S_1 and S_2 computed by climbing the taxonomy graph associated to S_1 and S_2 , i.e. such that $S = \uparrow(S_1) = \uparrow(S_2)$.

Definition 2. *A time sequence is a set of event descriptors $\mathbf{X} = (X_1, \dots, X_m)$, where each X_i is a tuple $X_i = (U_i, A_i, O_i, I_i, \delta_i)$; m is called size of the time sequence.*

Given a *time sequence* the aim of this work is to provide methods to reduce its data size, collapsing the descriptions of more events in a unique descriptor or in a smaller set of descriptors. These methods can act aggregating different events having common information about users, actions and objects. Moreover, when taxonomies are defined, taking up a loss information as side effect, is possible to process data before merging them in order to become mergeable some events that otherwise they would not be.

In order to reduce data size of the initial *time sequence*, we define a *merging operator* that can collapse intervals of *event descriptors* with identical label sets U , A and O . It is obvious that a merging process involves a change in the *covering index*.

Definition 3. Let Ω the set of all event descriptors and given two event descriptors $X_1 = (U, A, O, I_1, \delta_1)$ and $X_2 = (U, A, O, I_2, \delta_2)$ with the same label sets U , A and O , and let $I_1 = [t'_1, t''_1]$ and $I_2 = [t'_2, t''_2]$, we define the merging operator as

$$\begin{aligned} \oplus : \Omega \times \Omega &\rightarrow \Omega \\ ((U, A, O, I_1, \delta_1), (U, A, O, I_2, \delta_2)) &\mapsto (U, A, O, I, \delta) \end{aligned} \tag{1}$$

such that $I = [\min(t'_1, t'_2), \max(t''_1, t''_2)]$ and

$$\delta = \frac{\delta_1|I_1| + \delta_2|I_2| - \min(\delta_1|I_1|, \delta_2|I_2|, |I_1 \cap I_2|)}{|I|} \tag{2}$$

The *covering index* δ is computed according to a pessimistic approach. In fact, we assume the “worst” case situation in which events happening in both I_1 and I_2 coincide as much as possible; it is simple to prove that the *covering index* δ is always less than or equal to $\max(\delta_1, \delta_2)$. For this reason, applying the *merging operator* we have a loss of information about data.

Example 2. Let a web portal storing information about users and their actions in a log file, modeled by the time sequence $\mathbf{X} = \{X_1, X_2, X_3, X_4, X_5, X_6\}$, and let

$$\begin{aligned} X_1 &= (\{user_1, user_2\}, \{login\}, \{objA\}, [1, 1], 1.0), X_2 = (\{user_2\}, \{send\}, \{objA\}, [2, 2], 1.0), \\ X_3 &= (\{user_2\}, \{send\}, \{objA\}, [3, 3], 1.0), X_4 = (\{user_1\}, \{read\}, \{objA\}, [4, 4], 1.0), \\ X_5 &= (\{user_2\}, \{send\}, \{objA\}, [5, 5], 1.0), X_6 = (\{user_1, user_2\}, \{logout\}, \{objA\}, [6, 6], 1.0). \end{aligned} \tag{3}$$

The *merging operator* can be applied among X_2 , X_3 and X_5 obtaining

$$\begin{aligned} X_{23} &= X_2 \oplus X_3 = (\{user_2\}, \{send\}, \{objA\}, [2, 3], 1.0), \\ X_{235} &= X_{23} \oplus X_5 = (\{user_2\}, \{send\}, \{objA\}, [2, 5], 0.75). \end{aligned}$$

The reduction in data size has the side effect of a reduction in accuracy of the representation: the new *time sequence* $\mathbf{X} = \{X_1, X_{235}, X_4, X_6\}$ has a smaller size, but the information about exact location of events *read* has been lost.

Accordingly to Definition 3, no more *merging operation* can be applied. \square

Although the *merging operator* is very useful, it can be applied only between event descriptors having the same label sets.

In order to overcome this restriction, i.e. to reduce the size of a *time sequence* in which *event descriptors* have different label sets, we introduce an *abstraction operator* that, generalizing labels in the sets U , A and O , will make mergeable *event descriptors* that otherwise would not be.

Definition 4. Let Ω the set of all event descriptors and given an event descriptor $X = (U, A, O, I, \delta)$, the abstraction operator \uparrow_S is defined as

$$\begin{aligned} \uparrow_S: \Omega &\rightarrow \Omega \\ (U, A, O, I, \delta) &\mapsto (U', A', O', I, \delta) \end{aligned} \tag{4}$$

where $S \in \{U, A, O\}$ stating which set among U , A and O is modified and $S' = \uparrow(S)$ is obtained by means the abstraction operator \uparrow as defined above.

As said before, the *abstraction operator* can be applied to *event descriptors* having different label sets to generalize their labels.

Definition 5. Let $\{X_i : X_i = (U_i, A_i, O_i, I_i, \delta_i), i = 1, \dots, n\}$ a set of event descriptors, $X_i^* = (U, A, O, I_i, \delta_i)$ is the minimal abstracting event for X_i if each label set U , A , O is the minimal abstracting set respectively for $\{U_i\}$, $\{A_i\}$, $\{O_i\}$.

For instance, let consider the taxonomy graphs depicted in Fig 2, and let

$$\begin{aligned} X_1 &= (\{user_1\}, \{Create.folder, Save\}, \{log_1\}, I_1, \delta_1), \\ X_2 &= (\{user_1, user_2\}, \{Disk.op, Write.doc\}, \{log_1\}, I_2, \delta_2), \end{aligned}$$

the *abstraction process* can be applied to U_1 , U_2 , A_1 and A_2 in order to find the minimal abstracting sets U and A , obtaining the two (mergeable) *event descriptors* $X_1^* = (\{user\}, \{User.op\}, \{log_1\}, I_1, \delta_1)$ and $X_2^* = (\{user\}, \{User.op\}, \{log_1\}, I_2, \delta_2)$ that are respectively the *Minimal Abstracting Event* for X_1 and X_2 .

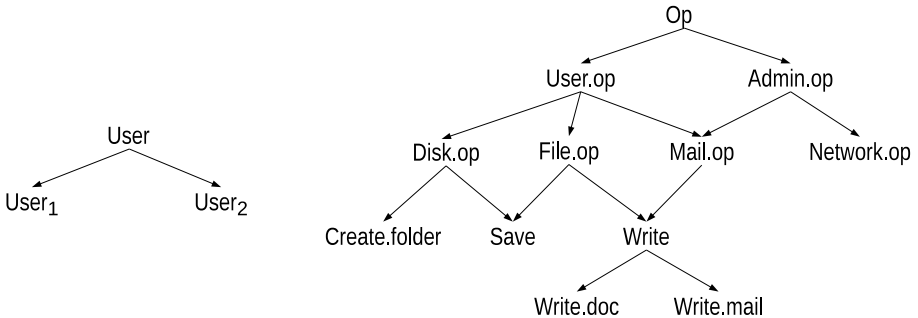


Fig. 2. An example of taxonomy graphs respectively over the sets U and A

It is clear that the new *event descriptors* X_1^* and X_2^* represent the original events with less precision than the original ones because the original data hold information about the exact users $user_1$ and $user_2$, while the new descriptors hold a more generic information about the group $user$. The same is for the sets A_1 and A_2 : in the original version of descriptors the stored information is more precise than the one represented in the *abstracted* version. So, considering that creating a summary is quite natural and common to have information loss, the final aim is to find a plan of summarization operations that maximize the reduction with respect to the initial volume of data minimizing the information loss.

It is clear that the optimal summarization (i.e. summarization with minimum information loss) is a question of tradeoff between application of the *merging operator* and the *abstraction operator* to the *event descriptors*.

3 The Optimal Summarization Problem

As presented in the previous section, given a time sequence \mathbf{X} , there are several possibilities to summarize it and reach the aimed data size.

The *summarization procedure* is oriented along two directions: reducing the data size and abstracting labels. The merging process combines together different *event descriptors*, obtaining a smaller set of descriptors with less *covering indexes*. The *abstraction process* transforms *event descriptors* generalizing them and making possible to apply further the *merging operators*.

Let \mathbf{X}_0 the time sequence of the initial volume of data and \mathbf{X} a summarized time sequence, we define some metrics to assess the quality of \mathbf{X} with respect to \mathbf{X}_0 .

Definition 6 (Compaction Gain)

The *compaction gain* of a time sequence \mathbf{X} is define as $\mathcal{C}(\mathbf{X}, \mathbf{X}_0) = \frac{|\mathbf{X}_0|}{|\mathbf{X}|}$.

Definition 7 (Covering Accuracy)

Let $I_i = [t'_i, t''_i]$ and $G_i = [t'_i, t'_{i+1}]$, we define the *covering accuracy* of a time sequence \mathbf{X} as

$$\mu(\mathbf{X}) = \frac{\sum_{i=1}^n \delta_i |I_i| + \sum_{i=1}^m |G_i|}{\sum_{i=1}^n |I_i| + \sum_{i=1}^m |G_i|}. \tag{5}$$

Note that the gap intervals G_i are considered as intervals with $\delta_{G_i} = 1.0$. In fact, in these intervals we have the maximum information we can have, i.e. we are sure that there are no events.

It easy to prove that $0 \leq \mu(\mathbf{X}) \leq 1$. In particular, $\mu(\mathbf{X}) = 1$ is verified if and only if $\delta_i = 1, \forall i = 1, \dots, n$.

Definition 8 (Description Accuracy)

Let \mathbf{X} a time sequence, the *description accuracy* of \mathbf{X} is defined as

$$\eta(\mathbf{X}) = \min_{X \in \mathbf{X}} (\min(\omega_U \eta(U), \omega_A \eta(A), \omega_O \eta(O))), \tag{6}$$

where $\omega_U, \omega_A, \omega_O \geq 0$ are constants that can be used to differ the weights of label sets, and

$$\eta(S) = \min_{n \in S} \frac{d(r, n)}{h(n)},$$

where $S \in \{U, A, O\}$, r is the root of the taxonomy T_S and $h(n)$ is the longest distance from n to a leaf.

Note that $0 \leq \eta(S) \leq 1$. In particular, $\eta(S) = 1$ is verified when n is a leaf, and $\eta(S) = 0$ when n coincides with the root.

Remark 1. Let \mathbf{X}_0 a time sequence and \mathbf{X}^* the summarized time sequence obtained from \mathbf{X}_0 replacing each X_i with its *minimal abstracting event* X_i^* , then it is easy to prove that \mathbf{X}^* as the *maximum description accuracy* among all the \mathbf{X} making mergeable all the descriptors in \mathbf{X}_0 , i.e. does not exist another time sequence \mathbf{X} making mergeable all the descriptors in \mathbf{X}_0 causing a less decrease in *description accuracy* than \mathbf{X}^* .

Definition 9 (Information Loss)

Let \mathbf{X}_0 a time sequence and \mathbf{X} a summarized time sequence obtained from \mathbf{X}_0 , the *information loss* of the summarization process is defined as

$$\mathcal{I}(\mathbf{X}, \mathbf{X}_0) = \alpha(\mu(\mathbf{X}_0) - \mu(\mathbf{X})) + \beta(\eta(\mathbf{X}_0) - \eta(\mathbf{X})). \tag{7}$$

An *optimal summarization problem* is the problem to find a sequence of operators \oplus and \uparrow such that it optimizes the tradeoff between $\mathcal{C}(\mathbf{X}, \mathbf{X}_0)$ and $\mathcal{I}(\mathbf{X}, \mathbf{X}_0)$.

Definition 10 (Optimal Summarization Problem)

Let the time sequence \mathbf{X}_0 and a real number $\gamma > 0$, we define *Optimal Summarized Time Sequence*, the time sequence $\bar{\mathbf{X}}$ such that the parameterized ratio between $\mathcal{I}(\mathbf{X}, \mathbf{X}_0)$ and $\mathcal{C}(\mathbf{X}, \mathbf{X}_0)$ is minimal, i.e.

$$\bar{\mathbf{X}} = \operatorname{argmin}_{\mathbf{X}} \frac{\mathcal{I}(\mathbf{X}, \mathbf{X}_0)}{[\mathcal{C}(\mathbf{X}, \mathbf{X}_0)]^\gamma}$$

4 Conclusion and Future Work

In this work the problem of summarizing data obtained by the log systems of applications with a lot of users is studied. In particular, a model to represent these raw data as temporal events collected in time sequences is proposed, methods to reduce the data size are provided and the optimal summarization problem is posed. This novel approach aims to understand, interpret and analyze temporal log data.

We have presented a new method to produce a concise summary of sequences of events related to time, based on the data size reduction obtained merging time intervals and collapsing the descriptions of more events in a unique descriptor or in a smaller set of descriptors. Moreover, in order to obtain a data representation

as compact as possible, an abstraction operation allowing an event generalization process is defined.

The time sequence summarization process proposed in this work takes as input a time sequence of chronologically ordered event descriptors, characterized by the label sets U, A, O and the time interval I and aims to produce a summarized time sequence that can substitute the original one in the data analysis. The reduction of the amount of data produces also, as a side effect, a decrease of their informative capability due to an information loss. For this reason, we have formally defined the summarization problem as an optimization problem that balances between shortness of the summary and accuracy of the data description.

Moreover, we are studying about the formalization and the implementation of an optimal algorithm for the *Optimal Summarization Problem*. Preliminary results using suboptimal algorithms are already obtained and they present encouraging results that deserve of further investigations.

References

1. Hadoop: an Open-Source MapReduce computing platform, <http://hadoop.apache.org>
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE), pp. 3–14 (1995)
3. Allen, J.F.: An interval-based representation of temporal knowledge. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, vol. 1, pp. 221–226 (1981)
4. Allen, J.F.: Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11), 832–843 (1983)
5. Chandola, V., Kumar, V.: Summarization—compressing data into an informative representation. *Knowledge and Information Systems* 12(3), 355–378 (2007)
6. Costantini, A., Tasso, S., Gervasi, O.: It Visualization and Web Services for Studying Molecular Properties. In: *Computational Science and Applications*, pp. 222–228 (2009) ISBN 978-0-7695-3701-6
7. Jiang, Y., Perng, C.S., Li, T.: Natural event summarization. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 765–774. ACM (2011)
8. Kiernan, J., Terzi, E.: Constructing comprehensive summaries of large event sequences. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 417–425. ACM (2008)
9. Kiernan, J., Terzi, E.: Constructing comprehensive summaries of large event sequences. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3(4), 21 (2009)
10. Kiernan, J., Terzi, E.: EventSummarizer: A tool for summarizing large event sequences. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 1136–1139. ACM (2009)
11. Pallottelli, S., Tasso, S., Pannacci, N., Costantini, A., Lago, N.F.: Distributed and Collaborative Learning Objects Repositories on Grid Networks. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B.O. (eds.) *ICCSA 2010. LNCS*, vol. 6019, pp. 29–40. Springer, Heidelberg (2010)

12. Peng, W., Perng, C., Li, T., Wang, H.: Event summarization for system management. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1028–1032 (2007)
13. Pham, Q.K., Raschia, G., Mouaddib, N., Saint-Paul, R., Benatallah, B.: Time sequence summarization to scale up chronology-dependent applications. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 1137–1146 (2009)
14. Povinelli, R.J.: Identifying temporal patterns for characterization and prediction of financial time series events. In: Temporal Spatial and SpatioTemporal Data Mining, pp. 46–61 (2001)
15. Saint-Paul, R., Raschia, G., Mouaddib, N.: General purpose database summarization. In: Proceedings of the 31st International Conference on Very Large Data Bases, pp. 733–744. VLDB Endowment (2005)
16. Tang, L., Li, T., Perng, C.S.: LogSig: Generating system events from raw textual logs. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 785–794. ACM (2011)
17. Tasso, S., Pallottelli, S., Bastianini, R., Lagana, A.: Federation of Distributed and Collaborative Repositories and Its Application on Science Learning Objects. In: Murgante, B., Gervasi, O., Iglesias, A., Tanar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part III. LNCS, vol. 6784, pp. 466–478. Springer, Heidelberg (2011)
18. Wang, J., Karypis, G.: On efficiently summarizing categorical databases. *Knowledge and Information Systems* 9(1), 19–37 (2006)
19. Wang, P., Wang, H., Liu, M., Wang, W.: An algorithmic approach to event summarization. In: Proceedings of the 2010 International Conference on Management of Data, pp. 183–194. ACM (2010)

User Modeling for Adaptive E-Learning Systems

Birol Cilogluligil¹ and Mustafa Murat Inceoglu²

¹ Ege University, Department of Computer Engineering,
35100 Bornova, Izmir, Turkey

² Ege University, Department of Computer Education and Instructional Technology,
35100 Bornova, Izmir, Turkey
{birol.cilogluligil,mustafa.inceoglu}@ege.edu.tr

Abstract. Adaptive systems have been a hot topic in various areas like hypermedia systems, e-commerce systems, e-learning environments and information retrieval. In order to provide adaptivity, these systems need to keep track of different types of information about their users. Therefore, user modeling is at the heart of the adaptation process. In this paper, different user modeling techniques will be reviewed with the focus on what needs to be modeled and how it will be modeled, i.e., the demographic information of the users are collected in most of these systems, however, how it will be used in the adaptation process depends on the methodology being followed. The evaluation of different user modeling approaches and examination of some recent adaptive e-learning systems' architectures will also be provided.

Keywords: Adaptive Systems, E-Learning, Ontologies, Personalization, User Modeling, Semantic Web.

1 Introduction

Adaptivity makes it possible for a system to behave in a different way for different users. In order to achieve that feature, adaptive systems need a user model which holds information about individual users. In other words, user models keep representations of information about different users in the system. In the e-learning context, a user model can be defined as an abstract image of the learner in the system [1]. There are two types of inputs when collecting data about users when creating a user model;

- Requesting direct input from users “explicitly”
- Observing user's interaction with the system and automatically collecting information “implicitly”

The success of a system's adaptivity depends on the user model and the properties of the data represented in user models depend on the design of the adaptation process. Therefore, user modeling and adaptation are two sides of the same coin [2].

User modeling and adaptivity are applied in different domains such as information retrieval, e-commerce systems, intelligent tutoring systems, adaptive educational

systems, adaptive hypermedia systems such as online news systems [3], museum guide systems [4] and many more. From here on, the focus will be on e-learning systems starting with a brief introduction to the e-learning field below.

The main aim of the studies in the e-learning field is to provide the opportunity of studying without time and place constrictions to the learners. With these systems, it is possible for a learner to access whichever course content he/she seeks, whenever he/she wants and wherever he/she is. With the rapid development at the field, nowadays most of the universities deliver online courses and/or offer e-learning programs. Most educational institutions use learning management systems (LMS) that provide huge benefits for the teachers to organize their courses and put the course materials online. However, from the learners' point of view, one of the main problems with LMSs and most e-learning systems is that the systems don't consider individual differences of the learners when courses are composed. Therefore, LMSs provide exactly the same course with the same content and structure to every user.

On the other hand, every user has different characteristics, interests, motivation, cognitive abilities, learning styles and different familiarity levels with the subject being taught. Thus, taking into account each user's needs is a must for the success of the adaptation process of the system.

Personalization looks for ways in which the learners' needs could be identified and the content could be adapted to suit those needs. Adaptation of the e-learning systems to support the learners' individual needs can be achieved in two different ways [5]:

- Those that allow the learner to change certain system parameters and adapt their behavior accordingly are called “adaptable” systems
- Those that adapt to the learners' needs in an intelligent form and are automatically based on the system's conjecture are called “adaptive” systems

In this study, the focus will be on the adaptive systems to examine adaptive e-learning system infrastructures that aim at taking the individual preferences of the learners into consideration and tailoring the courses and contents according to them.

In the next section, the contents of user modeling will be given with the focus on what is being modeled in user models to achieve adaptivity in e-learning systems. Section 3 discusses how these concepts of the user models are going to be modeled. A review of some of the recent adaptive e-learning systems' architectures in practice is provided in Section 4. Section 5 concludes the paper by discussing the challenges of developing user models and adaptive systems.

2 What Is Being Modeled?

User models can be classified by the nature and form of information contained in the model as well as the methods of working with it [6]. According to this approach, user models are analyzed in three layers:

- what is being modeled (nature)
- how this information is represented (structure)
- how different kinds of models are maintained (user modeling approaches)

In this paper, we are going to discuss these layers in Sections 2, 3 and 4 respectively, starting with this section. The first two layers are contained in sections with the names in correspondence with them. However, the third one will be reviewed in Section 4 with different system architectures.

Here in this section, the fundamental terms that compose the background of the study will be given as they are the concepts being modeled.

Some of the learner needs are intellectual ability, cognitive styles, learning styles, prior knowledge, anxiety, achievement motivation, and self-efficacy [7].

However, the learner needs can be classified as follows [2];

- User Knowledge
- User's Interests
- User's Goals and Tasks
- User's Background
- Individual Traits
- Context of Work

In order to provide personalization, personalization strategies that are composed of personalization parameters are needed [8]. Personalization parameters represent the learners' prior knowledge, interests, motivation, cognitive abilities, learning styles and so on.

The personalization parameters constitute the source for personalization of e-learning scenarios. 16 personalization parameters that are most commonly used in the e-learning domain are explained in [8]. These parameters are given in alphabetical order in Table 1 with the set of values for each personalization parameter.

The e-learning systems in the literature which support personalization are reviewed in [8] and it is shown that the most commonly used personalization parameter is the learners' level of knowledge applied in personalized e-learning systems such as ELM-ART [15], PERSO [16] and ActiveMath [17]. The other commonly used personalization parameters are learning goals and media preference.

Out of the 16 personalization parameters, most of the systems use at most three personalization parameters. Learning style models are being used in recent studies such as [13] and [18] where the Kolb learning cycle and Felder-Silverman learning style are used respectively. Considering the fact that several combinations of the personalization parameters have not been tested yet, determining which and how many personalization parameters to use when designing personalized e-learning systems is an open research problem [8].

There are many learning style models in the literature such as Kolb [21], Honey and Mumford [11], and Felder and Silverman [10]. Felder and Silverman learning styles model (FSLSM) is the most commonly used model in adaptive systems based on learning styles. FSLSM is a four-dimensional model where each dimension is a linguistic variable. These dimensions are active/reflective, sensing/intuitive, visual/verbal and sequential/global. FSLSM is the Cartesian product of the linguistic variables each representing its dimension. For example, a learner can be active, sensing, visual and global.

A comparison of learning style theories on the suitability for e-learning is provided by [22] and it suggests that FSLSM is the most appropriate model to use in adaptive e-learning systems.

Table 1. Examples of Values for Personalization Parameters [8]

Personalization parameter	Set of values
Cognitive traits [9]	{low, high working memory capacity} x {low, high inductive reasoning ability} x {low, high information processing speed} x {low, high associative learning skills}
Felder–Silverman learning style [10]	{sensing, intuiting} x {visual, verbal} x {active, reflective} x {sequential, global}
Honey–Mumford learning style [11]	{activist, reflector, theorist, pragmatist}
Information seeking task [12]	{learning the structure of SDP-TA, project planning, reverse engineering, following an activity}
Kolb learning cycle [13]	{Converger, Diverger, Assimilator, Accommodator}
La Garanderie learning style [14]	{competitive, cooperative, access on the avoidance, participative, dependant, independent}
Language preference [15]	{English, German}
Learner’s level of knowledge [16]	{beginner, intermediate, advanced}
Learning goals [17]	{knowledge, comprehension, application}
Media preference [16]	{video, sound, simulation, text/image}
Motivation level [13]	{low, moderate, high}
Navigation Preference [18]	{breadth-first, depth-first}
Participation balance [19]	{tooMuch, notEnough, acceptable}
Pedagogical approach [20]	{objectivist, competencies based, collaborative}
Progress on task [19]	{small, large}
Waiting for feedback [19]	{significant, medium, low}

3 How Is the Information Represented?

User modeling is important for both representing knowledge and providing the infrastructure for adaptivity. Adaptivity can be achieved in two ways [2]; on the presentation level by presenting the content adaptively or on the navigation level by manipulating the order of content (hiding, sorting, annotating).

In order to detect the learning styles of learners, indexes for learning style identification are the most popular way. On the other hand, some studies track students’ behaviors for automatic learning style identification, while some use both of them.

Learners may be less motivated to take indexes for learning style identification, which may be time consuming. Thus, the learning styles of the students may not be determined correctly. Automatic student modeling is generally considered more accurate, because no explicit information is requested from the learners and they can concentrate only on learning.

Student modeling can be done in two ways; static and dynamic. Most of the studies in the literature are static where the learning styles are detected at the beginning of user-system interaction and do not change. However, some researchers are working on dynamic student modeling by using a combination of the following inputs [23];

- tracking student behaviors
- collecting test scores
- student selections
- student preferences
- time spent on learning units
- user feedbacks

Information retrieval and filtering systems try to find the documents that are most relevant to the users' interests and order them by their relevance. The user models of these systems represent the users' interests in terms of keywords or concepts and are referred as "user profiles". On the other hand, Intelligent tutoring systems (ITS) try to select educational activities and deliver individual feedback that are most relevant to the users' level of knowledge. The user models in ITSs represent the users' knowledge of the subject in relation to an expert-level domain knowledge and are referred as "student models" [24].

Scalar models are the simplest forms of a user knowledge model, which try to estimate the level of user domain knowledge with a single value on a quantitative (such as a number ranging from 0 to 10) or qualitative (with classes such as low, moderate, high) scale. Scalar models, especially the qualitative ones, are quite similar to stereotype models. However, the scalar models differ by focusing exclusively on user knowledge and by being produced by user self-evaluation or objective testing, not by a stereotype-based modeling mechanism. Scalar models are simple but useful, for example learners may be classified according to their level of knowledge as beginner, intermediate or advanced, and the system may serve each learner depending on the class he/she belongs.

Even though scalar models are useful for reasonably small domains, they have the disadvantage of providing classification with low precision as the domain gets larger and the different parts of the domain may have different characteristics. For example, in word processing, a user may be an expert in using text annotation, but a novice in formula editing [25]. Since a scalar model averages the user knowledge of the domain in general, it is not sufficient when an advanced adaptation technique which is going to be applied on some specific part of the user knowledge is needed. For this reason, many of the adaptive systems use structural models.

Structural models provide the body of domain knowledge to be divided into independent fragments and the user knowledge of different fragments to be

represented independently. Structural models can be classified into two different sub-dimensions according to the nature of the represented knowledge as follows [2]:

- the type of represented knowledge (declarative vs. procedural)
- a comparison of the user's knowledge represented in the model to an expert's level of knowledge of the subject, referred to as domain model, expert model, or "ideal student" model

Overlay models are the most popular form of structural knowledge models, which represent each learner's knowledge as a subset of the domain mode. An overlay model stores the estimation of the learners' knowledge level of each fragment of the domain knowledge independently.

Overlay models provide better representation options than scalar models, but they are often criticized in the field of ITS for being "too simple". It has been argued that the state of user knowledge can never be an exact subset of expert knowledge, the user may also have misconceptions and his/her knowledge does not progress to an expert-level knowledge straightforwardly by "filling the gaps", but through a complex process of generalization and refinement. In order to model user misconceptions, overlay models can be expanded into bug models, representing both correct knowledge and misconceptions (known as buggy knowledge or "bugs"). Bug models are generally used to model user procedural problem solving knowledge [2].

Genetic models have richer modeling options by making it possible to reflect the development (genesis) of user knowledge from the simple to the complex and from the specific to the general [26]. Bug models and genetic models provide more powerful modeling than the overlay models, yet they are much harder to develop. Research on these models has contributed to the development of the fields of cognitive modeling and ITS [27], but the practical use of these models has been quite limited.

4 An Overview of Adaptive E-Learning System Architectures

In this section, some of the recent adaptive e-learning systems in practice will be discussed with their user modeling approaches and architectures as the basis.

An overview of adaptive distributed e-learning systems by focusing on how personalization can be achieved is provided in [28]. The basis of the study is the technologies used for establishing adaptive learning environments such as web services, multi-agent systems, semantic web and AI techniques like case-based reasoning, neural networks and Bayesian networks.

There are many studies in the literature to classify e-learning systems with different approaches. In this section, some of these classifications will be discussed first and architectures of three adaptive e-learning systems will be discussed in detail. These three systems are selected because of the different infrastructures they provide to support personalization. The focus of the examined systems is what they provide to accommodate learning styles.

The efforts to support adaptivity in the literature can be classified as follows in the order of their frequency [23];

- the studies to propose a framework or model to support adaptivity,
- the studies to automatically detect users' learning styles,
- the studies that examine the effectiveness of learning style-based adaptive systems.

Adaptive systems are classified in two categories depending on whether they provide users the opportunity to use outer web resources as open and closed corpus systems [6].

Several approaches to e-learning systems are investigated in [29] and the systems are classified as follows:

- Distributed learning repositories, which focus on the dynamic and networking aspects,
- Learning management systems, which focus on course delivery and administrative aspects,
- Adaptive web-based educational systems, which offer personalized access and presentation facilities to learning resources for specific application domains.

With this classification, LMSs do not provide adaptivity by nature but they can be modified and extended to be adaptive. For example, in order to present the most appropriate learning objects (LO) to the learners; the LMSs' functionality is extended by [30] with adaptivity based on the learning styles by using adaptive sorting and adaptive annotation.

The three extensions in the architecture of the LMS are shown in Figure 1 [30]. The first extension deals with detecting the learning styles of students and storing them in the student model. The Index of Learning Styles (ILS), a 44-item questionnaire developed by Felder and Soloman [31], is used for detecting learning

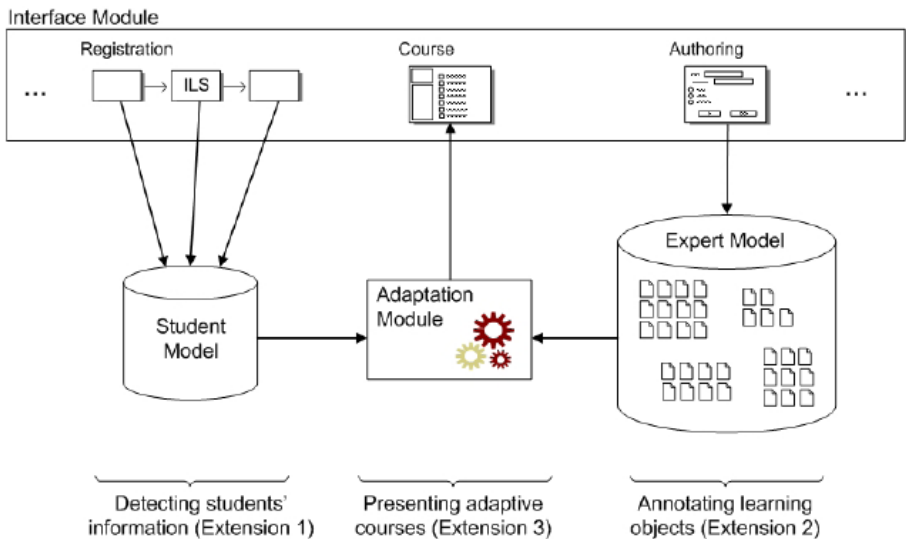


Fig. 1. Extensions of the LMS architecture proposed in [30]

styles. Second, the expert model and the authoring tool are extended to be able to distinguish between the different types of LOs. The third extension namely the adaptation module; uses the student model to calculate the values of each adaptation feature based on the students' learning styles; composes an individual course and by accessing the respective LOs through the expert model, presents it to the respective student via the interface in the LMS.

The second study to be examined in detail is a personalized multi-agent e-learning system based on item response theory (IRT) and artificial neural network (ANN) which presents adaptive tests (based on IRT) and personalized recommendations (based on ANN) proposed by [32]. It is stated that in most systems the importance of tests, constructing and adapting them to learner's ability have been neglected and those systems have only adapted the content and the personalized part of their systems is sequence and organization of the content. Test adaptability, the main contribution of the system, is achieved by integrating IRT and ANN into its multi-agent structure. The architecture of the system is shown in Figure 2 [32].

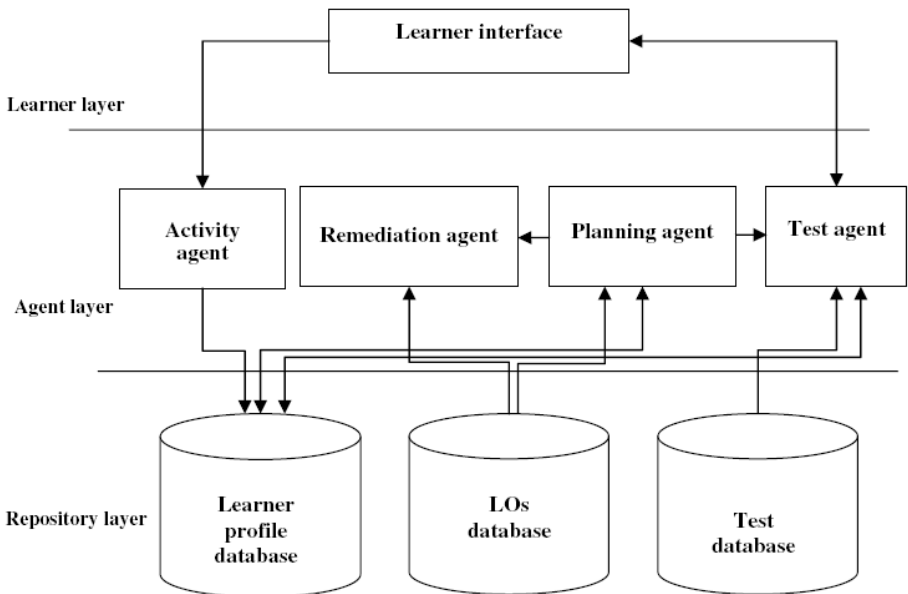


Fig. 2. System architecture proposed in [32]

The agent layer of the architecture contains four different types of agents:

- Activity agent records online learners' learning activities and stores them in the learner's profile.
- Planning agent plans the learning process by providing pre-tests, post-tests and review-tests in conjunction with the test agent and delivers session contents based on the learner's profile and presents the review test's responses to the remediation agent.

- Test agent based on the requests of planning agent, presents appropriate test type to the learner based on his/her ability.
- Remediation agent analyzes the results of review tests and diagnoses learner’s learning problems like a human instructor and then recommends the appropriate learning materials to the learner.

The third and the final study proposes a fully personalization strategy of e-learning scenarios. The two-level architecture of the personalized system capable of providing fully personalization strategies is given in Figure 3 [8]. The second level (ELP2) allows teachers to select personalization parameters and combine them flexibly to define different personalization strategies according to the specifics of courses. The first level (ELP1) is the application of the personalization strategies specified with ELP2. The system is implemented with the web service technology which provides interoperability of the system with other e-learning personalization systems.

This system was experimented with the courses personalized according to the learner’s level of knowledge and the sequential/global dimension of the Felder–Silverman learning style model and the results were promising for future evaluations with new personalization parameter combinations which compose personalization strategies.

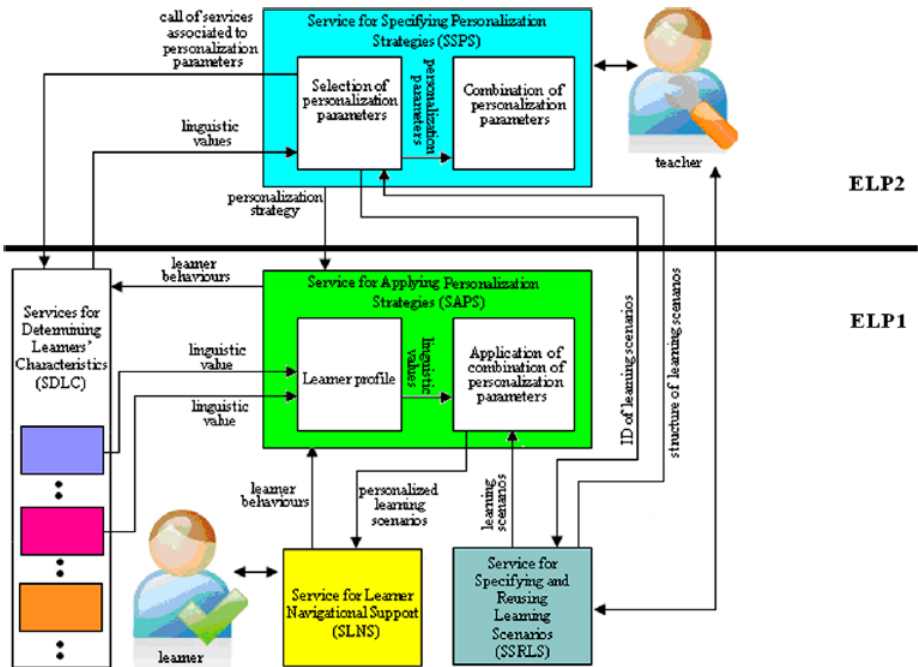


Fig. 3. Architecture of ELP1 + ELP2 for personalization in [8]

5 Conclusion

Some of the challenges of developing adaptive e-learning systems are discussed here as well as the open research areas and possible future works.

First of all, different functionalities of the three examined systems, namely, the personalization strategy determination mechanism of [8], extendibility of LMSs for adaptation in [30], and combinational usage of artificial intelligence techniques like IRT and ANN in MASs by [32] shows that there is a good potential for combining different technologies and getting effective results in the adaptive e-learning field.

Some of the major problems encountered in the area can be ordered as follows: low reusability level of learning objects, inconsistency of pedagogical techniques and low interoperability among different systems [23].

Another concern regarding the student modeling is the static or dynamic modeling dimension. In static modeling, student models are initialized only once, while dynamic student modeling is based on frequent updates with the user system interaction. Learning styles can change with time, thus, it is not accurate to measure it with a questionnaire once and ignore the possible changes.

Considering the user preferences may be misdiagnosed in a system and/or the preferences of the user are changing continuously during the interaction of the user with the system, if the user is unsuccessful at a subject with his/her current preferences, the system should be intelligent enough to offer different learning options (different learning materials according to the changing needs) to the user.

In most of the studies, more than one personalization parameter is used. Therefore, the effect of learning styles on the learning outcomes can not be measured alone, since it is not the only variable in the system.

Even though successful adaptive mechanisms have been developed, the success of these systems is relatively low. In other words, the reflection of these adaptive mechanisms on academic achievement is still relatively weak.

Although there have been a lot of successful studies providing personalization in e-learning environments, a commonly used standard or an effort to support reusability between different systems does not yet exist. Also, since personalization has not been adopted in widely used e-learning systems, this research area provides a huge potential in the educational domain for integration of the existing systems and reusability of the personalization mechanisms.

References

1. Holt, P., Dubs, S., et al.: The state of student modelling. In: *Student Modelling: The Key to Individualized Knowledge-Based Instruction*, pp. 3–35. Springer (1994)
2. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., Nejd, W. (eds.) *Adaptive Web 2007*. LNCS, vol. 4321, pp. 3–53. Springer, Heidelberg (2007)
3. Ardissono, L., Console, L., Torre, I.: An adaptive system for the personalised access to news. *AI Communications* 14, 129–147 (2001)

4. Goren-Bar, D., Graziola, I., Pianesi, F., Zancanaro, M.: The influence of personality factors on visitor attitudes towards adaptivity dimensions for mobile museum guides. *User Modeling and User Adapted Interaction* 16(1), 31–62 (2005)
5. Canales-Cruz, A., Sanchez-Arias, V.G., Cervantes-Perez, F., Peredo-Valderrama, R.: Multi-agent system for the making of intelligence and interactive decisions within the learner's learning process in a web-based education environment. *Journal of Applied Research and Technology* 7(3), 310–322 (2009)
6. Sleeman, D.H.: UMFE: a user modeling front end system. *International Journal on the Man-Machine Studies* 23, 71–88 (1985)
7. Park, O., Lee, J.: Adaptive instructional systems. In: Jonassen, D.H. (ed.) *Handbook of Research for Educational Communications and Technology*, pp. 651–685. Lawrence Erlbaum, Mahwah (2004)
8. Essalmi, F., Jemni Ben Ayed, L., Jemni, M., Kinshuk, Graf, S.: A fully personalization strategy of E-learning scenarios, *Computers in Human Behavior: Emerging and Scripted Roles in Computer-supported Collaborative Learning* 26(4), 581–591 (2010)
9. Kinshuk, Graf, S.: Considering cognitive traits and learning styles to open web-based learning to a larger student community. In: *The First International Conference on ICT & Accessibility*, Hammamet, Tunisia, pp. 21–26 (2007)
10. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. *Engineering Education* 78(7), 674–681 (1988)
11. Honey, P., Mumford, A.: A manual of learning styles. In: Honey, P., Maidenhead (eds.) *Learning Styles*. Engineering Subject Centre (1986)
12. Höök, K., Karlgren, J., Waern, A., Dahlbäck, N., Jansson, C.-G., Karlgren, K., et al.: A glass box approach to adaptive hypermedia. *User Modeling and User-Adapted Interaction* 6(2-3), 157–184 (1996)
13. Milosevic, D., Brkovic, M., Bjekic, D.: Designing lesson content in adaptive learning environments. *International Journal of Emerging Technologies in Learning* 1, 2 (2006)
14. La Granderie, A.: 1993 Les profils pédagogiques. In: Chabchoub, A. (ed.), *Enseigner à l'Université de la théorie à la pratique*. Publications de l'ATURED, Paris (2006)
15. Weber, G., Brusilovsky, P.: ELM-ART: An adaptive versatile system for Web-based instruction. *International Journal of Artificial Intelligence in Education* 12, 351–384 (2001)
16. Chorfi, H., Jemni, M.: PERSO: Towards an adaptive e-learning system. *Journal of Interactive Learning Research* 15, 433–447 (2004)
17. Melis, E., Andrès, E., Büdenbender, J., Frischauf, A., Gogvadze, G., Libbrecht, P., et al.: ActiveMath: A generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education* 12(4), 385–407 (2001)
18. Stash, N., Cristea, A., de Bra, P.: Adaptation to Learning Styles in ELearning: Approach evaluation. In: Reeves, T., Yamashita, S. (eds.) *Proceedings of World Conference on e-Learning in Corporate, Government, Healthcare, and Higher Education*, pp. 284–291. AACE, Chesapeake (2006)
19. Constantino-González, M., Suthers, D., Santos, J.: Coaching web-based collaborative learning based on problem solution differences and participation. *International Journal of Artificial Intelligence in Education* 13, 261–297 (2003)
20. Essalmi, F., Jemni Ben Ayed, L., Jemni, M.: A multi-parameters personalization approach of learning scenarios. In: *The 7th IEEE International Conference on Advanced Learning Technologies*, Niigata, Japan, pp. 90–91 (2007)

21. Kolb, D.A.: Experiential learning: Experience as the source of learning and development. In: Kolb, D.A., Boyatzis, R.E., Mainemelis, C. (eds.) *Experiential Learning Theory: Previous Research and New Directions*. Prentice-Hall, NJ (1984); Sternberg, R.J., Zhang, L.F. (eds.): *Perspectives on cognitive, learning, and thinking styles*. Lawrence Erlbaum, NJ (2000)
22. Kuljis, J., Liu, F.: A comparison of learning style theories on the suitability for elearning. In: Hamza, M.H. (ed.) *Proceedings of the IASTED Conference on Web Technologies, Applications, and Services*, pp. 191–197. ACTA Press, Calgary (2005)
23. Akbulut, Y., Cardak, C.S.: Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. *Computers & Education* 58, 835–842 (2012)
24. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User Profiles for Personalized Information Access. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *Adaptive Web 2007*. LNCS, vol. 4321, pp. 54–89. Springer, Heidelberg (2007)
25. Fischer, G.: User modeling in human-computer interaction. *User Modeling and User Adapted Interaction* 11(1-2), 65–86 (2001)
26. Goldstein, I.P.: The genetic graph: a representation for the evolution of procedural knowledge. In: Sleeman, D.H., Brown, J.S. (eds.) *Intelligent Tutoring Systems*, pp. 51–77. Academic Press, London (1982)
27. Van Lehn, K.: Student models. In: Polson, M.C., Richardson, J.J. (eds.) *Foundations of Intelligent Tutoring Systems*, pp. 55–78. Lawrence Erlbaum Associates, Hillsdale (1988)
28. Ciloglulil, B., Inceoglu, M.M.: Exploring the State of the Art in Adaptive Distributed Learning Environments. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B.O. (eds.) *ICCSA 2010, Part II*. LNCS, vol. 6017, pp. 556–569. Springer, Heidelberg (2010)
29. Dolog, P., Henze, N., Nejdl, W., Sintek, M.: Personalization in distributed elearning environments, pp. 170–179. ACM (2005)
30. Graf, S., Kinshuk, Ives, C.: A Flexible Mechanism for Providing Adaptivity Based on Learning Styles in Learning Management Systems. In: *Proceedings of the 2010 10th IEEE International Conference on Advanced Learning Technologies ICALT, July 05-07 (2010)*
31. Felder, R.M., Soloman, B.A.: Index of Learning Styles questionnaire (1997), <http://www.engr.ncsu.edu/learningstyles/ilsweb.html> (retrieved October 25, 2011)
32. Baylari, A., Montazer, G.A.: Design a personalized e-learning system based on item response theory and artificial neural network approach. *Expert Systems with Applications* 36(4), 8013–8021 (2009)

An Experimental Study of the Combination of Meta-Learning with Particle Swarm Algorithms for SVM Parameter Selection

Péricles B.C. de Miranda¹, Ricardo B.C. Prudêncio¹,
Andre Carlos P.L.F. de Carvalho², and Carlos Soares³

¹ Federal University of Pernambuco, CIn-UFPE

² University of São Paulo, USP

³ University of Porto, FEP

Abstract. Support Vector Machines (SVMs) have become a well succeed algorithm due to the good performance it achieves on different learning problems. However, to perform well the SVM formulation requires adjustments on its model. Avoiding the trial and error procedure, the automatic SVM parameter selection is a way to deal with this. The automatic parameter selection is commonly considered an optimization problem whose goal is to find suitable configuration of parameters which attends some learning problem.

In the current work, we propose a study of the combination of Meta-learning (ML) with Particle Swarm Optimization (PSO) algorithms to optimize the SVM model, seeking for combinations of parameters which maximize the success rate of SVM. ML is used to recommend SVM parameters, to a given input problem, based on well-succeeded parameters adopted in previous similar problems. In this combination, initial solutions provided by ML are possibly located in good regions in the search space. Hence, using a reduced number of candidate search points, in the search process, to find an adequate solution, would be less expensive.

In our work, we implemented five benchmarks PSO approaches applied to select two SVM parameters for classification. The experiments consist in comparing the performance of the search algorithms using a traditional random initialization and using ML suggestions as initial population. This research analysed the influence of meta-learning on convergence of the optimization algorithms, verifying that the combination of PSO techniques with ML obtained solutions with higher quality on a set of 40 classification problems.

Keywords: particle swarm optimization, meta-learning, svm parameter selection.

1 Introduction

SVMs have achieved a considerable attention due to its theoretical foundations and good empirical performance when compared to other learning algorithms in different applications [1]. However, the SVM performance strongly depends on

the adequate choice of its parameters including, for instance, the kernel function, the values of kernel parameters, the regularization parameter, among others [2]. An exhaustive trial-and-error procedure for selecting good values of parameters is obviously not practical [3].

The process of selecting SVM parameters is commonly treated by different authors as an optimization problem in which a search algorithm is used to find the adequate configurations of parameters on the problem at hand [4]. Although it represents an automatic mode to select SVM parameters, this approach can still be very expensive, since a large number of candidate configurations of parameters is often evaluated during the search process [1].

An alternative approach to SVM parameter selection is the use of Meta-Learning (ML), which treats the SVM parameter selection as a supervised learning task [1] [5]. Each training example for ML (i.e. each meta-example) stores the characteristics of a past problem and performance obtained by a set of candidate configurations of parameters on the problem. By receiving a set of such meta-examples as input, a meta-learner is able to predict the most suitable configuration of parameters for a new problem based on its characteristics. ML is a less expensive solution compared to the search approach. In fact, once the knowledge is acquired by the meta-learner, configurations of parameters can be suggested for new problems without the need of empirically evaluating different candidate configurations (as performed using search techniques). However, ML is very dependent on the quality of its meta-examples. In the literature, it is usually difficult obtaining good results since meta-features are in general very noisy and the number of problems available for meta-example generation is commonly limited. Hence, the performance of ML for SVM parameter selection may be not so good as the performance of search techniques [6].

Visualizing these drawbacks, a recent work, developed by [18], combined search techniques and ML to solve the SVM parameter selection problem for regression. In this proposal, configurations of parameters suggested by ML are adopted as initial solutions which will be later refined by the search technique. This work used as search technique the single objective version of PSO whose objective was to minimize the error rate. The results showed that the combination of ML with PSO generated better solutions in comparison to PSO with random initialization for SVM parameter selection in the cited problem. This work implemented a prototype of PSO, using the *inertia* weight and adopted *Star* as topology, and compared the PSO convergence with the convergence of PSO using ML, verifying that the suggestions of ML speed up and refine the algorithm's convergence.

Aiming to realize a more complete study of the influence of ML suggestions in the optimization process of search algorithms, we implemented five benchmarks PSO approaches for comparison: Basic PSO, Inertia Gbest (PSO using *inertia* weight and *Gbest* topology), Constricted Gbest (PSO using *constricted* factor and *Gbest* topology), Inertia Lbest (PSO using *inertia* weight and *Lbest* topology) and Constricted Lbest (PSO using *constricted* factor and *Lbest* topology). In order to evaluate our study, we adapted the prototypes to select two SVM

parameters: the parameter γ of the RBF kernel and the regularization C , which may have a strong influence in SVM performance [10].

In our work, a database of 40 meta-examples was produced from the evaluation of a set of 399 configurations of (γ, C) on 40 different classification problems. Each classification problem was described by a number of 8 meta-features proposed in [11] [1] [16]. All the implemented prototypes were used to optimize the parameters (γ, C) regarding the success rate on classification.

In our experiments, we, initially, performed an analysis comparing the performance of the benchmark algorithms using random initialization. After that, we realized a comparison between the benchmark algorithm which achieved the best performance with the same algorithm using ML (hybrid approach). The experiments' results revealed that the hybrid approach was able to generate better solutions along the generations when compared to the randomly initialized PSOs.

The paper is organized as follows: Section 2 brings a brief presentation on the SVM model selection. Section 3 presents details of the proposed work. Section 4 describes the experiments, obtained results and statistical analysis. Finally, Section 5 presents some conclusions and the future work.

2 SVM Parameter Selection

The SVM parameter selection task is often performed by evaluating a range of different combinations of parameters and retaining the best one in terms of performance [12]. Different authors have deployed search and optimization techniques aiming to automatize this process and to avoid an exhaustive or a random exploration of parameters [14] [4] [7] [8] [9] [12] [13].

In this context, solutions are combinations of parameters and the objective function corresponds to SVM execution. Different techniques were proposed as based on gradients [14], evolutionary algorithms [7] [8] [9], Tabu Search [12], and PSO [13] [18].

Although the use of search techniques automatize the parameter selection process, this solution may still be very expensive since for each configuration being evaluated during the search it is necessary to train the SVM [1]. The impact of this limitation can be even more drastic depending on the problem at hand and the number of parameters to be optimized.

ML is an alternative that has been studied in recent years to SVM parameter selection [1] [15] [11] [16] [17] [5]. In this approach, the choice of parameters for a problem is based on well-succeeded parameters adopted to previous similar problems. In ML, it is necessary to maintain a set of meta-examples where each meta-example stores: (1) a set of features (called meta-features) describing a learning problem; and (2) evaluations of a set of candidate parameters on the problem. A meta-learner is then used to acquire knowledge from a set of such meta-examples in order to recommend (or predict) the adequate configurations of parameters for new problems based on past problems' characteristics.

In this way, using ML, SVM parameters can be suggested for new problems without executing the SVM on each candidate configuration of parameters

making this approach more economic in terms of computational cost. However, ML is very dependent on the quality of its meta-examples. In the literature, it is usually difficult obtaining good results since meta-features are in general very noisy and the number of problems available for meta-example generation is commonly limited. Hence, the performance of ML for SVM parameter selection may be not so good as the performance of search techniques [6].

Thus, a recent work was performed combining ML with particle swarm optimization algorithms [18] in such a way that ML is used to recommend parameters which will be later refined by a search algorithm. This research handled the SVM parameter selection task as a single objective problem where the hybrid approach achieved good results for regression problems.

As it will be seen, in this study we propose a more complete analysis of the ML application to recommend parameters which will be later refined by search approaches applied for classification problems.

3 Developed Work

The work presented here proposes a comparison of benchmarks search techniques and hybrid search techniques (search techniques combined with ML) aiming to analyse the influence of suggestion of solutions in the search process. As context, we adopted the SVM model selection problem for classification.

Figure 1 depicts the general architecture [18] used to perform the combination of search techniques with ML. Initially, the Meta-Learner module retrieves a predefined number of past meta-examples stored in a Database (DB), selected on the basis of their similarity to the input problem. Following, the Search module adopts as initial search points the configurations of parameters which were well-succeeded on the retrieved meta-examples. The Search module iterates its search process by generating new candidate configurations to be evaluated in the SVM. The output configuration of parameters will be the best one generated by the Search module up to its convergence or another stopping criteria.

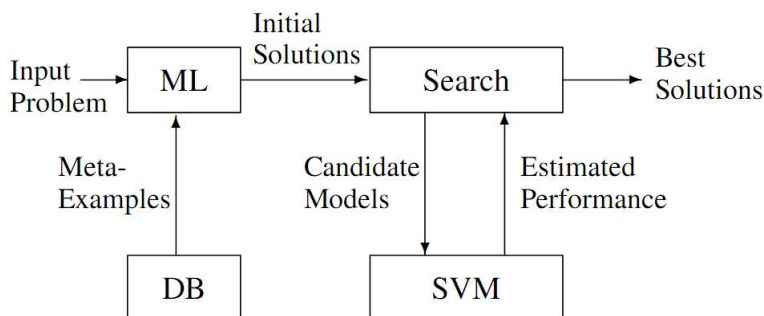


Fig. 1. Proposed Architecture

In the current work, we implemented five benchmark particle swarm optimization algorithms to select two specific SVM parameters: the γ parameter of RBF kernel and the regularization parameter C . The choice of RBF kernel is due to its flexibility in different problems compared to other kernels [19] [20]. It is known that the γ parameter has an important influence in learning performance since it controls the linearity of the induced SVM. The parameter C is also important for learning performance since it controls the complexity of the induced SVMs [20]. As it will be seen, the prototypes were implemented to select the parameters (γ , C) for classification problems according one objective: success rate. Details of implementation will be presented in the next subsections.

3.1 Search Module

In our study, we implemented five benchmark search algorithms based on PSO: basic PSO, Inertia *Gbest*, Inertia *Lbest*, Constricted *Gbest* and Constricted *Lbest*. This subsection presents, initially, how the basic PSO works and after that we present its variations.

The Main Forms of PSO. In our basic PSO implementation, each particle i represents a solution for a given problem, indicating the position of the particle in the search space. Each particle also has a velocity which indicates the current search direction performed by the particle. PSO basically works by updating the position and velocity of each particle in order to progressively explore the best regions in the search space. The update of position and velocity in the basic PSO is given by the following equations:

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) + c_1 r_1 (\mathbf{p}_i(t) - \mathbf{x}_i(t)) + c_2 r_2 (\mathbf{n}_i(t) - \mathbf{x}_i(t)), \quad (1)$$

$$\mathbf{x}_i(t+1) = \mathbf{x}_i + \mathbf{v}_i(t+1). \quad (2)$$

In equation (1), $\mathbf{p}_i(t)$ is the best position achieved by the particle so far, and $\mathbf{n}_i(t)$ is the best position achieved by any particle in the population so far. Hence, each particle is progressively moved in direction of the best *global* positions achieved by the population (the social component of the search) and the best local positions obtained by the particle (the *cognitive* component of the search). The parameters c_1 and c_2 are positive accelerators which control the trade-off between exploring good global regions in the search space and refining the search in local regions around the particle. In equation (1), r_1 and r_2 are random numbers used to enhance the diversity of particle positions.

As it was mentioned, two parameters (c_1 and c_2) accelerate the convergence process, however, this acceleration can lead the swarm to an explosion state, where the particles achieve high velocity values [21]. Thus, a few years after the initial PSO publications, a new parameter ω , called *inertia* weight, was

introduced in an effort to strike a better balance. The velocity update equation was altered to the form:

$$\mathbf{v}_i(t+1) = \omega\mathbf{v}_i(t) + c_1r_1(\mathbf{p}_i(t) - \mathbf{x}_i(t)) + c_2r_2(\mathbf{n}_i(t) - \mathbf{x}_i(t)). \tag{3}$$

By adjusting the value of ω , the swarm has a greater tendency to eventually constrict itself down to the area containing the best fitness and explore that area in detail.

Another method of balancing global and local searches known as *constriction* was being explored simultaneously with the *inertia* weight method. Similar to the *inertia* weight method, this method introduced a new parameter χ , known as the *constriction* factor. χ is derived from the existing constants in the velocity update equation:

$$\chi = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}, \varphi = c_1 + c_2. \tag{4}$$

It was found that when $\varphi < 4$, the swarm would slowly "spiral" toward and around the best found solution in the search space with no guarantee of convergence, while for $\varphi > 4$ convergence would be quick and guaranteed.

This *constriction* factor is applied to the entire velocity update equation:

$$\mathbf{v}_i(t+1) = \chi(\mathbf{v}_i(t) + c_1r_1(\mathbf{p}_i(t) - \mathbf{x}_i(t)) + c_2r_2(\mathbf{n}_i(t) - \mathbf{x}_i(t))). \tag{5}$$

The effects are similar to those of *inertia* weight, resulting in swarm behaviour that is eventually limited to a small area of the feasible search space containing the best known solution. In our prototypes, we fixed PSO parameters using c_1 and $c_2 = 2.05$ and the $\omega = 0.8$.

Besides changes in the velocity structure the way the information is exchanged among the particles is crucial for the convergence. To disseminate information within a swarm is the key of any swarm intelligence based algorithm. PSO, like others swarm algorithms, make use of its own information exchange methods to distribute the best positions found during the algorithm execution [21] [22]. The way used by the swarm to distribute this information is the social structure formed by the particles. Variations in this structure can improve the algorithm performance.

Even when using different types of velocity update equations, the algorithm can work better by exploring the information exchange mechanism inside the swarm. This information exchange influences the particles in the velocity evaluation. The most common communication mechanisms between particles are *Star* and *Ring* topologies, shown in Figure 2.

Particles can share information globally through a fully-connected structure called star topology represented by Figure 2 a. This topology uses a global

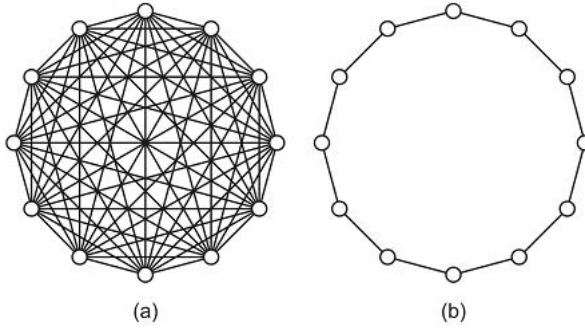


Fig. 2. Star and Ring topology

neighbourhood mechanism known as *Gbest* to share information. By using *Gbest*, particles can spread information quickly through the swarm. It is analogous to a large community where all decisions taken are instantaneously known by everyone. Each particle in this topology is attracted towards the best solution found by the entire swarm.

An information exchange mechanism based on local neighbourhood is known as *Lbest*. The particles only share information with their own neighbours. The structure used by this mechanism is ring topology and is illustrated in Figure 2b. Different regions of the search space can be explored at the same time; however, the successful information from these regions takes a longer time to be sent to all other particles.

PSO to SVM Model Selection. In the basic PSO and its variations, we adapted here to perform the search for configurations (γ, C) . The objective function evaluates the success rate (SR) of each configuration of parameters, trying to maximize it, on a given classification problem.

In our work, the implemented search techniques performed a search in a space represented by a discrete grid of SVM configurations, consisting of 399 different settings of parameters γ and C . By following the guidelines provided in [20], we considered the following exponentially growing sequences of γ and C as potentially good configurations: the parameter γ assumed 19 different values (from 2^{-15} to 2^3) and the parameter C assumed 21 different values (from 2^{-5} to 2^{15}), thus yielding $19 \times 21 = 399$ different combinations of parameters in the search space.

3.2 Meta-Database.

In order to generate meta-examples, we collected 40 datasets corresponding to 40 different classification problems, available in the WEKA project [23] and UCI Repository [24]. Each meta-example is related to a single classification problem and stores: (1) a vector of meta-features describing the problem; and (2) the

performance grid which stores the success rate obtained by the SVM in the search space of configurations (γ, C) . The performance grid consists of 399 different settings of parameters γ and C .

Meta-Features. In the developed work, a total number of 8 meta-features was used to describe the datasets of classification problems. These meta-features were based on the set of features defined in [25], which proposed 16 meta-features for classification problems. However, according the restrictions (numerical, no missing values, no binary attributes) of the selected databases, we adopted the meta-features listed in Table 1, which is divided in 3 parts: Simple, Statistical and Information Theory meta-features.

Table 1. Meta-Features for Classification Problems

Simple
Number of examples
Number of attributes
Number of classes
Statistical
Mean correlation of attributes
Skewness
Kurtosis
Geometric mean of attributes
Information Theory
Entropy of class

Values as *number of examples*, *attributes* and *classes* are data already discriminated in databases, being considered simple data. The group of statistical values is composed by the *mean correlation of attributes*; *Skewness*, which measure the asymmetry of the distribution regarding the central axis [25]; *Kurtosis*, which measure the dispersion (characterized by the flatness of the distribution curve) [25] and the *geometric mean of the attributes* which evaluates the mean of the data standard deviation. Ultimately, the information theory group measure the randomness of the instances; being composed by the *Entropy* which defines the degree of uncertainty of classification [25] [26].

Performance Grid. The performance grid stores the success rate obtained by the SVM on a problem considering different SVM configurations. For each of the 399 configurations, a 10-fold cross validation experiment was performed to collect SVM performance. The obtained 399 objective values were stored in the performance grid. In these experiments, we deployed the Scikits Learn library [20] to implement the SVMs and to perform the cross-validation experiments.

We highlight here that the performance grid is equivalent to the search space explored by PSO. By generating a performance grid for a problem, we can evaluate which configurations of parameters were the best ones in the problem (i.e., the best points in a search space) and we can use this information to guide the search process for new similar problems.

3.3 Meta-Learner

Given a new input problem described by the vector $\mathbf{x} = (x_1, \dots, x_p)$, the Meta-Learner selects the k most similar problems according to the distance between the meta-attributes. We applied this method to make the initial population more diverse (retrieving good solutions from different similar problems) to avoid suggesting local maximum regions. The distance function (*dist*) implemented was the Euclidean Distance, defined as:

$$\text{dist}(\text{vec}_i, \text{vec}_l) = \sum_{j=1}^p \sqrt{(i_j - l_j)^2}. \quad (6)$$

For each retrieved meta-example, the meta-learner selects in the performance grid the configuration of parameters (among the 399 candidates) that obtained the highest SR value, i.e. the best SVM configuration. Hence, the meta-learner will suggest as initial PSO population the set of k best configurations selected in the performance grids of the retrieved meta-examples.

4 Experiments

In this work, we realized two experiments: 1) comparison of the benchmark algorithms' performance and 2) comparison between the benchmark algorithm which achieved the best performance with the hybrid algorithm.

In the first experiment, we executed each one of the five benchmark approaches using random initialization, computing the mean of the success rate values for each generation of all problems. The output of this experiment is the mean curve, among all problems, representing the performance along the generations.

In the second experiment, to realize the comparison between the best benchmark algorithm with the hybrid algorithm, it was necessary to execute the hybrid technique following a leave-one-out methodology described as follows.

At each step of leave-one-out, one meta-example was left out to evaluate the implemented prototype and the remaining 39 meta-examples were considered in the DB to be selected by the ML module. Initially, a number of k configurations were suggested by the ML module as the initial PSO population (in our experiments, we adopted $k = 7$). The PSO then optimized the SVM configurations for the problem left out up to the number of 10 generations. In each generation, we recorded the highest SR value obtained so far (i.e. the best fitness). Hence, for each problem left out a curve of N values of success rate was generated aiming to analyse the search progress on the problem. Finally, the curves of SR values

were averaged over the 40 steps of the leave-one-out experiment in order to evaluate the quality of the PSO search on optimizing SVM parameters for the 40 classification problems considered.

After executing both, best benchmark algorithm and hybrid algorithm, we analysed their performance along the generations and, moreover, evaluated the number of wins of the algorithms per generation regarding all problems, where the algorithm which achieved better performance, according an specific problem, is the winner.

Finally, we highlight that each evaluated version of PSO was executed 30 times and the average results were recorded.

4.1 Results

Initially, we analysed the success rate curve of all PSO approaches using random initialization, and selected the technique which achieved the best performance to be used as basis of comparison with its own version, but, using ML suggestions as initial population.

Figure 3 shows the mean performance curve of all benchmark search techniques along the generations. As we can see, the basic PSO performance, after the fourth generation, outperformed the other techniques. As we adopt the policy of selecting the k most similar problems to augment the diversity, the basic PSO exploration converged faster with 10 generations. For the same reason, the other techniques which used the *Gbest* mechanism achieved better results than the approaches that used *Lbest* mechanism. The approaches using *Ring* topology were overcome by the other techniques since its policy of information exchange is based on exploitation, reducing the velocity of convergence.

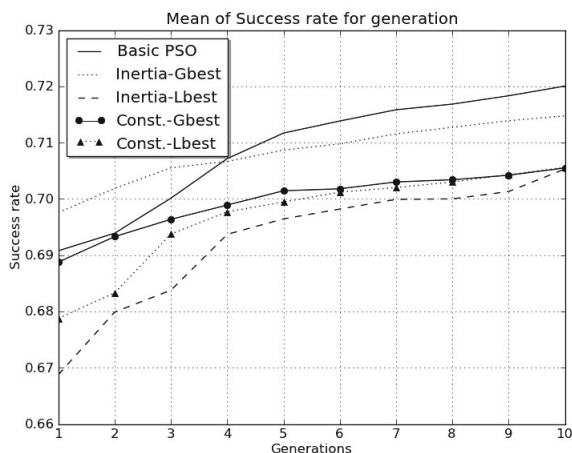


Fig. 3. Success rate curve of all PSO approaches using random initialization with $k = 7$

As it was discussed above, the approach which obtained the best results was the basic PSO. So, in order to analyse the influence of ML on search approaches performance, we compared the performance of the basic PSO with basic PSO using ML (hybrid PSO). Figure 4 makes a comparison between the basic PSO and the basic PSO using ML performance. As it can be seen, the combination of the basic PSO with ML makes the search process start in well succeed regions, for this reason in the first iteration the Hybrid PSO presented a higher performance than basic PSO. Along the generations, while the basic PSO sought by good regions, the hybrid PSO refined the found solutions augmenting the convergence.

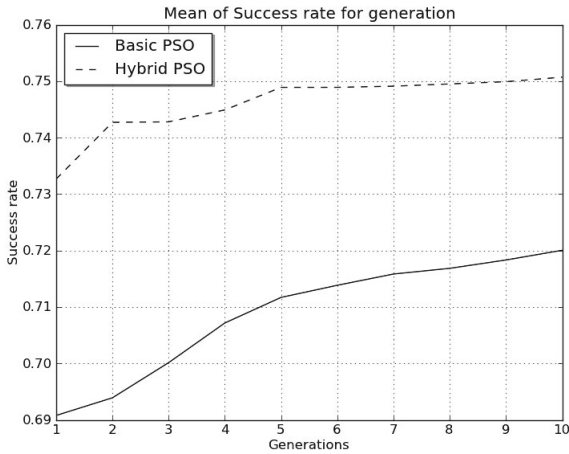


Fig. 4. Success rate curve of the basic PSO and the hybrid PSO with $k = 7$

The analysis realized above was about the mean of the success rate values for each generation of all problems. The next analysis intend to count the number of classification problems that each algorithm won per iteration, shown in Figure 5.

As it can be seen, the hybrid PSO won the basic PSO in most of the classification problems in all generations. It achieved a mean of victories of 78.5% considering all generations. Possibly, the problems the hybrid approach lost for the basic PSO are classification problems with few or none similar problems. So, the suggestion of solutions performed by the ML can be impaired impacting negatively in the search process.

Although, the hybrid approach has achieved clearly better results than the basic PSO, a visual representation is not enough to conclude which approach is considered better. In this way, we realized statistical tests to guarantee our experiments.

4.2 Statistical Analysis

Besides the analysis presented in last section, we realized two statistical tests: normality and hypothesis test; in order to evaluate which distribution,

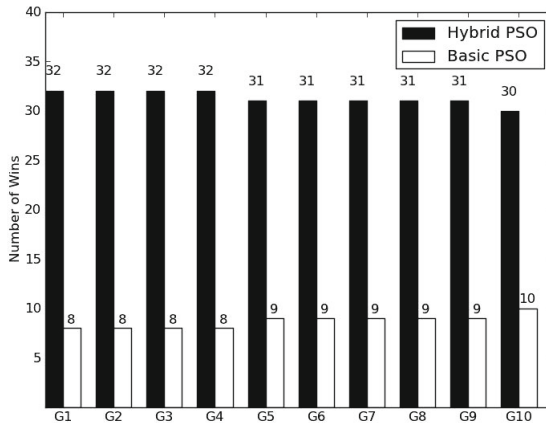


Fig. 5. Number of wins regarding success rate using $k = 7$ for 10 generations

presented by the basic and hybrid PSO, is considered better. The Anderson-Darling method was used to test the data's normality using 5% of significance. After verifying and assuring the non-normality of the data, we applied the Wilcoxon test, a non-parametric test, to verify the hypothesis.

To perform this test, the hybrid PSO sample is presented as μ_1 and basic PSO sample as μ_2 , using the following hypothesis:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_a &: \mu_1 > \mu_2. \end{aligned}$$

The null hypothesis is that both distributions are equal, represented by H_0 ; and the alternative one is that the results of μ_1 are better than μ_2 , represented by H_a ; using 5% of significance. If the p -value assumes a value less than 5%, the null hypothesis is rejected and the alternative one accepted.

After applying the Wilcoxon test, it returned p -value = 0.00578. As the p -value is less than 5%, the null hypothesis is rejected. Hence, we conclude μ_1 (hybrid PSO results) achieved better solutions, per iteration, in comparison to μ_2 (basic PSO results), with 95% of reliability.

5 Conclusions

In the current work, we analysed the influence of meta-learning in the search process of particle swarm algorithms for the problem of SVM parameter selection. To perform the experimental study, we implemented the main particle swarm approaches to select the parameter γ of the RBF kernel and the regularization parameter C . In our implementation, a number of 40 classification problems was used to generate meta-examples. In the performed experiments, we verified that the combination of meta-learning with particle swarm algorithms overcame all the benchmark results with 95% of reliability.

In future work, we intend to augment the number of meta-examples as we believe that the performance of the hybrid approaches can be improved as more meta-examples are considered. Also, other variations of search techniques can be considered in the future implementations.

References

1. Soares, C., Brazdil, P., Kuba, P.: A meta-learning approach to select the kernel width in support vector regression (4), 195–209 (2000)
2. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press (2000)
3. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. *Machine Learning*, 131–159 (2002)
4. Cristianini, N., Campbell, C., Shawe-Taylor, J.: Dynamically adapting kernels in support vector machines. In: *NIPS*, pp. 204–210 (1998)
5. Ali, S., Smith-Miles, K.: On optimal degree selection for polynomial kernel with support vector machines: Theoretical and empirical investigations. *KES Journal* 1(1), 1–18 (2007)
6. Narzisi, G.: An Experimental Multi-Objective Study of the SVM Model Selection problem
7. Lessmann, S., Stahlbock, R., Crone, S.: Genetic algorithms for support vector machine model selection. In: *International Joint Conference on Neural Networks*, pp. 3063–3069 (2006)
8. Friedrichs, F., Igel, C.: Evolutionary tuning of multiple svm parameters. *Neurocomputing* (2005)
9. Lorena, A., de Carvalho, A.: Evolutionary tuning of svm parameter values in multiclass problems. *Neurocomputing*, 16–18
10. Keerthi, S.S.: Efficient tuning of svm hyperparameters using radius/margin bound and iterative algorithms. *IEEE Transactions on Neural Networks* (2002)
11. Kuba, P., Brazdil, P., Soares, C., Woznica, A.: Exploiting sampling and meta-learning for parameter setting support vector machines. In: *Proceedings of the IBERAMIA*, pp. 217–225 (2001)
12. Cawley, G.: Model selection for support vector machines via adaptive step-size tabu search. In: *Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms*, vol. 28(8), pp. 434–437 (2001)
13. de Souza, B., de Carvalho, A., Ishii, R.: Multiclass svm model selection using particle swarm optimization. In: *Sixth International Conference on Hybrid Intelligent Systems* (2006)
14. Glasmachers, T., Igel, C.: Gradient-based adaptation of general gaussian kernels. *Neural Comput.* (2005)
15. Ali, S., Smith-Miles, K.A.: A meta-learning approach to automatic kernel selection for support vector machines. *Neurocomputing*, 173–186 (2006)
16. Soares, C., Brazdil, P.: Selecting parameters of svm using meta-learning and kernel matrix-based meta-features. In: *SAC* (2006)
17. Ali, S., Smith, K.A.: Matching svm kernel's suitability to data characteristics using tree by fuzzy c-means clustering. In: *Third International Conference on Hybrid Intelligent Systems* (2010)
18. Gomes, T., Prudencio, R.B.C., Soares, C., Rossi, A., Carvalho, A.: Combining meta-learning and search techniques to svm parameter selection. In: *Brazilian Symposium on Neural Networks*, pp. 79–84 (2010)

19. Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with gaussian kernel. In: Imielinski, T., Korth, H. (eds.) *Mobile Computing*, pp. 153–181. Kluwer Academic Publishers (1996)
20. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: *Scikit-Learn: A Toolkit of Machine Learning in Python*, <http://scikit-learn.org>
21. Engelbrecht, A.P.: *Fundamentals of Computational Swarm Intelligence*. John Wiley & Sons (2005)
22. Sutton, A.M., Whitley, D., Lunacek, M., Howe, A.: PSO and multi-funnel landscapes: how cooperation might limit exploration. In: *GECCO 2006: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, July 8-12, vol. 1, pp. 75–82. ACM Press, Seattle (2006)
23. WEKA, The University of Waikato, <http://www.cs.waikato.ac.nz/ml/weka/>
24. UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml/>
25. Smith-Miles, K. A.: *Cross-Disciplinary Perspectives on Meta-Learning for Algorithm Selection* (2005)
26. Prudêncio, R.B.C., Ludermir, T.B.: Selective generation of training examples in active meta-learning. *International Journal of Hybrid Intelligent Systems* (2008)

An Investigation into Agile Methods in Embedded Systems Development

Caroline Oliveira Albuquerque¹, Pablo Oliveira Antonino¹,
and Elisa Yumi Nakagawa²

¹ Fraunhofer Institute for Experimental Software Engineering
Fraunhofer-Platz 1, 67663, Kaiserslautern, Germany
{caroline.albuquerque,pablo.antonino}@iese.fraunhofer.de

² Dept. of Computer Systems, University of São Paulo - USP
PO Box 668, 13560-970, São Carlos, SP, Brazil
elisa@icmc.usp.br

Abstract. Embedded systems are widely used in diverse areas, such as avionics, consumer electronics, and medical equipments, causing a considerable impact on modern society. Since these systems sometimes deal directly with human lives, and require a considerable level of quality, their development should be subject to a rigorous process. In another perspective, agile methods (or agile processes) have been adopted by the software industry as a lightweight, iterative, and collaborative approach for developing software systems. Although agile methods do not seem to be suitable to embedded systems, they have been successfully used for building such systems. However, there exists no detailed and analytical overview of the use of such methods in the embedded systems domain. The main objective of this paper is to present a detailed view of how agile methods have been used in the development of embedded systems, and to describe their benefits, challenges, and limitations. For this, we have applied Systematic Review, a technique for systematically exploring, organizing, summarizing, and assessing potentially all works conducted in a specific research area. As the main result, we have observed that agile methods have brought advantages to embedded systems development; however, more studies should be conducted. Furthermore, this work is also intended to contribute to the identification of important new research lines.

1 Introduction

Embedded systems refer to computational systems designed to perform dedicated functions, sometimes as part of a complete device often including hardware and mechanical parts [1]. They play an important role all over modern society; for instance, in automobiles, aircrafts, mobile devices, and devices with a lot of built-in intelligence. Therefore, it is observed that the demand for embedded systems has increased dramatically. In parallel, they have become more and more sophisticated and complex, and are increasingly being used in situations where human lives are directly involved. In the face of these facts, considerable attention must be given to their development process in order to provide a high level

of quality, e.g., in terms of efficiency, reliability, and safety, which are quality attributes strongly demanded by such systems.

In parallel, different software processes have been proposed to organize software development activities. At one extreme, there are industry-tested processes aimed at more effective project management, such as RUP¹ (Rational Unified Process) and V-Model². Due to their rigorousness, they have been explored in the development of quality-centered software systems, such as embedded systems. At the other extreme, there are agile methods (also called agile processes), such as XP³ (eXtreme Programming) and Scrum⁴, which have also attracted attention from the software industry, since by adopting these methods, industry has been able to accelerate the time to market of their products [2]. In particular, XP presents a set of activities, values, principles, and agile practices; these practices are derived from the best practices of software engineering; therefore, they have been widely used when adopting XP as a development process. In general, agile methods aim at achieving reduced time to market by simplifying the release scopes with the intent to reduce effort for complexity and management. They also aim at keeping the customer continuously satisfied and present during the software development [3].

Due to the relevance and dissemination of agile methods and, although these methods do not seem to be suitable to the development of embedded systems at first glance, we can find important initiatives aimed at exploring these methods for use in the development of these systems. However, there is a lack of a detailed overview that could support a good understanding of why, how, when, and which agile methods have currently been adopted for the development of embedded systems. This overview could also be the basis from which new, essential research lines on this topic could be identified, which could contribute to improve the way these systems are developed. In this context, the main objective of this paper is to present a detailed overview of agile methods used for the development of embedded systems. For this, we conducted a systematic review [4], a technique that originates from Evidence Based Software Engineering (EBSE) and makes it possible to explore, organize, summarize, and assess all contributions of the current state of a research area. As the results of our systematic review, we can point out the benefits that can be gained by exploring agile methods within the context of embedded systems. However, we have observed that agile methods and embedded systems should be further investigated together, with the goal to explore their advantages in a coordinated manner. Moreover, this work is intended to be used to identify interesting perspectives for future research in this area.

This paper is organized as follows. In Section 2, we present the background on agile methods, embedded systems, and systematic review. In Section 3, we present the systematic review we conducted. In Section 4, we discuss results,

¹ <http://www-01.ibm.com/software/awdtools/rup/>

² <http://www.v-modell.iabg.de/>

³ <http://www.extremeprogramming.org/>

⁴ <http://www.scrumalliance.org/>

lessons learned, and limitations of this work. Conclusions and future directions are summarized in Section 5.

2 Background

According to Kamal [5], an embedded system is composed of dedicated-purpose software embedded in computer hardware, i.e., it is a computer-based system for a specific application or product, performing some specific tasks. It may also be an independent system or part of a larger system. For Wilmshurst [6], an embedded system can be defined as a system whose main function is not computational, but which is controlled by a computer (likely a microprocessor or microcontroller). Thus, automobiles, industrial machinery, medical equipment, airplanes, toys, and mobile devices are examples of possible hosts of embedded systems. Due to the increasing use, dissemination, and relevance of embedded systems, companies need to increase their productivity and reduce time to market. At the same time, since these systems are sometimes used in critical environments, their development process must be rigorous in order to insure the required quality attributes. It is worth highlighting that acceleration of time to market, increase of productivity, and the arise of quality are also concerns shared with agile methods [2].

The first association of a software process with the word *agile* was used in 1998 [7], although many claim that agile methods themselves existed long before they were formalized. Later in 2001, a group of practitioners of those methods gathered and summarized their experience in what was named *The Agile Manifesto* [8]. In short, its main principles are [3]: individuals and interactions over processes and tools; working software over comprehensive documentation; customer collaboration over contract negotiation; and responding to change over following a plan. However, it is not possible to exactly define agile methods, since specific practices vary with the different methods. In spite of that, short-time boxed iterations with adaptive, evolutionary refinement of plans and goals is a basic practice that these methods share [3,8]. Although this apparently contradicts the agile method principles, the use of these methods in the development of embedded systems has also been explored [9,10,11]. The aim of these experiments has been to explore and adapt such methods in order to build embedded systems more effectively. Important, but isolated contributions, can be found. There is also a lack of a more analytical and detailed work putting together these contributions and presenting an overview of this promising research area.

In another perspective, it has been noticed that as a research area matures, there is almost always an increase in the number of reports and results made available. During the study of a new knowledge area, researchers usually conduct a bibliographical review (almost always an informal review) to identify publications related to a specific subject. However, this kind of review does not use a systematic approach and does not offer any support to avoid bias during the selection of the publications. Thus, the use of mechanisms for summarizing

⁵ <http://agilemanifesto.org>

and providing an overview of an area or topic of interest becomes important. For this, EBSE has investigated and proposed the use of the Systematic Review technique [4]. In this context, an individual evidence (for instance, a case study or an experimental study reported in a publication/paper) which contributes to a systematic review is called a primary study, while the result of a systematic review is a secondary study. Systematic review aims at providing an overview of a research area to assess the quantity, quality, and type of primary studies existing on a topic of interest. In short, a systematic review is conducted via planning, conducting the search, and screening primary studies using inclusion and exclusion criteria [4]. Besides that, a systematic review also performs data extraction and conducts a quantitative and qualitative analysis of the primary study. Therefore, the use of the systematic review technique appears adequate for achieving the goal of this work.

3 Systematic Review Conducted

The main objective of our systematic review was to identify all primary studies that explore agile methods in the context of embedded system development. This systematic review was conducted from August 2011 to September 2011 and involved three people (one researcher in software engineering and embedded systems, one specialist in systematic review, and one graduate student). In order to conduct the systematic review, we used the process presented in Figure 1. In short, it is composed of three steps [4]: (i) Planning of the systematic review; (ii) Conduction of the search and data extraction; (iii) Reporting of the results of the qualitative and quantitative analysis. These steps are explained in more detail during the presentation of our systematic review. Below, we describe each step in detail.

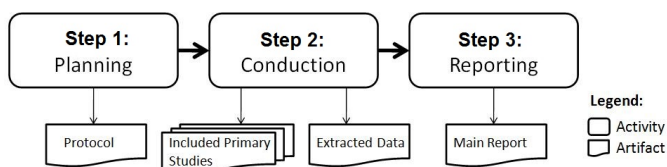


Fig. 1. Systematic review process (Adapted from [4])

3.1 Step 1: Planning

In this phase, we established the systematic review plan. For this, we specified: (i) research questions; (ii) search strategy; (iii) inclusion and exclusion criteria; and (iv) data extraction and synthesis methods.

- (i) **Research Questions:** These questions are structured according to the objective that is intended to be achieved with the systematic review and drive the subsequent steps of the review. Considering that our objective was a detailed overview involving embedded systems and agile methods, the following research questions (RQ) were established:

RQ1: What is the current state of agile methods in the development of embedded systems?

RQ2: Which agile methods are used most often in the development of embedded systems?

RQ3: What are the benefits, challenges, and limitations provided by the adoption of agile methods in the development of embedded systems?

- (ii) **Search Strategy:** In order to establish the search strategy based on the research questions, we initially identified the main keywords separated by area: “Agile”, “Scrum”, and “Extreme Programming” (for the Agile Methods area) and “Embedded” and “Electrical and Electronic System” (for the Embedded Systems area). It is worth highlighting that the keywords chosen needed to be simple enough to bring many results and, at the same time, rigorous enough to cover only the desired research topic. The search string was then built based on the keywords and using Boolean connectors (such as AND/OR). The search string used in our systematic review was: (‘Agile’ OR ‘Scrum’ OR ‘Extreme Programming’) AND (‘Embedded’ OR ‘Electrical and Electronic System’).

We also selected larger publications databases as sources of primary studies: IEEE Xplore⁶, ACM Digital Library⁷, Springer Link⁸, Scirus⁹, Web of Science¹⁰, ScienceDirect¹¹, and SCOPUS¹². Those sources are cited in [12] as trusted in the software engineering context. Furthermore, we decided that only papers written in English would be considered, since English is widely used to write scientific papers.

- (iii) **Inclusion and exclusion criteria:** The selected primary studies needed to be assessed for their relevance, enabling the inclusion of studies that provide evidence for the research questions as well as the exclusion of studies that do not. This is achieved by the definition of Inclusion Criteria (IC) and Exclusion Criteria (EC). Thus, our inclusion criteria were:

IC1: The study involves an agile method in the development of embedded systems;

IC2: The study presents an agile method being used in the development of embedded systems; and

IC3: The study presents benefits, challenges, and limitations of agile methods in the development of embedded systems.

⁶ <http://www.ieeexplore.ieee.org>

⁷ <http://www.portal.acm.org>

⁸ <http://www.springer.com/lncs>

⁹ <http://www.scirus.com/>

¹⁰ <http://www.isiknowledge.com>

¹¹ <http://www.sciencedirect.com>

¹² <http://www.scopus.com>

The exclusion criteria established were:

EC1: The study does not address the development of embedded systems using an agile method;

EC2: The study presents an abstract and/or an introductory section apparently related to embedded systems and agile methods; however, the rest of the text is not, in fact, related;

EC3: The study does not present any abstract or it is not available for further reading;

EC4: The study is written in a language other than English;

EC5: The study is directly related to another primary study by the same author; and

EC6: The study consists of a compilation of work, for instance, from a conference or workshop.

- (iv) **Data extraction and synthesis methods:** In order to extract relevant data, we checked and analyzed each primary study, aiming at relating it to the research questions. After that, we synthesized the results in order to obtain a reliable, comprehensive, and detailed overview of the research topic of our systematic review.

3.2 Step 2: Conduction

In this step, we conducted the search for the primary studies according to the previously established plan. This was achieved by performing a search in the selected databases using the previously built search string. In order to enable the correct results to be delivered, the search string was adjusted for each specific database and its search mechanisms. As a result, a total of 494 primary studies were found in these databases. The selection criteria (i.e., inclusion and exclusion criteria) were then applied to select the relevant primary studies. For this, the title and abstract of each primary study were read first and, if interesting, the study was initially selected. Next, we read each primary study in full and again applied the selection criteria in order to decide whether to include the study as relevant for our systematic review. Table 1 summarizes the total number of primary studies obtained from each database (column 4), as well as the number of studies included after applying the selection criteria (column 2). Out of a total of 494 primary studies previously identified, 51 studies were included. However, among the 51 studies, repetitive studies were identified, since these databases sometimes used the same primary studies. At the end, 23 unique studies were included as relevant. Regarding the relevance of each database, we can say that Scopus was the most efficient source, returning 14 studies out of 23 studies included. Thus, if only this database had been used in our systematic review, at least 60.8 % of the relevant studies would have been found.

Analyzing from another perspective, we found 182 repetitive studies (out of a total of 494 studies) and, thus, we in fact analyzed 312 studies. This means that the 23 unique studies refer to 7.37 % of the non-repetitive studies. In Table 2, we present the 23 studies (S1 to S23), specifically their authors, titles, and inclusion criteria, organized according to their publication years.

To support the organization and manipulation of the primary studies, we used EndNote [13](http://www.endnote.com/), a software tool for publishing and managing bibliographies, in combination with a spreadsheet software.

Table 1. Number of primary studies included, excluded, and found

Source	Included	Excluded	Total Found
ACM Digital Library	4	97	101
IEEE Xplore	11	83	94
Scirus	6	41	47
Springer Link	3	20	23
ISI Web of Knowledge	12	100	112
ScienceDirect	1	10	11
Scopus	14	92	106
Total	51	443	494

Table 2. Included primary studies

#	Authors	Title	Year	IC
S1	Mueller, G. and Borzuchowski, J. [13]	Extreme embedded a report from the front line	2002	IC1, IC2, IC3
S2	Karlstroem, D. and Runeson, P. [14]	Scaling Extreme Programming in a Market Driven Development Context	2003	IC1, IC2, IC3
S3	Ronkainen, J. and Abrahamsson, P. [15]	Software Development under Stringent Hardware Constraints: Do Agile Methods Have a Chance?	2003	IC2, IC3
S4	Greene, B. [9]	Agile Methods Applied to Embedded Firmware Development	2004	IC1
S5	Manhart, P. and Schneider, K. [16]	Breaking the Ice for Agile Development of Embedded Software: An Industry Experience Report	2004	IC1
S6	Kettunen, P. and Laanti, M. [17]	How to steer an embedded software project: tactics for selecting the software process model	2004	IC1, IC2
S7	Chae, H.; Lee, D.; Park, J.; and In, H. P. [18]	The Partitioning Methodology in Hardware/Software Co-Design Using Extreme Programming: Evaluation through the Lego Robot Project	2006	IC1
S8	Cordeiro, L.; Barreto, R.; Barcelos, R.; Oliveira, M.; Lucena, V.; and Maciel, P. [19]	TXM: An Agile HW/SW Development Methodology for Building Medical Devices	2007	IC1, IC2, IC3

¹³ <http://www.endnote.com/>

Table 2. (continued)

#	Authors	Title	Year	IC
S9	Fletcher, M.; Bereza, W.; Karlesky, M.; and Williams, G. [20]	Evolving into Embedded Development	2007	IC1
S10	Cordeiro, L.; Barreto, R.; Barcelos, R.; Oliveira, M.; Lucena, V.; and Maciel, P. [21]	Agile Development Methodology for Embedded Systems: A Platform-Based Design Approach	2007	IC1, IC2, IC3
S11	Wang, Z. [22]	Fuxi: An Agile Development Environment for Embedded Systems	2007	IC1
S12	Cordeiro, L.; Mar, C.; Valentin, E.; Cruz, F.; Patrick, D.; Barreto, R.; and Lucena, V. [23]	An Agile Development Methodology Applied to Embedded Control Software under Stringent Hardware Constraints	2008	IC1, IC2, IC3
S13	Cordeiro, L.; Mar, C.; Valentin, E.; Cruz, F.; Patrick, D.; Barreto, R.; and Lucena, V. [24]	A Platform-Based Software Design Methodology for Embedded Control Systems: An Agile Toolkit	2008	IC1, IC2, IC3
S14	Cordeiro, L.; Barreto, R.; and Oliveira, M. [25]	Towards A Semiformal Development Methodology for Embedded Systems	2008	IC1, IC2, IC3
S15	Salo, O. and Abrahamsson, P. [26]	Agile Methods in European Embedded Software Development Organisations: a Survey on the Actual Use and Usefulness of Extreme Programming and Scrum	2008	IC1
S16	Wilking, D. [10]	Empirical Studies for the Application of Agile Methods to Embedded Systems	2008	IC1
S17	Hill, J. [7]	Agile Techniques for Developing and Evaluating Large-scale Component-based Distributed Real-time and Embedded Systems	2009	IC1, IC2, IC3
S18	Smith, M.; Miller, J.; and Daeninck, S. [27]	A Test-oriented Embedded System Production Methodology	2009	IC1, IC2, IC3
S19	Smith, M.; Miller, J.; Huang, L.; and Tran, A. [28]	A More Agile Approach to Embedded System Development	2009	IC1, IC2, IC3
S20	Srinivasan, J.; Dobrin, R.; and Lundqvist, K. [29]	'State of the Art' in Using Agile Methods for Embedded Systems Development	2009	IC1, IC2, IC3
S21	Heidenberg, J.; Hirkman, P.; Matinlassi, M.; Partanen, J.; and Pikkarainen, M. [11]	Systematic Piloting of Agile Methods in the Large: Two Cases in Embedded Systems Development	2010	IC1, IC2, IC3
S22	Cawley, O.; Wang, X.; and Richardson, I. [30]	Lean/Agile Software Development Methodologies in Regulated Environments - State of the Art	2010	IC1, IC2, IC3
S23	Savolainen, J.; Kuusela, J.; and Vilavaara, A. [31]	Transition to Agile Development - Rediscovery of Important Requirements Engineering Practices	2010	IC1, IC2, IC3

3.3 Step 3: Reporting

In this step, a qualitative and descriptive analysis was conducted on each selected primary study, enabling us to achieve more accurate answers to our research questions. First of all, it was observed that studies involving agile methods for building embedded systems are quite recent. As presented in Figure 2, out of a total of 23 included primary studies – the first published in 2002 and the others in the following years – the increase in the number of studies does not indicate a trend statistically, but a growth in the interest for this research area, with studies more concentrated in the last years. It is important to state that only studies published until September 2011 were considered in our systematic review. Bellow, we discuss each research question established previously:

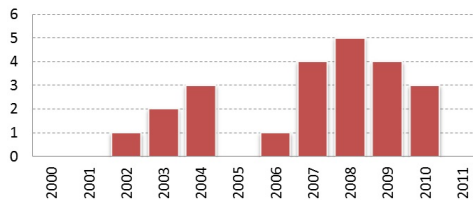


Fig. 2. Distribution of the relevant primary studies through the years

RQ1: This question is related to the current state of the adoption of agile methods in the development of embedded systems. Through the primary studies identified in our systematic review, we can observe that agile methods, their adaptations, and their practices have been experimented within real projects of industrial organizations. Greene [9] (study S4) describes an experience in applying agile approaches to the development of firmware for the Intel Itanium processor family. Studies S12 and S13 present the development of a digital soft-starter [23] and of a induction motor simulator prototypes, respectively [24]. Besides that, studies S8 and S10 present the development of a pulse oximeter [19,21]. Moreover, studies S2, S6, S17, and S21 focused on the development of large embedded systems. Thus, according to Ronkainen and Abrahamsson [15] (study S3), Wilking [10] (study S16), and Oisín et al. [30] (study S22), agile methods can be successfully applied in the embedded system domain.

Another interesting finding was that the application of agile methods exactly as they were established by their authors is not found in the embedded systems domain. On the other hand, the use of a subset of agile practices or the adaptation of particular methods based on agile methods were identified. Regarding the use of agile practices, we found that a considerable number of primary studies identified in our systematic review (in particular, studies S1, S4, S5, S7, S8, S10, S12, S13, S15, S18, S19, S20, and S21) present experiences regarding the use of a single or a set of agile practices. Regarding adaptation, we found studies S11, S14, and S20, which are related to the adaptation of a particular method or the proposal of a new method for embedded systems based on agile methods. Wang

[22] (study S11) proposes an architecture-based, aspect-oriented method for agile software development, which uses expressivity and efficiency as two major concerns of embedded systems. This method is based on separation of concerns (i.e., separating system functionality from system technicality). Study S20 also makes use of separation of concerns to effectively adopt agile methods in the development of embedded systems; do to so, this study considers two distinguished concerns [29]: (i) technical issues (such as requirements and testing); and (ii) organizational issues (such as process tailoring, knowledge sharing and transfer, culture change, and support for infrastructure development). In another study, Cordeiro et al. [25] (study S14) proposes a method – called TXM – that adapts agile principles in order to build embedded systems focusing on the issues related to the system’s constraints and safety. This method is composed of practices from “traditional” software engineering and agile methods (in particular, Scrum and XP), and aims at minimizing the main problems in software development (i.e., requirement volatility and risk management), and of practices that are needed to achieve hardware and software development. TXM was also successfully applied in industry projects [19,21,23,24]. We also found a unique primary study (study S23) that investigates a specific software development phase, i.e., requirements engineering [31]. This study shows how requirements engineering should be treated in the agile development process and illustrates this with two case studies done in industry.

Study S6 presents a comparative analysis of a range of software process models/methods (such as Rational Unified Process (RUP) [32], XP, Feature-Driven Development (FDD) [33], Waterfall, and Spiral model). This analysis is meant to support the selection of one of these models for the development of embedded systems, based on the particular characteristics of the embedded systems to be built. This means that there is concern regarding better understanding of process models and their characteristics in order to allow adopting a more suitable process/method for embedded systems.

As a result, we have observed that there are currently important, but isolated initiatives on exploring agile methods for the development of embedded systems, also in industry. However, more work should be conducted to allow more widespread and efficient application of these methods for that area.

RQ2: This research question refers to which agile methods have been used in the development of embedded systems. Although there are several agile methods proposed in the literature, we have observed that only Scrum and XP have been investigated and used in projects involving embedded systems. Salo et al. [26] (study S15) conducted a first survey in 2008 to address the level of adoption and usefulness of Scrum and XP in embedded systems development environments. The survey was conducted in 13 industrial organizations and showed that 54% of the organizations systematically, mostly, or sometimes used the practices of XP, whereas 27% used Scrum. It is worth highlighting that, according to another previous work [2], in the software industry in general, XP and Scrum have also been pointed out as the most frequently used agile methods. Therefore, the findings of our systematic review are in line with this previous work. While Scrum

outlines a process for developing software systems, XP also proposes a valorous set of agile practices; therefore, they can be seen as complementary approaches and could be explored more in a coordinated way in order to develop embedded systems. In our systematic review, we found several studies that combine XP and Scrum, such as studies S4, S8, S10, S12, S13, and S15.

Instead of using the agile methods as proposed by their authors, as stated above, a subset of agile practices have been adopted in the embedded systems domain. This is discussed in more detail in studies S1, S4, S5, S7, S8, S10, S12, S13, S15, S18, S19, S20, and S21. In study S21, the following practices were used in pilot projects [11]: time-boxed deliveries, increased face-to-face communication through meetings (planning meeting, daily meeting, and sprint demos), product (a list of requirements typically from the customer) and sprint backlogs (a list of prioritized requirements to be developed), and early testing (also test before coding or TDD Test-Driven Development). A similar set of agile practices was also considered in study S1, where Mueller and Borzuchowski [13] experimented with the main XP practices — i.e., pair programming (a practice where two programmers work together at one workstation, one as the driver who types in the code while the other acts as an observer), encouraging test before coding, short meetings, and use of white boards for task status and task schedules — in a real industrial project. Chae et al. [18] (study S7) also applied pair programming, together with other practices: improvement of communication among hardware and software team members, delivery of small releases, unit testing, refactoring, and continuous integration. They combine these practices with hardware/software co-design in order to manage the cost, development time, and risk of a project involving embedded systems. With respect to pair programming, Greene [9] (study S4) says that it is the most controversial practice, whereas the most valuable practice is unit testing. This study also combined the practices of Scrum and XP into its own running method. In particular, the Scrum practices used were sprints (30-day iterations), sprint planning meeting, daily Scrum, and sprint review (retrospective). A combination of Scrum and XP was also experimented by Salo et al. [26] (study S15); in this work, the most often used XP practices were open office space, coding standards, 40h week, continuous integration, and collective ownership, whereas the most often used Scrum practice was the product backlog. Studies S8, S10, S12, and S13 also used a combination of XP and Scrum practices in the method TXM (discussed above) and, in particular, they used sprints (30-day iterations). In another work, Manhart and Schneider (study S5) [16] also added agile practices (specifically, unit testing and test before coding) to an existing, running software development process. Besides that, studies S18 and S19 also explored the use of XP practices, specially, unit testing and refactoring, respectively. Finally, regarding adoption of specific agile practices, these practices as initially proposed do not appear to be totally suitable for embedded environments but rather need to be tailored. However, we can observe that the most common agile practice used in embedded systems projects is TDD. In this context, Srinivasan (S20) also realized that

TDD increases the overall quality of the development software and helps to elicit design decisions that were made as part of the hardware development process.

We observed that there is still no consensus regarding which agile methods and/or practices are most adequate for the embedded systems domain. There are thus important open opportunities for research in that direction. Moreover, other important agile methods, such as Open Unified Process (OpenUP) [34] and Feature-Driven Development (FDD), have not been cited or investigated in the embedded systems domain. OpenUP, in particular, is an interesting, lean Unified Process that applies an iterative, incremental approach within a structured process; it embraces a pragmatic, agile philosophy that focuses on the collaborative nature of software development [34]. Therefore, efforts should be dedicated to investigating whether the agility provided by these other methods could contribute to more effective development of embedded systems.

RQ3: This research question refers to the benefits, challenges, and limitations resulting from the adoption of agile methods to the development of embedded systems. The majority of the studies have indicate benefits in adopting agile methods to the building of embedded systems. In general, amongst the benefits achieved, the most cited ones are decrease of development time, improvement of productivity, and reduction of the error rate, i.e., improvement in the overall quality of the systems. In particular, according to the experiments presented in [25] (study S14), the application of a method based on agile principles substantially reduced the development time of the product. The results of other experiments presented by Chae et al. [18] (study S7) showed that productivity was improved and the error rate decreased. Furthermore, studies S1, S4, S5, S18, and S19 showed also that agile methods can be employed successfully on embedded projects. Mueller and Borzuchowski [13] (study S1) stated that XP, in particular, is a productive and valuable method for developing software for embedded applications.

Greene [9] (study S4) stated that agile methods present good outcomes when dealing with changes and uncertainty and, at the same time, embedded systems can also experience unexpected changes in requirements and hardware dependencies. Thus, agile methods are suited well for embedded systems in which engineers are averse to oppressive processes. Greene also points out also problems in traditional development that could be solved with agile methods [9]: schedule not followed, lack of test coverage, poor code maintainability, and little cross-training (when one team member is trained to do the tasks of another member). Some other advantages that can be perceived are: (i) delivery of small releases decrease the risk that the system will not meet the customers' intentions; and (ii) the probability to find bugs in later phases of development is decreased by using unit testing and continuous integration.

In terms of challenges regarding the adoption of agile methods, the main one is the cultural change, i.e., change in how software is developed in the organization. In study S4, Greene [9] stated that developers are reluctant to adopt new software development methods, mainly when this method is quite different from the traditional one. Besides that, according to Mueller and Borzuchowski [13]

(study S1), other challenges are scarce management support and the inexperience of the team regarding agile practices. Moreover, existing codified knowledge (available, for instance, in books and papers) does not explicitly address domain-specific characteristics, as discussed in [29] (study S20); thus, one should tailor agile methods such that they address the industry and particular characteristics of embedded systems. In spite of this, Salo et al. [26] (study S15) argue that embedded software development organizations are able to apply agile methods and their practices in their projects and achieve positive results.

With respect to the limitations of using agile methods, we have observed that these are more related to a lack of understanding of the change management required to introduce a new process than to the context of an embedded system itself. As mentioned in [13] (study S1), having a culture inside the company that encourages, promotes, and supports new ideas and changes is extremely important for the successful use of new methods. Another important limitation in using agile methods for the development of embedded systems is that many companies need to use quality certifications that demand, for instance, independent testing procedures not supported by these methods. Furthermore, according to Ronkainen and Abrahamsson [15] (study S3), in order to establish a foothold in the development of embedded systems, agile methods have to focus on specific embedded domain requirements such as: (i) techniques needed for determining specification and documentation needs; (ii) techniques needed for progressively increasing code maturity; (iii) techniques needed for recognizing and managing change-prone requirements; (iv) coordination and communication methods for inter-team work; and (v) techniques for building an optimal test suite.

4 Discussion

Through our systematic review, the existing knowledge about the application of agile methods in the development of embedded systems has been widely and thoroughly mapped, as the Systematic Review technique provided the mechanism for achieving it. We have found that agile methods and their practices have had a positive impact on embedded systems development. However, their use is still not widespread. Moreover, researchers have been conducting isolated work, as observed in Figure 3, where blue dots represent the authors and black dots the studies. Therefore, integration and collaborations could be promoted in order to achieve more effective results.

As far as using the knowledge presented in this work for further research is concerned, we have identified important, interesting research lines that should be investigated soon: (i) consensus about agile methods and their practices that best fit to the development of embedded systems, considering the specific characteristics of such systems; (ii) conduction of qualitative and quantitative analysis regarding, for instance, cost/effort reduction and quality improvement resulting from the use of agile methods; (iii) investigation on how to adopt agile methods when specific standards, such as ISO 26262 [35] (a functional safety standard), must be followed; and (iv) investigation on how agile methods impact different

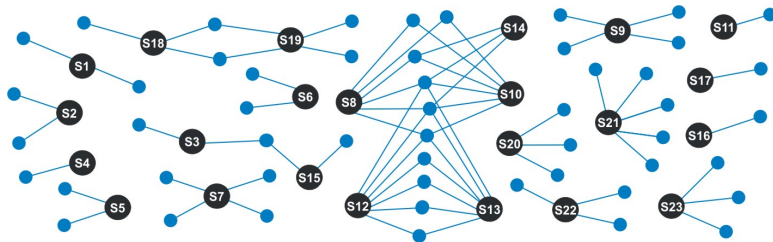


Fig. 3. Relationship among studies and authors

phases of the development process, such as requirement management, design, and architecting of embedded systems. The objective should be to establish a more suitable method for such systems.

Considering the positive results achieved, our systematic review could be conducted again. In such a repeat study, primary studies published between September 2011 and now should also be included. In order to permit repetition of this study, the systematic review protocol is explicitly divulged in this work. Besides that, relevant primary studies written in languages other than English could be considered. Furthermore, an investigation into the adoption of agile methods in different segments, such as avionics and consumer electronics, could be conducted. Although the databases used in our systematic review are usually considered quite efficient sources for software engineering, other databases, such as Compendex¹⁴ and Google Scholar¹⁵, could be included.

5 Conclusion and Future Work

As embedded systems are becoming increasingly large and complex, and require a high level of quality, considerable attention needs to be given to how they are developed. Although agile methods initially not appear to be suitable for these systems, they have provided a new, interesting perspective for developing such systems. The main contribution of this work is to present a comprehensive, detailed overview of the adoption of agile methods in the context of embedded systems. We systematically applied a set of steps provided by the Systematic Review technique. As main result, we have observed that there are already important, effective contributions exploring the agility provided by these methods. However, more attention should be given to this topic, especially regarding the establishment of a consensus among researchers about which agile methods and practices work best for the embedded systems domain. Based on these results, we have also identified some lines of research that must be explored yet. Furthermore, we want this overview to open other new, important research lines contributing to more effective and successful development of embedded systems.

¹⁴ <http://www.engineeringvillage.com>

¹⁵ <http://www.scholar.google.com>

Acknowledgments. This work was supported by the Brazilian funding agencies FAPESP, CNPq, and CAPES.

References

1. Vahid, F., Givargis, T.D.: *Embedded System Design: A Unified Hardware/Software Introduction*, 1st edn. Wiley Higher Education (2002)
2. West, D., Grant, T.: Agile development: Mainstream adoption has changed agility. In: *World Wide Web (2011)*, (On-line) http://www.forrester.com/rb/Research/agile_development_mainstream_adoption_has_changed_agility/q/id/56100/t/2 (accessed October 02, 2011)
3. Larman, C.: *Agile and Iterative Development: A Manager's Guide*. Pearson Education (2003)
4. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele Univ. and Durham Univ. (2007)
5. Kamal, R.: *Embedded systems: architecture, programming and design*. McGraw-Hill (2003)
6. Wilmshurst, T.: *Designing embedded systems with PIC microcontrollers: principles and applications*. Elsevier (2007)
7. Hill, J.H.: *Agile Techniques for Developing and Evaluating Large-scale Component-based Distributed Real-time and Embedded Systems*. PhD thesis, Graduate School of Vanderbilt University (May 2009)
8. Vijayarathy, L., Turk, D.: Drivers of agile software development use. *Information and Software Technology* 54(2), 137–148 (2012)
9. Greene, B.: Agile methods applied to embedded firmware development. In: *ADC 2004*, Salt Lake City, Utah, pp. 71–77 (2004)
10. Wilking, D.: *Empirical studies for the application of agile methods to embedded systems*. Master's thesis, Aachen University (2008)
11. Heidenberg, J., Matinlassi, M., Pikkarainen, M., Hirkman, P., Partanen, J.: Systematic Piloting of Agile Methods in the Large: Two Cases in Embedded Systems Development. In: Ali Babar, M., Vierimaa, M., Oivo, M. (eds.) *PROFES 2010*. LNCS, vol. 6156, pp. 47–61. Springer, Heidelberg (2010)
12. Dyba, T., Dingsoyr, T., Hanssen, G.K.: Applying systematic reviews to diverse study types. In: *ESEM 2007*, Los Alamitos, USA, pp. 225–234 (2007)
13. Mueller, G., Borzuchowski, J.: Extreme embedded a report from the front line. In: *Practitioners Reports, OOPSLA 2002*, Seattle, USA (2002)
14. Karlstroem, D., Runeson, P.: Scaling Extreme Programming in a Market Driven Development Context. In: Marchesi, M., Succi, G. (eds.) *XP 2003*. LNCS, vol. 2675, pp. 363–365. Springer, Heidelberg (2003)
15. Ronkainen, J., Abrahamsson, P.: Software Development Under Stringent Hardware Constraints: Do Agile Methods Have a Chance? In: Marchesi, M., Succi, G. (eds.) *XP 2003*. LNCS, vol. 2675, pp. 73–79. Springer, Heidelberg (2003)
16. Manhart, P., Schneider, K.: Breaking the ice for agile development of embedded software. In: *ICSE 2004*, Edinburgh, Scotland, pp. 378–386 (2004)
17. Kettunen, P., Laanti, M.: How to steer an embedded software project: tactics for selecting the software process model. *Inf. Software Technology* 47(9), 587–608
18. Chae, H., Lee, D., Park, J., In, H.P.: The partitioning methodology in hardware/software co-design using extreme programming: Evaluation through the Lego robot project. In: *CIT 2006*, Washington, DC, USA, p. 187 (2006)

19. Cordeiro, L., Barreto, R., Barcelos, R., Oliveira, M., Lucena, V., Maciel, P.: Txm: an agile hw/sw development methodology for building medical devices. SIGSOFT Softw. Eng. Notes 32(6) (November 2007)
20. Fletcher, M., Bereza, W., Karlesky, M., Williams, G.: Evolving into embedded development. In: AGILE 2007, Aalborg, Denmark, pp. 150–155 (2007)
21. Cordeiro, L., Barreto, R., Barcelos, R., Oliveira, M., Lucena, V., Maciel, P.: Agile development methodology for embedded systems: A platform-based design approach. In: ECBS 2007, Tucson, USA, pp. 195–202 (2007)
22. Wang, Z.: Fuxi: An agile development environment for embedded systems. In: COMPSAC 2007, Beijing, China, pp. 631–632 (2007)
23. Cordeiro, L., Mar, C., Valentin, E., Cruz, F., Patrick, D., Barreto, R., Lucena, V.: An agile development methodology applied to embedded control software under stringent hardware constraints. SIGSOFT Softw. Eng. Notes 33(5), 1–10 (2008)
24. Cordeiro, L., Mar, C., Valentin, E., Cruz, F., Patrick, D., Barreto, R., Lucena, V.: A platform-based software design methodology for embedded control systems: An agile toolkit. In: ECBS 2008, Belfast, Northern Ireland, pp. 408–417 (2008)
25. Cordeiro, L., Barreto, R., Oliveira, M.: Towards a semiformal development methodology for embedded systems. In: ENASE 2008, Funchal, Portugal, pp. 5–12 (2008)
26. Salo, O., Abrahamsson, P.: Agile methods in european embedded software development organisations: a survey on the actual use and usefulness of eXtreme programming and Scrum. IET Software 2(1), 58–64 (2008)
27. Smith, M., Miller, J., Daeninck, S.: A test-oriented embedded system production methodology. Journal of Signal Processing Systems 56(1), 69–89 (2009)
28. Smith, M., Miller, J., Huang, L., Tran, A.: A more agile approach to embedded system development. IEEE Software 26(3), 50–57 (2009)
29. Srinivasan, J., Dobrin, R., Lundqvist, K.: State of the Art in using agile methods for embedded systems development. In: COMPSAC, Seattle, USA, pp. 522–527 (2009)
30. Cawley, O., Wang, X., Richardson, I.: Lean/Agile Software Development Methodologies in Regulated Environments – State of the Art. In: Abrahamsson, P., Oza, N. (eds.) LESS 2010. LNBIIP, vol. 65, pp. 31–36. Springer, Heidelberg (2010)
31. Savolainen, J., Kuusela, J., Vilavaara, A.: Transition to agile development - re-discovery of important requirements engineering practices. In: RE 2010, Sydney, Australia, pp. 289–294 (2010)
32. Kruchten, P.: The Rational Unified Process: An Introduction, 3rd edn. The Addison-Wesley Object Technology Series. Addison-Wesley (2003)
33. Palmer, S.R., Felsing, J.M.: A Practical Guide to Feature-Driven Development. Prentice Hall (2002)
34. Eclipse: OpenUP, World Wide Web (2011) (On-line), <http://epf.eclipse.org/wikis/openup/> (accessed December 29, 2011)
35. International Organization for Standardization: ISO/DIS 26262 Software compliance: Achieving functional safety in the automotive industry (2011)

Heap Slicing Using Type Systems

Mohamed A. El-Zawawy

College of Computer and Information Sciences, Al-Imam M. I.-S. I. University
Riyadh, Kingdom of Saudi Arabia

and

Department of Mathematics, Faculty of Science, Cairo University
Giza 12613, Egypt
maelzawawy@cu.edu.eg

Abstract. Using type systems, this paper treats heap slicing which is a technique transforming a program into a new one that produces the same result while working on a heap sliced into independent regions. Heap slicing is a common approach to handle the problem of modifying the heap layout without changing the program semantics. Heap slicing has applications in the areas of performance optimization and security.

Towards solving the problem of heap slicing, this paper introduces three type systems. The first type system does a pointer analysis and annotates program points with pointer information. This type system is an augmentation of a previously developed type system by the author. The second type system does a region analysis and refines the result of the first type system by augmenting the pointer information with region information. The region information approximately specifies at each program point for each memory cell the region where the cell exists. The third type system uses the information gathered by the region type system to do the principal transformation of heap slicing.

The paper also presents two operational semantics; one for single-region heap scenario and the other for multi-regions heap scenario. These semantics are used to prove the soundness of the type systems.

Keywords: heap slicing, type systems, semantics of programming languages, operational semantics, region analysis, pointer analysis.

1 Introduction

Heap slicing [28,31] is a technique that transforms a program into a new one that produces the same result while working on a heap sliced into independent regions. This transformation enables an optimizing compiler to figure out memory cells that must lie in different slices of the heap. The input to this technique is a program in which integer argument-expressions in statements allocating memory cells are annotated with slice (region) names. Every slice only contains data that was annotated with the slice name. Arithmetic and Boolean operations are allowed only between arguments in the same slice. Usually, it is assumed that no cell in a slice is allowed to point to a cell in a different slice.

<ol style="list-style-type: none"> 1. $x := \text{cons}(1 : R_1, 2 : R_2);$ 2. $y := \text{cons}(x, 3 : R_2, 4 : R_3, 5 : R_1);$ 3. $z := \text{cons}(y, 6 : R_3, 7 : R_2);$ 4. $w := [x + 1];$ 5. $t := [y + 2];$ 6. $[z + 1] := t;$ 	\hookrightarrow	$x := \text{cons}'(1 : R_1, 2 : R_2);$ $y := \text{cons}'(x : \{1\}, 3 : R_2, 4 : R_3, 5 : R_1);$ $z := \text{cons}'(y : \{1\}, 6 : R_3, 7 : R_2);$ $w :=_{\{2\}} [x + 1];$ $t :=_{\{3\}} [y + 2];$ $[z + 1] :=_{\{3\}} t;$
---	-------------------	--

Fig. 1. A motivating example

Very often while maintaining a large software, it becomes apparent that a change to the heap layout (e.g. adding arguments to an allocation statement) is necessary. The amount of code depending on the heap layout can make the process of introducing such a change, even when it is very little, very tricky. Introducing changes in such situations can be time-consuming and it scarifies the software correctness as it may call bugs. The heap slicing techniques are good tools to address the problem of altering the heap layout without changing the program semantics.

Heap slicing has applications in the areas of performance optimization and security [29][3]. The instance interleaving optimization is a static analysis [20] technique that rearranges the memory cells (or fields of different data structures) to improve cache performance via letting frequently-accessed fields (or cells) belong to the same cache line. Heap slicing techniques provide good implementations for instance interleaving optimization [20]. In security, heap slicing can be used to hide function pointers in a heap slice (region) preventing attackers from accessing them.

Motivating Example

Figure 1 shows a motivating example of our work. Consider the program on the l.h.s. of the figure. The integer-expressions of the allocation statements are annotated with their region names. For example the first allocation statement allocates an array of length two: the first of which belongs to region 1 and the second of which belongs to region 2. The goal of our research is to automatically transform such a program into the program on the r.h.s. of the figure. In the new program: (a) the address expressions (expressions evaluates to addresses) of allocation statements are annotated with their region names, and (b) mutation and look-up statements are annotated with reign names where the statements are allowed to be executed.

While the original program is assumed to be executed on a one-slice heap, the new program is executed on a heap that physical sliced into 3 regions. The number of the regions is fixed in the programming language. Figures 2 and 3 show the heaps of the original and new programs, respectively, after executing the allocation statements.

Moreover, we want to associate each of such program transformation with a proof that original and new programs have the same semantics: compute the same result. This proof is required in many application like *proof-carrying code* [19][22].

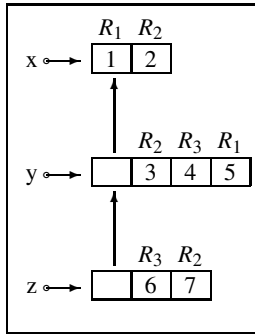


Fig. 2. One-slice heap

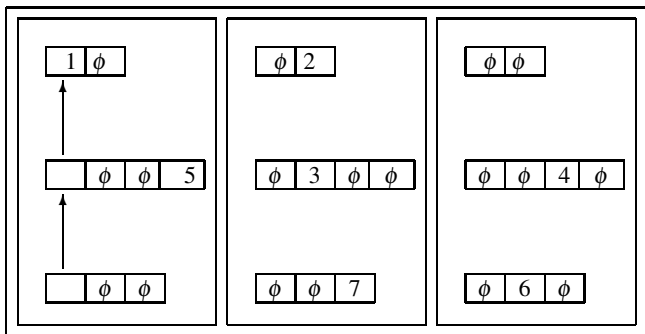


Fig. 3. Sliced heap

Algorithm

The transformation process described above on the motivating example is achieved in this paper using a 3-steps algorithm. Each of the 3 steps is accomplished by a type system. The first step is a pointer analysis to the input program. This analysis results in annotating the program points with points-to information in the form of types. The points-to information at a given point specifies approximately for each store (a variable or a memory cell) the address that has a chance of going into that store. The second step is a region analysis to the program resulting from the first step. This analysis results in augmenting the pointer information with region information in the form of types. The region information at a given point specifies approximately the region for each memory cell. Also the region information at a given point specifies approximately for each variable the source region of the variable’s content. The third step does the transformation step using the information gathered in the previous steps.

The justification (proof) that the source and the new programs are semantically equivalent takes the form of a type derivation.

$$\begin{aligned}
& \oplus \in \{+, -, \times\}, x \in \text{Var}, \text{ and } \{i, j\}, R_s \subseteq \{1, \dots, \gamma\} \\
e \in \text{Aexprs} & ::= x \mid n \mid e_1 \oplus e_2 \mid \text{Cast}(R_i \leftrightarrow R_j)e \\
d \in \text{Allo-exprs} & ::= e : R_i \mid e \\
b \in \text{Bexprs} & ::= \text{true} \mid \text{false} \mid \neg b \mid e_1 = e_2 \mid e_1 \leq e_2 \mid b_1 \wedge b_2 \mid b_1 \vee b_2 \\
S \in \text{Stmts} & ::= x := e \mid x := \text{cons}(d_1, \dots, d_n) \mid x := [e] \mid [e_1] := e_2 \mid \text{dispose}(e) \mid \\
& \quad \text{skip} \mid S_1; S_2 \mid \text{if } b \text{ then } S_f \text{ else } S_f \mid \text{while } b \text{ do } S_f \\
S' \in \text{Stmts}' & ::= x := e \mid x := \text{cons}'(e_1 : R_{s_1}, \dots, e_n : R_{s_n}) \mid x :=_{R_s} [e] \mid [e_1] :=_{R_s} e_2 \\
& \quad \mid \text{dispose}'(e) \mid \text{skip} \mid S'_1; S'_2 \mid \text{if } b \text{ then } S'_f \text{ else } S'_f \mid \text{while } b \text{ do } S'_f
\end{aligned}$$

Fig. 4. Our language for studying heap slicing

Contributions

Contributions of this paper are the following.

1. A type system for pointer analysis of the language presented in this paper. This type system is an augmented version of that we presented in [13].
2. A novel approach for region analysis (in the form of a type system as well).
3. An original technique for heap slicing.
4. Two new operational semantics; one for single-region heap scenario and the other for multi-regions heap scenario.

Organization

The rest of the paper is organized as follows. A toy programming language together with two operational semantics (one for single-region heap scenario and the other for multi-regions heap scenario) are presented in Section 2. Type systems for flow-sensitive pointer and region analyses are presented in Sections 3 and 4, respectively. The type system carrying program optimization is introduced in Section 5. A brief survey of related work and future work are presented in Section 6.

2 Programming Language and Two Operational Semantics

This section presents the programming language used to study heap slicing. The section also presents an operational semantics [23] for a one-slice heap executions and another operational semantics for γ -slices heap executions. The number of regions or slices in memory is fixed in our language and is denoted by γ .

We have two memory models; one for the single-slice heap scenario and the other for the γ -slices heap scenario. In our single-heap model, we assume that for any $m \in \mathbb{N}^+$ the memory has an infinite number of arrays of length m with addresses $\{a_{m,1}^1, a_{m,2}^1, \dots, a_{m,m}^1, a_{m,1}^2, a_{m,2}^2, \dots, a_{m,m}^2, \dots\}$. Therefore the set of address, *Addr*s, has the form presented in Figure 5. In order to facilitate evaluating inequalities we assume

$$\begin{aligned}
\text{Atoms} &\subseteq \text{Integers.} \\
\text{Addr} &= \{a'_{j,k} \mid i, j, k \in \mathbb{N}^+, k \leq j\} \\
&= \{a^1_{1,1}, a^1_{2,1}, a^1_{2,2}, a^1_{3,1}, a^1_{3,2}, a^1_{3,3}, \dots, a^2_{1,1}, a^2_{2,1}, a^2_{2,2}, a^2_{3,1}, a^2_{3,2}, a^2_{3,3}, \dots\}. \\
\mathcal{R} &= \{1, \dots, \gamma\}. \\
\text{Values} &= \mathbb{Z} \cup \text{Addr}. \\
\text{Values}^+ &= \text{Values} \cup \{\phi\}.
\end{aligned}$$

Fig. 5. Entities of our memory model

that the set *Values* is equipped with an order. We assume that our γ -slices memory model consists of γ separated regions each of which has the single-slice model. The value ϕ in the set *Values*⁺ goes into cells that are inactive in a region. Arithmetic and Boolean operations are only allowed between arguments of the same region.

The language (Figure 4) that we study is based on the programming language usually used to introduce separation logic [24]. There are two additions to the separation logic language. The first addition is that the arithmetic expression is extended with a cast statement permitting handling a value that we obtained from region i as it is obtained from region j . This is useful in many situations like if the programmer is interested in copying a value from a private slice of a memory to a public slice. The other addition is to annotate arguments (the ones evaluates to integers) of the allocation statement with region names. *Stmt'* presents the syntax of transformed programs. A clue to meaning of *Stmt'*-commands is given by the motivating example above and a precise meaning is given below by operational semantics.

The states of our operational semantics are defined as follows.

- Definition 1.**
1. $s \in \text{Stacks} = \{(s_v, s_r) \mid s_v : \text{Var} \rightarrow \text{Values} \text{ and } s_r : \text{Var} \rightarrow \mathcal{R} \cup \{\perp\}\}$.
 2. $h \in \text{Heaps} = \{(h_v, h_r) \mid h_v : A \rightarrow \text{Values}, h_r : A \rightarrow \mathcal{R}, \text{ and } A \subseteq_{\text{fin}} \text{Addr}\}$.
 3. A sliced heap \tilde{h} is a γ -tuple $(\tilde{h}_1, \dots, \tilde{h}_\gamma)$ of finite partial maps from *Addr* to *Values*⁺ such that:
 - (a) these maps share the same domain, and
 - (b) for any $a \in \text{dom}(\tilde{h}_1)$ there is a unique $i \in [1, \gamma]$ such that $\tilde{h}_i(a) \neq \phi$.

- Definition 2.**
1. A state is an abort or a pair of a stack and a heap (s, h) .
 2. A sliced state is an abort or a pair of a stack and a sliced heap (s, \tilde{h}) .

2.1 One-Slice Heap Semantics

This section presents an operational semantics for the input program of our transformation technique. The states of the semantics are defined in Definition 2.1.

The semantics of arithmetic and Boolean expressions are defined as follows:

$$\begin{aligned}
\llbracket d \rrbracket \in \text{States} &\rightarrow \text{Values} \times (\mathcal{R} \cup \{\perp\}) \\
\llbracket n \rrbracket(s, h) &= (n, \perp) \quad \llbracket x \rrbracket(s, h) = (s_v(x), s_r(x)) \quad \llbracket e_1 \oplus e_2 \rrbracket(s, h) = \eta_n(\llbracket e_1 \rrbracket(s, h) \oplus \llbracket e_2 \rrbracket(s, h))
\end{aligned}$$

where,

$$\eta_h(\alpha, \beta) \begin{cases} (\alpha, \beta), & \text{if } \alpha \in \mathbb{Z}; \\ (\alpha, h_r(\alpha)), & \text{if } \alpha \in \text{dom}(h_r); \\ \text{undefined}, & \text{otherwise.} \end{cases} \quad \llbracket e : R_i \rrbracket(s, h) = \begin{cases} (n, i) & \text{if } \llbracket e \rrbracket(s, h) \in \{(n, i), (n, \perp)\} \\ \text{undefined} & \text{otherwise.} \end{cases}$$

$$\llbracket \text{cast}(R_i \hookrightarrow R_j)e \rrbracket(s, h) = \begin{cases} (n, j) & \text{if } \llbracket e \rrbracket(s, h) \in \{(n, j), (n, i)\}, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

The semantics of the operation \oplus is defined as usual if both of its operands are integers and otherwise as follows:

$$v_1 \oplus v_2 = \begin{cases} (n \oplus m, \perp), & \text{if } v_1 = (n, \perp) \text{ and } v_2 = (m, \perp); \\ (n \oplus m, i), & \text{if } v_1 = (n, i) \text{ and } (v_2 = (m, i) \text{ or } v_2 = (m, \perp)); \\ (\alpha_{s,t \oplus n}^i), & \text{if } v_1 = (\alpha_{s,t}^i), (v_2 = (n, i) \text{ or } v_2 = (n, \perp)), \text{ and } 1 \leq t \oplus n \leq s; \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

Boolean operations are only allowed between values from the same region.
The inference rules of the semantics are defined as follows.

$$\frac{}{\text{skip} : (s, h) \rightarrow (s, h)} \quad \frac{\llbracket e \rrbracket(s, h) \text{ is undefined}}{x := e : (s, h) \rightarrow \text{abort}} \quad \frac{\llbracket e \rrbracket(s, h) = (\alpha, \beta)}{x := e : (s, h) \rightarrow ([s_v \mid x : \alpha], [s_r \mid x : \beta], h)}$$

$$\frac{u = \min\{t \mid \{a_{n,1}^t, \dots, a_{n,n}^t\} \cap \text{dom}(h) = \emptyset\} \quad \forall 1 \leq i \leq n (\llbracket d_i \rrbracket(s, h) = (\alpha_i, \beta_i))}{\begin{array}{c} x := \text{cons}(d_1, \dots, d_n) : (s, h) \rightarrow \\ ([s_v \mid x : a_{n,1}^u], [s_r \mid x : \beta_1], [h_v \mid a_{n,1}^u : \alpha_1 \mid \dots \mid a_{n,n}^u : \alpha_n], [h_r \mid a_{n,1}^u : \beta_1 \mid \dots \mid a_{n,n}^u : \beta_n]) \end{array}}$$

$$\frac{\exists 1 \leq i \leq n (\llbracket d_i \rrbracket(s, h) \text{ is undefined})}{x := \text{cons}(d_1, \dots, d_n) : (s, h) \rightarrow \text{abort}} \quad \frac{\begin{array}{l} \llbracket e \rrbracket(s, h) \text{ is undefined, or} \\ \llbracket e \rrbracket(s, h) = (\alpha, _)\wedge \alpha \notin \text{dom}(h) \end{array}}{\text{dispose}(e) : (s, h) \rightarrow \text{abort}}$$

$$\frac{}{x := [e] : (s, h) \rightarrow \begin{cases} ([s_v \mid x : h_v(\alpha)], [s_r \mid x : \beta], h), & \text{if } \llbracket e \rrbracket(s, h) = (\alpha, \beta) \text{ and } \alpha \in \text{dom}(h); \\ \text{abort}, & \text{otherwise.} \end{cases}}$$

$$\frac{}{[e_1] := e_2 : (s, h) \rightarrow \begin{cases} (s, [h_v \mid \alpha_1 : \alpha_2], [h_r \mid \alpha_1 : \beta]), & \text{if } \llbracket e_1 \rrbracket(s, h) = (\alpha_1, \beta) \text{ and } \alpha_1 \in \text{dom}(h); \\ \text{abort}, & \text{otherwise.} \end{cases}}$$

$$\frac{\llbracket e \rrbracket(s, h) = (\alpha, _)\wedge \alpha \in \text{dom}(h)}{\text{dispose}(e) : (s, h) \rightarrow (s, h_v \setminus \{\alpha\}, h_r \setminus \{\alpha\})} \quad \frac{S_1 : (s, h) \rightarrow (s', h') \quad S_2 : (s', h') \rightarrow \text{st}}{S_1; S_2 : (s, h) \rightarrow \text{st}}$$

$$\frac{S_1 : (s, h) \rightarrow \text{abort} \quad S_2 \in \text{Smts}}{S_1; S_2 : (s, h) \rightarrow \text{abort}} \quad \frac{\llbracket b \rrbracket(s, h) \text{ is undefined}}{\text{if } b \text{ then } S_t \text{ else } S_f : (s, h) \rightarrow \text{abort}} \quad \frac{\llbracket b \rrbracket(s, h) = \text{false}}{\text{if } b \text{ then } S_t \text{ else } S_f : (s, h) \rightarrow \text{st}}$$

$$\frac{\llbracket b \rrbracket(s, h) = \text{true}}{\text{if } b \text{ then } S_t \text{ else } S_f : (s, h) \rightarrow \text{st}} \quad \frac{\llbracket b \rrbracket(s, h) \text{ is undefined}}{\text{while } b \text{ do } S_t : (s, h) \rightarrow \text{abort}} \quad \frac{\llbracket b \rrbracket(s, h) = \text{true}}{\text{while } b \text{ do } S_t : (s, h) \rightarrow \text{abort}}$$

$$\frac{\llbracket b \rrbracket(s, h) = \text{false}}{\text{while } b \text{ do } S_t : (s, h) \rightarrow (s, h)} \quad \frac{\llbracket b \rrbracket(s, h) = \text{true} \quad S_t : (s, h) \rightarrow (s', h')}{\text{while } b \text{ do } S_t : (s, h) \rightarrow \text{st}}$$

If f is a map and A is a set, $f \upharpoonright A$ denotes the restriction of f on A and $[f \mid x : A]$ denotes the function whose domain is $\text{dom}(f) \cup \{x\}$ and whose definition is $\lambda y. \text{if } y = x \text{ then } A \text{ else } f(y)$.

Lemma 1. *Suppose $\llbracket e \rrbracket(s, h) = (\alpha, \beta)$. If $\alpha \in \text{Addrs}$ then $\beta = h_r(\alpha)$.*

2.2 γ -Slices Heap Semantics

This section presents an operational semantics for programs resulted by our proposed transformation technique. The semantics uses a memory model where the memory is physically sliced into γ regions. The states of the semantics are introduced in Definition 2.

The semantics of arithmetic and Boolean expressions is defined similarly to the one-heap semantics except that η_h is replaced with $\eta_{\tilde{h}}$:

$$\llbracket d \rrbracket \in \text{Sliced States} \rightarrow \text{Values} \times (\mathcal{R} \cup \{\perp\})$$

$$\eta_{\tilde{h}}(\alpha, \beta) = \begin{cases} (\alpha, \beta), & \text{if } \alpha \in \mathbb{Z}; \\ (\alpha, i), & \text{if } \alpha \in \text{dom}(\tilde{h}_i) \text{ and } \exists i \in \{1, \dots, \gamma\}. \tilde{h}_i(\alpha) \neq \phi; \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

The inference rules of the semantics are defined as follows.

$$\frac{}{\text{skip} : (s, \tilde{h}) \rightsquigarrow (s, \tilde{h})} \quad \frac{\llbracket e \rrbracket(s, \tilde{h}) \text{ is undefined}}{x := e : (s, \tilde{h}) \rightsquigarrow \text{abort}} \quad \frac{\llbracket e \rrbracket(s, \tilde{h}) = (\alpha, \beta)}{x := e : (s, \tilde{h}) \rightsquigarrow ([s_v | x : \alpha], [s_r | x : \beta], \tilde{h})}$$

$$\frac{u = \min\{t \mid \{a'_{n,1}, \dots, a'_{n,n}\} \cap \text{dom}(\tilde{h}_i) = \emptyset\} \quad \zeta_i(\alpha_j, \beta_j) = \begin{cases} \alpha_j, & \text{if } i = \beta_j; \\ \phi, & \text{otherwise.} \end{cases}}{x := \text{cons}'(d_1 : R_{S_1}, \dots, d_n : R_{S_n}) : (s, \tilde{h}) \rightsquigarrow \begin{cases} ([s_v | x : a'_{n,1}], [s_r | x : \beta_1], \dots, [\tilde{h}_i | a'_{n,1} : \zeta_i(\alpha_1, \beta_1) | \dots | a'_{n,n} : \zeta_i(\alpha_n, \beta_n)], \dots), \\ \text{if } \llbracket d_i \rrbracket(s, \tilde{h}) = (\alpha_i, \beta_i) \quad \beta_i \in R_{S_i}; \\ \text{otherwise.} \\ \text{abort,} \end{cases}}$$

$$\frac{}{x :=_{R_S} [e] : (s, \tilde{h}) \rightsquigarrow \begin{cases} ([s_r | x : \tilde{h}_\beta(\alpha)], [s_r | x : \beta], \tilde{h}), & \text{if } \llbracket e \rrbracket(s, \tilde{h}) = (\alpha, \beta) \quad \beta \in R_S; \\ \text{abort,} & \text{otherwise.} \end{cases}}$$

$$\frac{}{\begin{cases} [e_1] :=_{R_S} e_2 : (s, \tilde{h}) \rightsquigarrow \\ \begin{cases} (s, \dots, \tilde{h}_1, [\tilde{h}_\beta | \alpha_1 : \alpha_2], \dots, \tilde{h}_\gamma), & \text{if } \llbracket e_i \rrbracket(s, \tilde{h}) = (\alpha_i, \beta), \tilde{h}_\beta(\alpha_i) \neq \phi, \text{ and } \beta \in R_S; \\ \text{abort,} & \text{otherwise.} \end{cases} \end{cases}}$$

$$\frac{}{\text{dispose}(e) : (s, \tilde{h}) \rightsquigarrow \begin{cases} (s, \dots, \tilde{h}_i | (\text{dom}(\tilde{h}_i) \setminus \{\alpha\}), \dots), & \text{if } \llbracket e \rrbracket(s, \tilde{h}) = (\alpha, \beta) \text{ and } \alpha \in \text{dom}(\tilde{h}_i); \\ \text{otherwise.} \end{cases}}$$

$$\frac{S_1 : (s, \tilde{h}) \rightsquigarrow (s', \tilde{h}') \quad S_1 : (s, \tilde{h}) \rightsquigarrow \text{abort} \quad \llbracket b \rrbracket(s, \tilde{h}) \text{ is undefined}}{S_2 : (s', \tilde{h}') \rightsquigarrow \text{st} \quad S_2 \in \text{Stmts}} \quad \frac{}{S_1; S_2 : (s, \tilde{h}) \rightsquigarrow \text{st} \quad S_1; S_2 : (s, \tilde{h}) \rightsquigarrow \text{abort} \quad \text{if } b \text{ then } S_i \text{ else } S_f : (s, \tilde{h}) \rightsquigarrow \text{abort}}$$

$$\frac{\llbracket b \rrbracket(s, \tilde{h}) = \text{true} \quad S_i : (s, \tilde{h}) \rightsquigarrow \text{st}}{S_i : (s, \tilde{h}) \rightsquigarrow \text{st}} \quad \frac{\llbracket b \rrbracket(s, \tilde{h}) = \text{false} \quad S_f : (s, \tilde{h}) \rightsquigarrow \text{st}}{S_f : (s, \tilde{h}) \rightsquigarrow \text{st}} \quad \frac{\llbracket b \rrbracket(s, \tilde{h}) \text{ is undefined}}{S_i : (s, \tilde{h}) \rightsquigarrow \text{st}} \quad \frac{}{\text{if } b \text{ then } S_i \text{ else } S_f : (s, \tilde{h}) \rightsquigarrow \text{st} \quad \text{if } b \text{ then } S_i \text{ else } S_f : (s, \tilde{h}) \rightsquigarrow \text{st} \quad \text{while } b \text{ do } S_i : (s, \tilde{h}) \rightsquigarrow \text{abort}}$$

$$\frac{\llbracket b \rrbracket(s, \tilde{h}) = \text{true} \quad S_i : (s, \tilde{h}) \rightsquigarrow \text{abort}}{\text{while } b \text{ do } S_i : (s, \tilde{h}) \rightsquigarrow \text{abort}} \quad \frac{\llbracket b \rrbracket(s, \tilde{h}) = \text{false}}{\text{while } b \text{ do } S_i : (s, \tilde{h}) \rightsquigarrow (s, \tilde{h})} \quad \frac{\llbracket b \rrbracket(s, \tilde{h}) = \text{true} \quad S_i : (s, \tilde{h}) \rightsquigarrow (s', \tilde{h}')}{S_i : (s, \tilde{h}) \rightsquigarrow (s', \tilde{h}')} \quad \frac{}{\text{while } b \text{ do } S_i : (s, \tilde{h}) \rightsquigarrow \text{st} \quad \text{while } b \text{ do } S_i : (s, \tilde{h}) \rightsquigarrow \text{st}}$$

Lemma 2. Suppose $\llbracket e \rrbracket(s, \tilde{h}) = (\alpha, \beta)$. If $\alpha \in \text{Addr}_s$ then $\tilde{h}_\beta(\alpha) \neq \phi$.

Lemma 3. The semantics introduced in this section are well defined.

3 Pointer Analysis

This section presents a type system for pointer analysis [13,11,16,12,14,9] which is a flow-sensitive forward analysis. The analysis presented in this section is an augmented version of the type system we presented in [13]. We include the system here for the following reasons; (a) to make the current manuscript self-contained, (b) to show how differences between the language of this paper and that of [13] are treated, and (c) the following sections are built on this type system. The proof of the soundness for the type system presented here can be built by revising that presented in [13] peering in mind that the operational semantics used in both cases are different. The augmentation mentioned above is related to arithmetic expressions. The analysis annotates program points with partial maps (types of our type system) that approximatively specifies for each store the addresses that can go into the store.

The set of points-to types, PTS , and the sub-typing relation are defined as follows.

- Definition 3.** 1. $pts = \{pts \mid pts : Var \cup A \rightarrow 2^{Addr} \mid A \subseteq Addr\}$. The bottom type is denoted by \perp .
2. $pts \leq pts' \stackrel{\text{def}}{\iff} dom(pts) \subseteq dom(pts')$ and $\forall t \in dom(pts). pts(t) \subseteq pts'(t)$.
3. A state (s, h) has type pts , denoted by $(s, h) \models pts$, if
- $dom(h) \subseteq dom(pts)$,
 - $\forall x \in Var. s_v(x) \in Addr \implies s_v(x) \in pts(x)$, and
 - $\forall a \in dom(h). h_v(a) \in Addr \implies h_v(a) \in pts(a)$.

The pointer analysis of a program takes the form of a post-type derivation for a given pre-type. Typically \perp , the bottom type, is the pre-type.

The judgement of an arithmetic expression e has the form $e : pts \rightarrow V$. The set V is either a set of addresses or a singleton of an integer. The intended meaning, which is formalized in Lemma 4, of this judgement is that V captures any address that e evaluates to in a state of type pts . In particular if V is a set of addresses, then e is either an address from V or any integer.

The judgement of a statement S has the form $S : pts \rightarrow pts'$. The intuition, which is formalized in Theorem 1, of this judgement is that if S is executed in a state of type pts , then any state (rather than *abort*) where the execution ends is of type pts' .

The inference rules of our type system for pointer analysis are the following:

$$\begin{array}{c}
 \frac{}{n : pts \rightarrow \{n\}} \quad \frac{}{x : pts \rightarrow pts(x)} \quad \frac{}{cast(R_i \hookrightarrow R_j)e : pts \rightarrow \emptyset} \quad \frac{}{e : R_i : pts \rightarrow \emptyset} \\
 \\
 \frac{e_1 : pts \rightarrow V_1 \quad e_2 : pts \rightarrow V_2}{e_1 \oplus e_2 : pts \rightarrow \begin{cases} \{n \oplus m\} & \text{if } V_1 = \{n\} \wedge V_2 = \{m\}, \\ \{a_{i,j}^m \mid a_{i,j}^m \in V_2 \wedge 1 \leq j \oplus n \leq i\} & \text{if } V_1 = \{n\} \wedge V_2 \subseteq Addr, \\ \{a_{i,j}^m \mid a_{i,j}^m \in V_1 \wedge 1 \leq j \oplus n \leq i\} & \text{if } V_2 = \{n\} \wedge V_1 \subseteq Addr, \\ \{a_{i,j}^m \mid j = 1, \dots, i \text{ and for some } j, a_{i,j}^m \in V_1 \cup V_2\} & \text{otherwise.} \end{cases}}
 \end{array}$$

In the rest of the paper when $e : pts \rightarrow V$, we let V' denotes $V \cap Addr$.

$$\frac{}{skip : pts \rightarrow pts} \quad \frac{e : pts \rightarrow V}{x := e : pts \rightarrow [pts \mid x : V']} (ass^p)$$

$$\begin{array}{c}
\frac{v = \min\{t \mid \{d_{n,1}^t, \dots, d_{n,n}^t\} \cap \text{dom}(pts) = \emptyset\} \quad \forall 1 \leq i \leq n. d_i : pts \rightarrow V_i}{x := \text{cons}(d_1, \dots, d_n) : pts \rightarrow \cup_{1 \leq i \leq n} [pts \mid x : \{a_{n,1}^i\} \mid a_{n,1}^i : V_1' \mid \dots \mid a_{n,n}^i : V_n']]}{e : pts \rightarrow V} \text{(con}^p\text{)} \\
\frac{}{x := [e] : pts \rightarrow [pts \mid x : \cup_{a \in V'} pts(a)]} \text{(lok}^p\text{)} \quad \frac{\forall 1 \leq i \leq 2. e_i : pts \rightarrow V_i}{[e_1] := e_2 : pts \rightarrow \cup_{a \in V'} [pts \mid a : V_2']} \text{(mut}^p\text{)} \\
\frac{}{\text{dispose}(e) : pts \rightarrow pts} \text{(dis}^p\text{)} \quad \frac{S_1 : pts \rightarrow pts'' \quad S_2 : pts'' \rightarrow pts'}{S_1; S_2 : pts \rightarrow pts'} \text{(seq}^p\text{)} \quad \frac{S_f : pts \rightarrow pts'}{S_f : pts \rightarrow pts'} \text{(if}^p\text{)} \\
\text{if } b \text{ then } S_f \text{ else } S_f : pts \rightarrow pts' \\
\frac{S_f : pts \rightarrow pts}{\text{while } b \text{ do } S_f : pts \rightarrow pts} \text{(whl}^p\text{)} \quad \frac{pts_1' \leq pts_1 \quad S : pts_1 \rightarrow pts_2 \quad pts_2 \leq pts_2'}{S : pts_1' \rightarrow pts_2'} \text{(csq}^p\text{)}
\end{array}$$

Lemma 4. Suppose that $(s, h) \models pts$, $\llbracket d \rrbracket(s, h) = (\alpha, \beta)$ and $d : pts \rightarrow V$. Then

1. $V \subseteq \text{Addrs}$ or $\exists n \in \mathbb{Z}. V = \{n\}$,
2. $\forall n \in \mathbb{Z}. V = \{n\} \implies \alpha = n$, and
3. $\alpha \in \text{Addrs} \implies \alpha \in V$.

The soundness of the type system is stated in the following theorem whose proof can be driven from the corresponding theorem in [13].

Theorem 1. 1. $pts \leq pts' \iff (\forall (s, h), (s, h) \models pts \implies (s, h) \models pts')$.

2. Suppose that $S : pts \rightarrow pts'$ and $S : (s, h) \rightarrow (s', h')$. Then $(s, h) \models pts$ implies $(s', h') \models pts'$.

4 Region Analysis

In this section, we introduce a type system for region analysis which is a flow-sensitive, forward, and may analysis. The analysis annotates program points with region information in the form of partial maps from variables and memory locations to the power set of regions. Under these maps, the image of an address is an over-approximate set of regions where this address may exist. The image of a variable is an over-approximate set of regions from which the variable gets its value. We recall that the set of regions $\mathcal{R} = \{1, \dots, \gamma\}$.

The set of region types, *PTS-REG*, and the sub-typing relation are defined as follows.

Definition 4. 1. $REG = \{reg \mid reg : \text{Var} \cup A \rightarrow 2^{\mathcal{R}} \mid A \subseteq \text{Addrs}\}$.

2. $PTS\text{-}REG = \{(pts, reg) \in pts \times reg \mid \text{dom}(pts) = \text{dom}(reg)\}$.

3. $reg \leq reg' \iff \text{dom}(reg) \subseteq \text{dom}(reg')$ and $\forall t \in \text{dom}(reg). reg(t) \subseteq reg'(t)$.

4. $(pts, reg) \leq (pts', reg') \iff \text{def} \text{ } pts \leq pts' \text{ and } reg \leq reg'$.

5. A state (s, h) has type *reg*, denoted by $(s, h) \models reg$, if

– $\text{dom}(h_r) \subseteq \text{dom}(reg)$,

– $\forall t \in \text{Var}. s_r(t) = \beta \implies \beta \in reg(t)$, and $s_r(t) = \perp \implies reg(t) = \{1, \dots, \gamma\}$, and

– $\forall t \in \text{dom}(h_r). h_r(t) = \beta \implies \beta \in reg(t)$.

6. A state (s, h) has type (pts, reg) , denoted by $(s, h) \models (pts, reg)$, if
- $dom(pts) = dom(reg)$,
 - $(s, h) \models pts$, and
 - $(s, h) \models reg$.

The inference rules of our type system for region analysis are the following:

$$\begin{array}{c}
\frac{}{n : (pts, reg) \rightarrow \{1, \dots, \gamma\}} \quad \frac{}{x : (pts, reg) \rightarrow reg(x)} \quad \frac{}{Cast(R_i \hookrightarrow R_j) e : (pts, reg) \rightarrow \{j\}} \\
\frac{e_1 : (pts, reg) \rightarrow Rs_1 \quad e_2 : (pts, reg) \rightarrow Rs_2 \quad e_1 \oplus e_2 : pts \rightarrow V}{e_1 \oplus e_2 : (pts, reg) \rightarrow (Rs_1 \cap Rs_2) \cup (\cup_{a \in V'} reg(a))} \quad \frac{}{e : R_i : (pts, reg) \rightarrow \{i\}} \\
\frac{x := e : pts \rightarrow pts' \quad e : (pts, reg) \rightarrow Rs}{x := e : (pts, reg) \rightarrow (pts', [reg \mid x : Rs])} \quad (ass^R) \quad \frac{}{dispose(e) : (pts, reg) \rightarrow (pts, reg)} \quad (dis^R) \\
\frac{}{skip : (pts, reg) \rightarrow (pts, reg)} \quad \frac{x := [e] : pts \rightarrow pts' \quad e : (pts, reg) \rightarrow Rs}{x := [e] : (pts, reg) \rightarrow (pts', [reg \mid x : Rs])} \quad (lok^R) \\
\frac{v = \min\{t \mid \{d_{n,1}^t, \dots, d_{n,n}^t\} \cap dom(reg) = \emptyset\} \quad \forall 1 \leq i \leq n. d_i : (pts, reg) \rightarrow Rs_i}{x := cons(d_1, \dots, d_n) : (pts, reg) \rightarrow (pts', \cup_{1 \leq i \leq v} [reg \mid x : Rs_1 \mid d_{n,1}^i : Rs_1 \mid \dots \mid d_{n,n}^i : Rs_n])} \quad (con^R) \\
\frac{\frac{[e_1] := e_2 : pts \rightarrow pts' \quad e_2 : (pts, reg) \rightarrow Rs \quad e_1 : pts \rightarrow V}{[e_1] := e_2 : (pts, reg) \rightarrow (pts', \cup_{a \in V'} [reg \mid a : Rs])} \quad (mut^R) \quad \frac{S_1 : (pts, reg) \rightarrow (pts'', reg'') \quad S_2 : (pts'', reg'') \rightarrow (pts', reg')}{S_1; S_2 : (pts, reg) \rightarrow (pts', reg')} \quad (seq^R)}{\frac{S_r : (pts, reg) \rightarrow (pts', reg') \quad S_f : (pts, reg) \rightarrow (pts', reg')}{if \ b \ then \ S_r \ else \ S_f : (pts, reg) \rightarrow (pts', reg')} \quad (if^R) \quad \frac{S_r : (pts, reg) \rightarrow (pts, reg)}{while \ b \ do \ S_r : (pts, reg) \rightarrow (pts, reg)} \quad (whl^R)}{\frac{(pts'_1, reg'_1) \leq (pts_1, reg_1) \quad S : (pts_1, reg_1) \rightarrow (pts_2, reg_2) \quad (pts_2, reg_2) \leq (pts'_2, reg'_2)}{S : (pts'_1, reg'_1) \rightarrow (pts'_2, reg'_2)} \quad (csq^R)}
\end{array}$$

The following lemma is needed in the proof of the following theorem which proves the soundness of the type system.

Lemma 5. *Suppose that $(s, h) \models (pts, reg)$, $\llbracket d \rrbracket = (\alpha, \beta)$, and $d : (pts, reg) \rightarrow Rs$. Then*

1. $\beta \in \mathcal{R} \implies \beta \in Rs$.
2. $\beta = \perp \implies Rs = \mathcal{R} = \{1, \dots, \gamma\}$.

Proof. The proof is by induction on the structure of d as follows:

1. If $d = n$, then by definition $\beta = \perp$ and $Rs = \mathcal{R}$ as required.
2. If $d = x$, then $\beta = s_r(x)$ and the required holds because $(s, h) \models reg$.
3. If $d = e : R_j$ or $d = Cast(R_i \hookrightarrow R_j) e$ then by definition $\beta = \{j\}$ and $Rs = \{j\}$ as required.
4. If $d = e_1 \oplus e_2$, then there are three subcases:
 - (a) α is an integer and $\beta = \perp$. In this case $\llbracket e_1 \rrbracket = (\alpha_1, \perp)$, $\llbracket e_2 \rrbracket = (\alpha_2, \perp)$, and $\alpha = \alpha_1 \oplus \alpha_2$, where α_1 and α_2 are integers. Therefore by the induction hypothesis $V_1 = V_2 = \mathcal{R}$. Hence $\mathcal{R} \subseteq Rs \subseteq \mathcal{R}$ implying $Rs = \mathcal{R}$.

- (b) α is an integer and $\beta \in \mathcal{R}$. In this case $\llbracket e_1 \rrbracket = (\alpha_1, \beta)$, $\llbracket e_2 \rrbracket = (\alpha_2, \perp)$, and $\alpha = \alpha_1 \oplus \alpha_2$, where α_1 and α_2 are integers. Therefore by the induction hypothesis $\beta \in V_1 \cap V_2 \subseteq Rs$.
- (c) α is address. Then by Lemma 11 $\beta \in \mathcal{R}$ and $\beta = h_r(\alpha)$. In this case, $\beta \in \text{reg}(\alpha)$ because $(s, h) \models \text{reg}$ and $\alpha \in V'$ because $(s, h) \models \text{pts}$. Therefore $\beta \in \cup_{a \in V'} \text{reg}(a) \subseteq RS$.

Theorem 2. 1. $(\text{pts}, \text{reg}) \leq (\text{pts}', \text{reg}') \implies (\forall (s, h), (s, h) \models (\text{pts}, \text{reg}) \implies (s, h) \models (\text{pts}', \text{reg}'))$.

2. $(S : (\text{pts}, \text{reg}) \rightarrow (\text{pts}', \text{reg}')) \implies (S : \text{pts} \rightarrow \text{pts}')$.

3. Suppose that $S : (\text{pts}, \text{reg}) \rightarrow (\text{pts}', \text{reg}')$ and $S : (s, h) \rightarrow (s', h')$. Then $(s, h) \models (\text{pts}, \text{reg})$ implies $(s', h') \models (\text{pts}', \text{reg}')$.

Proof. The first two items are obvious. For the last item and by (2), it is enough to prove that $(s', h') \models \text{reg}'$. This is proved by induction on the structure of type derivation as follows:

1. The type derivation has the form (*ass*^R). In this case, $\text{reg}' = [\text{reg} \mid x : Rs]$ and $(s', h') = ([s_v \mid x : \alpha], [s_r \mid x : \beta], h)$, where $\llbracket e \rrbracket(s, h) = (\alpha, \beta)$. By 2 and Theorem 11 $(s', h') \models \text{pts}'$. By Lemma 5 $(s', h') \models \text{reg}'$. Clearly $\text{dom}(\text{pts}') = \text{dom}(\text{reg}')$ and hence $(s', h') \models (\text{pts}', \text{reg}')$.
2. The type derivation has the form (*con*^R). In this case, $\text{reg}' = \cup_{1 \leq i \leq v} [\text{reg} \mid x : Rs_1 \mid a_{n,1}^i : Rs_1 \mid \dots \mid a_{n,n}^i : Rs_n]$ and $(s', h') = ([s_v \mid x : a_{n,1}^u], [s_r \mid x : \beta_1], [h_v \mid a_{n,1}^u : \alpha_1 \mid \dots \mid a_{n,n}^u : \alpha_n], [h_r \mid a_{n,1}^u : \beta_1 \mid \dots \mid a_{n,n}^u : \beta_n])$. Clearly, $1 \leq u \leq v$. For every $1 \leq i \leq n$ by Lemma 5 if $\beta_i \in \mathcal{R}$ then $\beta_i \in Rs_i$ and if $\beta_i = \perp$ then $Rs_i = \mathcal{R}$. We have $s'_r(x) = \beta_1 \in Rs_1 = \text{reg}'(x)$. We also have that $\text{dom}(h') \subseteq \text{dom}(\text{reg}')$ because $\text{dom}(h) \subseteq \text{dom}(\text{reg})$ ($(s, h) \models \text{reg}$) and $1 \leq u \leq v$. It is obvious that for any $x \neq y \in \text{Var}$ and $a \in \text{dom}(h') \setminus \{a_{n,1}^u, \dots, a_{n,n}^u\}$,
 - $s'_r(y) \in \mathcal{R}$ implies $s'_r(y) \in \text{reg}'(y)$,
 - $s'_r(y) = \perp$ implies $\text{reg}'(y) = \mathcal{R}$, and
 - $h'_r(a) \in \mathcal{R}$ implies $h'_r(a) = h_r(a) \in \text{reg}(a) \subseteq \text{reg}'(a)$.

For every $1 \leq i \leq n$, if $h_r(a_{n,i}^u) \in \mathcal{R}$, then $h_r(a_{n,i}^u) = \beta_i \in Rs_i \subseteq \text{reg}'(a_{n,i}^u)$. Hence $(s', h') \models \text{reg}'$.

3. The type derivation has the form (*lok*^R). In this case, $\text{reg}' = [\text{reg} \mid x : Rs]$ and $(s', h') = ([s_v \mid x : h_v(\alpha)], [s_r \mid x : \beta], h)$, where $\llbracket e \rrbracket(s, h) = (\alpha, \beta)$. By Lemma 5 $\beta \in Rs$. Also we have $\alpha \in \text{Addrs} \cap \text{dom}(h)$ and hence $\alpha \in V'$ by Lemma 4.
4. The type derivation has the form (*mut*^R). In this case, $\text{reg}' = \cup_{a \in V'} [\text{reg} \mid a : Rs]$ and $(s', h') = (s, [h_v \mid \alpha_1 : \alpha_2], [h_r \mid \alpha_1 : \beta])$, where $\llbracket e_i \rrbracket(s, h) = (\alpha_i, \beta)$. We have $\alpha_1 \in \text{dom}(h) \cap V_1$ and $\beta \in Rs$ by Lemma 5. Therefore $h_r(\alpha_1) \in \text{reg}'(\alpha_1)$.

The remaining cases are straightforward to check.

5 Data Slicing

This section presents a technique for solving the principal problem, heap slicing, motivating the paper. The basic instrument of the technique is a type system which is an

enrichment of the type system for region analysis with a transformation component. This transformation is that of heap slicing. In this section it is also shown that the transformation presented by the type system is sound in the sense that the original program and that results from the transformation produce the same result.

Definition 5. A sliced heap (s, \tilde{h}) is a valid slicing of a state (s, h) , denoted by $(s, h) \sim (s, \tilde{h})$, if

1. $\text{dom}(h) = \text{dom}(\tilde{h}_1)$, and
2. $(\forall a \in \text{dom}(h)) (h_v(a), h_r(a)) = (\alpha, \beta) \implies \tilde{h}_\beta(a) = \alpha$ and $(\forall i \neq \beta) h_i(a) = \phi$.

Definition 6. 1. $\text{Slice} : \text{Heaps} \rightarrow \text{Sliced Heaps} : h \mapsto (h_1, \dots, h_\gamma)$, where for every $i \in [1, \gamma]$,

$$h_i : \text{dom}(h) \rightarrow \text{Values}^+ : a \mapsto \begin{cases} h_v(a), & \text{if } h_r(a) = i; \\ \phi, & \text{otherwise.} \end{cases}$$

2. $\text{Con} : \text{Sliced Heaps} \rightarrow \text{Heaps} : \tilde{h} \mapsto (h_v, h_r)$, where

$$h_v : \text{dom}(\tilde{h}) \rightarrow \text{Values} : a \mapsto \tilde{h}_{i_a}(a) \quad h_r : \text{dom}(\tilde{h}) \rightarrow \mathcal{R} : a \mapsto i_a, \text{ where}$$

i_a is the unique index such that $\tilde{h}_{i_a}(a) \neq \phi$.

3. $\text{Slice}_S : \text{States} \rightarrow \text{Sliced States} : (s, h) \mapsto (s, \text{Slice}(H))$.
4. $\text{Cons}_S : \text{Sliced States} \rightarrow \text{States} : (s, \tilde{h}) \mapsto (s, \text{Con}(\tilde{H}))$

Lemma 6. The maps of the previous definitions are well-defined. Moreover Slice_S and Cons_S are inverses to each other.

The inference rules of our type system are the following:

$$\begin{array}{c} \rho(d_i, R_{S_i}) = \begin{cases} d_i : R_{S_i}, & \text{if } d_i = e_i; \\ d_i, & \text{otherwise.} \end{cases} \quad \frac{}{x := e : (pts, reg) \rightarrow (pts', reg') \hookrightarrow x := e} \quad \frac{}{skip : (pts, reg) \rightarrow (pts, reg) \hookrightarrow skip} \\ \\ x := \text{cons}(d_1, \dots, d_n) : (pts, reg) \rightarrow (pts', reg') \quad \forall 1 \leq i \leq n. d_i : (pts, reg) \rightarrow R_{S_i} \\ \hline x := \text{cons}(d_1, \dots, d_n) : (pts, reg) \rightarrow (pts', reg') \hookrightarrow x := \text{cons}(\rho(d_1, R_{S_1}), \dots, \rho(d_n, R_{S_n})) \\ \\ \frac{x := [e] : (pts, reg) \rightarrow (pts', reg')}{e : (pts, reg) \rightarrow R_S} \quad \frac{}{dispose(e) : (pts, reg) \rightarrow (pts, reg') \hookrightarrow dispose'(e)} \\ \hline x := [e] : (pts, reg) \rightarrow (pts', reg') \hookrightarrow x :=_{R_S} [e] \quad \frac{[e_1] := e_2 : (pts, reg) \rightarrow (pts', reg') \quad e_2 : (pts, reg) \rightarrow R_{S_2}}{[e_1] := e_2 : (pts, reg) \rightarrow (pts', reg') \hookrightarrow [e_1] :=_{R_{S_1} \cap R_{S_2}} e_2} \\ \\ \frac{S_1 : (pts, reg) \rightarrow (pts'', reg'') \hookrightarrow S'_1 \quad S_2 : (pts'', reg'') \rightarrow (pts', reg') \hookrightarrow S'_2}{S_1; S_2 : (pts, reg) \rightarrow (pts', reg') \hookrightarrow S'_1; S'_2} \quad \frac{S_f : (pts, reg) \rightarrow (pts, reg) \hookrightarrow S'_f}{\text{while } b \text{ do } S_f : (pts, reg) \rightarrow (pts, reg') \hookrightarrow \text{while } b \text{ do } S'_f} \\ \\ \frac{S_f : (pts, reg) \rightarrow (pts', reg') \hookrightarrow S'_f \quad S_g : (pts'', reg'') \rightarrow (pts', reg') \hookrightarrow S'_g}{\text{if } b \text{ then } S_f \text{ else } S_g : (pts, reg) \rightarrow (pts', reg') \hookrightarrow \text{if } b \text{ then } S'_f \text{ else } S'_g} \\ \\ \frac{(pts'_1, reg'_1) \leq (pts_1, reg_1) \quad S : (pts_1, reg_1) \rightarrow (pts_2, reg_2) \hookrightarrow S' \quad (pts_2, reg_2) \leq (pts'_2, reg'_2)}{S : (pts'_1, reg'_1) \rightarrow (pts'_2, reg'_2) \hookrightarrow S'} \end{array}$$

Theorem 3. (*Soundness*) Suppose that $S : (pts, reg) \rightarrow (pts', reg') \hookrightarrow S'$ and $(s, h) \sim (s, \tilde{h})$. Then

1. If $S : (s, h) \rightarrow (s', h')$, then there exists a state (s', \tilde{h}') such that $S' : (s, \tilde{h}) \rightsquigarrow (s', \tilde{h}')$ and $(s', h') \sim (s', \tilde{h}')$.
2. If $S' : (s, \tilde{h}) \rightsquigarrow (s', \tilde{h}')$, then there exists a state (s', h') such that $S : (s, h) \rightarrow (s', h')$ and $(s', h') \sim (s', \tilde{h}')$.

Proof. The proof is by induction on the structure of type derivation. For the base cases in the proof of (1), take $(s', \tilde{h}') = Slice_s((s', h'))$. For the base cases in the proof of (2), take $(s', h') = Con_s((s', \tilde{h}'))$.

6 Related and Future Work

In [6], Condit et al. present data slicing [28,31], a program transformation which divides the heap into separate regions, for a C-like language. The basic idea in [6] is to syntactically slice structures defined in a given program. Then, the slicing of the program commands is calculated using sliced versions of program structures. The physical slicing of the program heap follows upon executing the sliced program.

Related concepts to data slicing are program slicing, intentional polymorphism, structure splitting. Program slicing [11,18,4,26] finds the program portions that contribute to evaluating the value of a given variable at a given program point. In other words, program slicing [25] is a practicable technique to bound the focus of a job to certain part of a program. Program slicing is used in program comprehension, testing, restructuring, debugging, and optimizing. A technique to compile polymorphism while still being able to use types information at run time is intentional polymorphism [7,17,8]. The similarity to data slicing comes from the fact that intentional polymorphism enables the compiler of preserving type safety and efficiently representing types. An alternative approach to data slicing, is structure splitting [25]. This approach marks the non-active fields of data structures by adding new pointers to data structures. Clearly this pointer addition does sacrifices the backward compatibility. Therefore data slicing is advantageous over structure splitting.

Among advantages of data slicing is preserving backward compatibility. As an alternative, splay trees [30,21,27] can be used to preserve backward compatibility. However some research like [6] concludes that the use of splay trees is more expensive in terms of time and complexity of the system used in implementation.

A typical approach for heap slicing is the algorithmic style. However the use of type systems in program analysis (in general) [13,11,16,12], rather than classical algorithms, and in data slicing (in particular) is very useful for applications like certified code or proof-carrying code. The catch of the type systems approach is that type derivations serve as proofs for the technique result.

Programs and data structures can mathematically be represented by mathematical domains and maps between domains. This representation is called denotational semantics of programs. An important direction for future research is to transfer concepts of data and program slicing to the side of denotational semantics [15,10]. This enables us to mathematically study in deep heap slicing and translates back obtained results to the side of programs and data structures.

References

1. Barraclough, R.W., Binkley, D., Danicic, S., Harman, M., Hierons, R.M., Kiss, Á., Laurence, M., Ouarbya, L.: A trajectory-based strict semantics for program slicing. *Theor. Comput. Sci.* 411(11-13), 1372–1386 (2010)
2. Carrillo, S., Siegel, J., Li, X.: A control-structure splitting optimization for gpgpu. In: Johnson, G., Trinitis, C., Gaydadjiev, G., Veidenbaum, A.V. (eds.) *Conf. Computing Frontiers*, pp. 147–150. ACM (2009)
3. Chen, C.-L., Lin, S.-H.: Formulating and solving a class of optimization problems for high-performance gray world automatic white balance. *Appl. Soft Comput.* 11(1), 523–533 (2011)
4. Cheney, J.: Program slicing and data provenance. *IEEE Data Eng. Bull.* 30(4), 22–28 (2007)
5. Chilimbi, T.M., Davidson, B., Larus, J.R.: Cache-conscious structure definition. In: *PLDI*, pp. 13–24 (1999)
6. Condit, J., Necula, G.C.: Data Slicing: Separating the Heap into Independent Regions. In: Bodik, R. (ed.) *CC 2005. LNCS*, vol. 3443, pp. 172–187. Springer, Heidelberg (2005)
7. Cray, K., Weirich, S., Gregory Morrisett, J.: Intensional polymorphism in type-erasure semantics. *J. Funct. Program.* 12(6), 567–600 (2002)
8. Duggan, D.: Dynamic typing for distributed programming in polymorphic languages. *ACM Trans. Program. Lang. Syst.* 21(1), 11–45 (1999)
9. El-Zawawy, M., Daoud, N.: New error-recovery techniques for faulty-calls of functions. *Computer and Information Science* 4(3) (May 2012)
10. El-Zawawy, M.A.: Semantic spaces in Priestley form. PhD thesis, University of Birmingham, UK (January 2007)
11. El-Zawawy, M.A.: Flow Sensitive-Insensitive Pointer Analysis Based Memory Safety for Multithreaded Programs. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) *ICCSA 2011, Part V. LNCS*, vol. 6786, pp. 355–369. Springer, Heidelberg (2011)
12. El-Zawawy, M.A.: Probabilistic pointer analysis for multithreaded programs. *ScienceAsia* 37(4) (December 2011)
13. El-Zawawy, M.A.: Program optimization based pointer analysis and live stack-heap analysis. *International Journal of Computer Science Issues* 8(2) (March 2011)
14. El-Zawawy, M.A.: Dead code elimination based pointer analysis for multi-threaded programs. *Journal of the Egyptian Mathematical Society* (January 2012), doi:10.1016/j.joems.2011.12.011
15. El-Zawawy, M.A., Jung, A.: Priestley duality for strong proximity lattices. *Electr. Notes Theor. Comput. Sci.* 158, 199–217 (2006)
16. El-Zawawy, M.A., Nayel, H.A.: Partial redundancy elimination for multi-threaded programs. *IJCSNS International Journal of Computer Science and Network Security* 11(10) (October 2011)
17. Harper, R., Gregory Morrisett, J.: Compiling polymorphism using intensional type analysis. In: *POPL*, pp. 130–141 (1995)
18. Gaikovina Kula, R., Fushida, K., Kawaguchi, S., Iida, H.: Analysis of Bug Fixing Processes Using Program Slicing Metrics. In: Ali Babar, M., Vierimaa, M., Oivo, M. (eds.) *PROFES 2010. LNCS*, vol. 6156, pp. 32–46. Springer, Heidelberg (2010)
19. George, C.: Proof-carrying code. In: Henk, C., van Tilborg, H.C.A., Jajodia, S. (eds.) *Encyclopedia of Cryptography and Security*, 2nd edn., pp. 984–986. Springer (2011)
20. Nielson, F., Nielson, H.R., Hankin, C.L.: *Principles of Program Analysis*. Springer (1999); second printing (2005)
21. Pettie, S.: Splay trees, davenport-schinzel sequences, and the deque conjecture. In: Teng, S.-H. (ed.) *SODA*, pp. 1115–1124. SIAM (2008)

22. Pfenning, F., Caires, L., Toninho, B.: Proof-Carrying Code in a Session-Typed Process Calculus. In: Jouannaud, J.-P., Shao, Z. (eds.) CPP 2011. LNCS, vol. 7086, pp. 21–36. Springer, Heidelberg (2011)
23. Prasad, S., Arun-Kumar, S.: Introduction to operational semantics. In: The Compiler Design Handbook, pp. 841–890 (2002)
24. Reynolds, J.C.: Separation logic: A logic for shared mutable data structures. In: Symposium on Logic in Computer Science, p. 55 (2002)
25. Sasirekha, N., Edwin Robert, A., Hemalatha, M.: Program slicing techniques and its applications. CoRR, abs/1108.1352 (2011)
26. Tip, F.: A survey of program slicing techniques. *J. Prog. Lang.* 3(3) (1995)
27. Weiser, M.: Program slicing. *IEEE Trans. Software Eng.* 10(4), 352–357 (1984)
28. Xin, B., Zhang, X.: Memory slicing. In: Rothermel, G., Dillon, L.K. (eds.) ISSTA, pp. 165–176. ACM (2009)
29. Ye, X., Li, P.: Parallel program performance modeling for runtime optimization of multi-algorithm circuit simulation. In: Sapatnekar, S.S. (ed.) DAC, pp. 561–566. ACM (2010)
30. Zhang, S., Cui, Z., Gong, S.-R., Liu, Q., Fan, J.-X.: A data aggregation algorithm based on splay tree for wireless sensor networks. *JCP* 5(4), 492–499 (2010)
31. Zhang, X., Gupta, R., Zhang, Y.: Cost and precision tradeoffs of dynamic data slicing algorithms. *ACM Trans. Program. Lang. Syst.* 27(4), 631–661 (2005)

Using Autonomous Search for Generating Good Enumeration Strategy Blends in Constraint Programming

Ricardo Soto¹, Broderick Crawford¹, Eric Monfroy², and Víctor Bustos¹

¹ Pontificia Universidad Católica de Valparaíso, Chile

² CNRS, LINA, Université de Nantes, France

{broderick.crawford,ricardo.soto}@ucv.cl,

{victor.bustos.a}@mail.pucv.cl,

{eric.monfroy}@univ-nantes.fr

Abstract. In Constraint Programming, enumeration strategies play an important role, they can significantly impact the performance of the solving process. However, choosing the right strategy is not simple as its behavior is commonly unpredictable. Autonomous search aims at tackling this concern, it proposes to replace bad performing strategies by more promising ones during the resolution. This process yields a combination of enumeration strategies that worked during the search phase. In this paper, we focus on the study of this combination by carefully tracking the resolution. Our preliminary goal is to find good enumeration strategy blends for a given Constraint Satisfaction Problem.

Keywords: Artificial Intelligence, Constraint Programming, Autonomous Search.

1 Introduction

Constraint Programming (CP) is a powerful programming paradigm devoted to the efficient resolution of constraint-based problems. It gathers and combines ideas from different domains, among others, from operational research, numerical analysis, artificial intelligence, and programming languages. Currently, CP is largely used in different application areas, for instance, in computer graphics to express geometric coherence, in engineering design for the conception of complex mechanical structures, in database systems to ensure and/or restore data consistency, in electrical engineering to locate faults, and even for sequencing the DNA in molecular biology.

In a CP context, a problem is formulated as a Constraint Satisfaction Problem (CSP), which is a formal representation mainly consisting in a sequence of variables lying in a domain and a set of constraints. The goal is to find a complete variable-instantiation that satisfies the whole set of constraints. The basic idea for solving CSPs is to build a tree data structure holding the potential solutions by using a backtracking-based procedure. In general, two main phases

are involved: enumeration and propagation. The enumeration phase instantiates variables in order to create branches of the tree. The propagation phase tries to prune the tree by filtering from domains the values that do not lead to any solution. This is possible by using the so-called consistency properties [12].

In the enumeration phase, there are two important decisions to be made: the order in which the variables and values are selected. This selection refers to the variable and value ordering heuristics, and jointly constitutes the enumeration strategy. Enumeration strategies are known to be a key component of the resolution, in fact choosing the right one can dramatically impact the solving process. However, making the correct decision is not simple as the behavior of a given strategy is commonly unpredictable. Autonomous Search (AS) aims at tackling this concern [11], it proposes to replace bad performing strategies –known as dynamic or “on the fly” replacement– by more promising ones during the resolution.

The integration of AS in CP is indeed a recent trend, only a few works have been reported, which are mostly centered on defining theoretical frameworks [4] and on architectures [8,6] for performing dynamic replacements. However, little is known about which is the best combination of enumeration strategies for a given problem. In this paper, we focus on the study of this combination, namely enumeration strategy blends, by carefully tracking the resolution. We pay special attention in which part of the process the strategies participate, and how long they act. Our preliminary goal is to illustrate good strategy blends for a set of well-known benchmarks.

This paper is organized as follows. Section 2 presents the basic notions of CP and CSP solving. The related work is presented in Section 3. The AS+CP framework is described in Section 4. Experiments are presented in Section 5, followed by the conclusion and future work.

2 Constraint Programming

As previously mentioned, in a CP context, problems are formulated as CSPs. Formally, a CSP \mathcal{P} is defined by a triple $\mathcal{P} = \langle \mathcal{X}, \mathcal{D}, \mathcal{C} \rangle$ where:

- \mathcal{X} is an n -tuple of variables $\mathcal{X} = \langle x_1, x_2, \dots, x_n \rangle$.
- \mathcal{D} is a corresponding n -tuple of domains $\mathcal{D} = \langle D_1, D_2, \dots, D_n \rangle$ such that $x_i \in D_i$, and D_i is a set of values, for $i = 1, \dots, n$.
- \mathcal{C} is an m -tuple of constraints $\mathcal{C} = \langle C_1, C_2, \dots, C_m \rangle$, and a constraint C_j is defined as a subset of the Cartesian product of domains $D_{j_1} \times \dots \times D_{j_{n_j}}$, for $j = 1, \dots, m$.

A solution to a CSP is an assignment $\{x_1 \rightarrow a_1, \dots, x_n \rightarrow a_n\}$ such that $a_i \in D_i$ for $i = 1, \dots, n$ and $(a_{j_1}, \dots, a_{j_{n_j}}) \in C_j$, for $j = 1, \dots, m$.

Algorithm 1 represents a general procedure for solving CSPs. The goal is to iteratively generate partial solutions, backtracking when an inconsistency is detected, until a result is reached. The algorithm begins by loading the CSP

model. Then, a while loop encloses a set of actions to be performed until fixing all the variables (i.e. assigning a consistent value) or a failure is detected (i.e. no solution is found). The first two enclosed actions correspond to the variable and value selection. The third action is a call to a propagation procedure, which is responsible for attempting to prune the tree. Finally two conditions are included to perform backtracks. A shallow backtrack corresponds to try the next value available from the domain of the current variable, and the backtracking returns to the most recently instantiated variable that has still values to reach a solution.

Algorithm 1. A general procedure for solving CSPs

```

1: load_CSP()
2: while not all_variables_fixed or failure do
3:   heuristic_variable_selection()
4:   heuristic_value_selection()
5:   propagate()
6:   if empty_domain_in_future_variable() then
7:     shallow_backtrack()
8:   end if
9:   if empty_domain_in_current_variable() then
10:    backtrack()
11:  end if
12: end while

```

3 Related Work

In CP, selecting in advance an appropriate enumeration strategy is known to be quite complex, since its effects are hard to predict. During the last years, there is a trend to analyze the state of progress of the solving process in order to automatically identify strategies that work well. For instance, the Adaptive Constraint Engine (ACE) [9] is a framework that learns ordering heuristics by gathering the experience from problem solving processes. The idea of this approach is to manage a set of advisors that recommend, in the form of comment, a given action to perform. The reliability and utility of advisors is controlled by weights, which are determined by a DWL (Digression-based Weight Learning) algorithm. The algorithm learns by examining the solution's trace of problems successfully solved. The final decision is computed as a weighted combination of the comments done by the advisors in a process called voting.

Another interesting approach following a similar goal is the weighted degree heuristic [3]. The idea is to associate weights to constraints, which are incremented during propagation whenever this causes a domain wipeout. The sum of weights is computed for each variable involved in constraints and the variable with the largest sum is selected.

The random probing method [10,14] proposes two changes to the weighted degree heuristic. On one hand, the initial choices are made without information

on edge weights, and on the other, the weighted degree is biased by the path of the search. This makes the approach too sensitive to local instead of to global conditions of failure. The random probing method proposes to perform sampling during an initial gathering phase arguing that initial choices are often the most important.

The integration of AS in CP is another approach to tackle the aforementioned concern. It proposes an “on the fly” replacement of bad performing strategies by more promising ones during the resolution. In this context, only a few works have been reported, which have mostly been centered on defining theoretical frameworks [4] and on architectures [8,6]. However, the current knowledge about which is the best combination of enumeration strategies for a given problem is very limited. In this paper, we focus on that topic. Indeed, our preliminary goal is to illustrate good strategy blends for a set of well-known benchmarks.

4 AS+CP Framework

The state-of-the-art framework for performing AS in CP roughly consists in 4 components (see Figure 1): SOLVE, OBSERVATION, ANALYSIS and UPDATE.

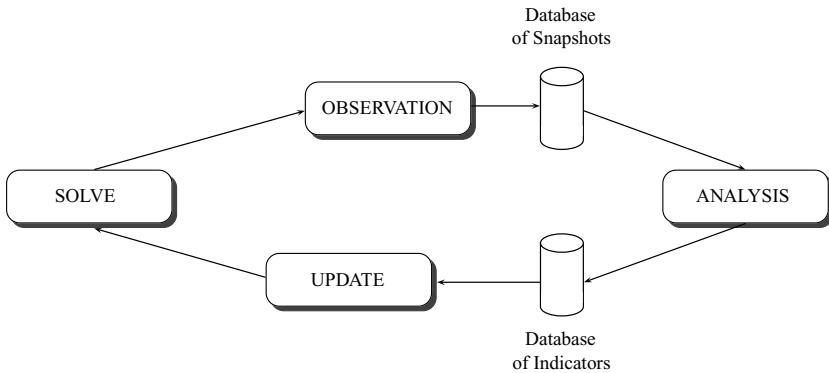


Fig. 1. Framework schema

- The SOLVE component runs a generic CSP solving algorithm performing a depth-first search by alternating constraint propagation with enumeration phases. SOLVE has a set of basic enumeration strategies each one characterized by a priority that evolves during computation: the UPDATE component

evaluate strategies and update their priorities. For each enumeration, the dynamic enumeration strategy selects the basic strategy to be used based on the attached priorities. SOLVE is also able to perform metabacktracks (jump back of a sequence of several enumerations and propagation phases) in order to repair a “desperate” state of resolution, i.e., when changing strategies is not sufficient due to several very bad previous choices.

- The OBSERVATION component aims at regarding and recording some information about the current search tree, i.e., it spies the resolution process in the SOLVE component. These observations (called snapshots) are not performed continuously, and they can be seen as an abstraction of the resolution state at a time t . Taking a snapshot consists in extracting (since search trees are too large) and recording some information from a resolution state.
- The ANALYSIS component studies the snapshots taken by the OBSERVATION. It evaluates the different strategies, and provides indicators (described in Table 1) to the UPDATE component. Indicators can be extracted, computed, or deduced from one or several snapshots from the database of snapshots.
- The UPDATE component makes decisions using a special choice function. The choice function determines the performance of a given strategy in a given amount of time. It is calculated based on the indicators given by the ANALYSIS component and a set of control parameters computed by a genetic algorithm (details about the genetic algorithm used can be seen in [5,8]).

Table 1. Search process indicators

Name	Description
VFP	Number of variables fixed by propagation
n	Number of steps or decision points (n increments each time a variable is fixed enumeration)
$T_n(S_j)$	Number of steps since the last time that an enumeration strategy S_j was used until step n^{th}
SB	Number of Shallow Backtracks [2]
B	Number of Backtracks
In1	Represents a Variation of the Maximum Depth. It is calculated as: $CurrentMaximumDepth - PreviousMaximumDepth$
In2	Calculated as: $CurrentDepth - PreviousDepth$. A positive value means that the current node is deeper than the one explored at the previous step
B-real	Number of backtracks considering also the number of shallow backtracks
d	Current depth in the search tree
Thrash	The solving process alternates enumerations and backtracks on a few number of variables without succeeding in having a strong orientation. It is calculated as: $d_{t-1} - VFP_{t-1}$

4.1 Choice Function

The choice function [13] attempts to capture the correspondence between the historical performance of each enumeration strategy and the decision point currently being investigated. Here, a decision point or step is every time the solver is invoked to fix a variable by enumeration.

The choice function is used to rank and choose between different enumeration strategies at each step. For any enumeration strategy S_j , the choice function f in step n for S_j is defined by equation 1, where l is the number of indicators considered and α is a parameter to control the relevance of the indicator within the choice function.

$$f_n(S_j) = \sum_{i=1}^l \alpha_i f_{i_n}(S_j) \quad (1)$$

Additionally, to control the relevance of an indicator i for an strategy S_j in a period of time, a popular statistical technique –called exponential smoothing– is used for producing smoothed time series. The idea is to associate, for some indicators, greater importance to recent performance by exponentially decreasing weights to older observations. In this way, recent observations give relatively more weight than older ones. The exponential smoothing is applied to the computation of $f_{i_n}(S_j)$, which is defined by equations 2 and 3, where x_0 is the value of the indicator i for the strategy S_j in time 1, n is a given step of the process, β is the smoothing factor, and $0 < \beta < 1$

$$f_{i_1}(S_j) = x_0 \quad (2)$$

$$f_{i_n}(S_j) = x_{n-1} + \beta f_{i_{n-1}}(S_j) \quad (3)$$

Let us note that the speed at which the older observations are smoothed (dampened) depends on β . When β is close to 0, dampening is quick and when it is close to 1, dampening is slow.

The general solving procedure including AS can be seen in Algorithm 2. Three new function calls have been included: for calculating the indicators (line 11), the choice function (line 12), and for choosing promising strategies (line 13), that is, the ones with highest choice function 1. They are called after constraint propagation to compute the real effects of the strategy (some indicators may be impacted by the propagation).

¹ When strategies have the same score, one is selected randomly.

Algorithm 2. A procedure for solving CSPs including autonomous search

```

1: while not all_variables_fixed or failure do
2:   heuristic_variable_selection()
3:   heuristic_value_selection()
4:   propagate()
5:   if empty_domain_in_future_variable() then
6:     shallow_backtrack()
7:   end if
8:   if empty_domain_in_current_variable() then
9:     backtrack()
10:  end if
11:  calculate_indicators()
12:  calculate_choice_function()
13:  enum_strategy_selection()
14: end while

```

5 Experiments

We have performed a set of experiments in order to identify the best strategy blends for a set of well-known benchmarks. Experiments have been performed on a 2.33GHz Intel Core2 Duo with 2Gb RAM running Windows XP, and the benchmarks are: N-Queens ($N=\{8,10,12,15,20,50,75\}$), 10 and 20 Linear Equations, Magic Squares ($N=\{3,4,5\}$), Sudoku, Knight Tournament ($N=\{5,6\}$). A portfolio of 8 enumerations strategies has been used, which is detailed in Table 2.

Figure 2 depicts a chart illustrating the best strategy blend found by the AS+CP framework for the 8-Queens problem. X-axis denotes the strategy id and Y-axis denotes the percentage of solving time that the strategy acts until is replaced (e.g. firstly, strategy 7 participated about 8% of the solving time, then strategy 6 participated about 2%, then strategy 3 participated about 2%,

Table 2. Portfolio used

Id	Variable ordering	Value ordering
1	First variable of the list	min. value in domain
2	The variable with the largest domain	min. value in domain
3	The variable with the smallest domain	min. value in domain
4	The variable with the largest number of attached constraints	min. value in domain
5	First variable of the list	max. value in domain
6	The variable with the largest domain	max. value in domain
7	The variable with the smallest domain	max. value in domain
8	The variable with the largest number of attached constraints	max. value in domain

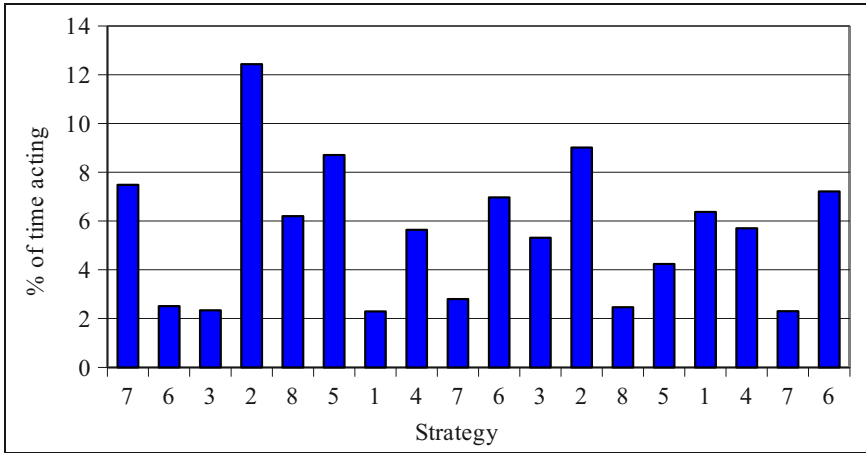


Fig. 2. Strategy blend for the 8-queens problem

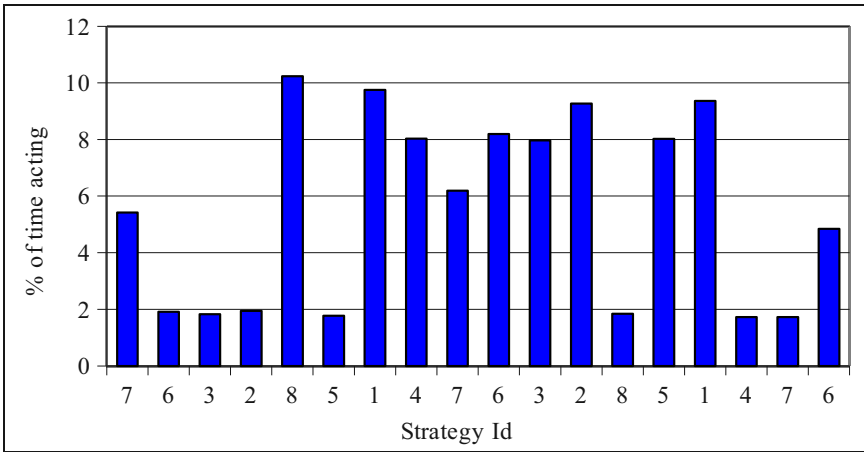


Fig. 3. Strategy blend for the 10-queens problem

and so on). In Figure 3, the 10-Queens problem exhibits a similar sequence of strategies involved in the process, however having different percentages of time acting.

Figure 4 illustrates the best strategy blend found for the Magic Squares problem (N=3), here only 3 strategies participated, strategy 7 being the more active. For space reasons, we omit the complete set of experiments (bigger instances involve more than 1000 strategy replacements), however they are available at [1].

Table 3 depicts solving times in seconds for N-Queens (N={8,10,20,50,75}), Magic Squares (N={4,5}), and for Knight Tournament (N={5,6}). First row gives the best time reached by using a single strategy from the portfolio. Second

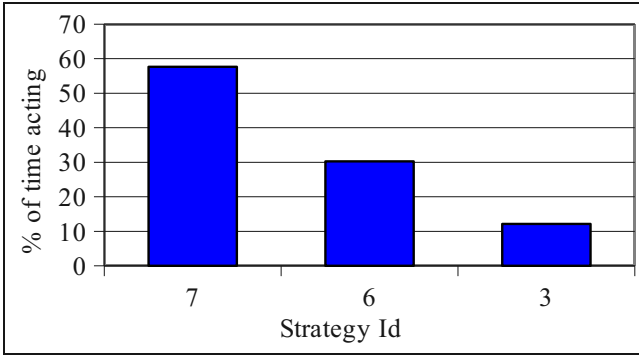


Fig. 4. Strategy blend for the Magic Squares problem (N=3)

Table 3. Solving times (in seconds).

	8-Q	10-Q	20-Q	50-Q	75-Q	4-MS	5-MS	5-Knight	6-Knight
Best strategy	0	0	0.031	1.031	8.562	0.015	0.516	2.578	t.o.
Average	0.006	0.006	13.644	t.o.	t.o.	0.534	t.o.	t.o.	t.o.
Best blend	1.89	1.89	7.875	24.343	49.859	2.203	2.875	7.422	114.906

row shows the average time of using single strategies from the portfolio. Third row includes the time reached by the best strategy blend including the cost of the choice function. The stop criterion is 65535 steps. A step corresponds to every time the solver is invoked to fix a variable by enumeration.

For smaller instances of the N-queens problem ($N=\{8,10\}$) as well as for Magic Squares ($N=4$) the overhead is nearly 2 seconds w.r.t. the average runtime (1.89-0.006=1.884 for 8-Q and 10-Q; and 2.203-0.534=1.69 for 4-MS). We estimate that such a cost is reasonable, considering the strong work done by the genetic algorithm (details about the cost of the choice function can be seen in [7]). However, for harder problems, the overhead begins to be less important, for instance for 20-Queens, the runtime of the strategy blend is about 7 seconds slower than the best time reached by a single strategy, but about 6 seconds faster than the average. For the N-Queens ($N=\{50,75\}$), Magic Squares ($N=5$), and for the Knight problem ($N=5$) the average time overpasses the stop criterion, while the best blend is one of the few able to solve them. Finally, the best blend is unique in solving the Knight problem ($N=6$), and as a consequence the only one that solves the complete set of problems.

6 Conclusion and Future Work

In this work, we have presented a preliminary study about enumeration strategy blends for CP. Such an study has been supported by a novel architecture, which consist in four components: SOLVE, OBSERVATION, ANALYSIS and

UPDATE. This framework allows one to introduce Autonomous Search to a common CP solving process. The idea is to replace bad performing strategies by more promising ones during the resolution. This process produces the so-called strategy blend, which is the combination of enumeration strategies that participated during the resolution. In this study, we have found the best strategy blends for a set of well-known benchmarks. The experiments exhibited that for small problems the cost of using Autonomous Search is noticeable but reasonable. However, for harder problems the overhead becomes less relevant.

The results presented here corresponds to ongoing work, and it they can certainly be extended by testing bigger instances and harder problems. Another interesting research direction is about the choice function, which can be implemented by using other optimization techniques (particle swarm, ant colony, etc).

References

1. Best Blends Experiments, http://www.inf.ucv.cl/~rsoto/best_blends (visited November 2011)
2. Barták, R., Rudová, H.: Limited assignments: A new cutoff strategy for incomplete depth-first search. In: Proceedings of the 20th ACM Symposium on Applied Computing (SAC), pp. 388–392 (2005)
3. Boussemart, F., Hemery, F., Lecoutre, C., Sais, L.: Boosting systematic search by weighting constraints. In: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI), pp. 146–150. IOS Press (2004)
4. Crawford, B., Castro, C., Monfroy, E.: Using a Choice Function for Guiding Enumeration in Constraint Solving. In: Proceedings of the 9th Mexican International Conference on Artificial Intelligence (MICAI), pp. 37–42. IEEE Computer Society (2010)
5. Crawford, B., Soto, R., Castro, C., Monfroy, E.: A Hyperheuristic Approach for Dynamic Enumeration Strategy Selection in Constraint Satisfaction. In: Ferrández, J.M., Álvarez Sánchez, J.R., de la Paz, F., Toledo, F.J. (eds.) IWINAC 2011, Part II. LNCS, vol. 6687, pp. 295–304. Springer, Heidelberg (2011)
6. Crawford, B., Soto, R., Castro, C., Monfroy, E.: Extensible CP-Based Autonomous Search. In: Stephanidis, C. (ed.) Posters, HCI 2011, Part I. CCIS, vol. 173, pp. 561–565. Springer, Heidelberg (2011)
7. Crawford, B., Soto, R., Castro, C., Monfroy, E., Paredes, F.: An Extensible Autonomous Search Framework for Constraint Programming. *Int. J. Phys. Sci.* 6(14), 3369–3376 (2010)
8. Crawford, B., Soto, R., Montecinos, M., Castro, C., Monfroy, E.: A Framework for Autonomous Search in the Eclⁱps^e Solver. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011, Part I. LNCS, vol. 6703, pp. 79–84. Springer, Heidelberg (2011)
9. Epstein, S.L., Freuder, E.C., Wallace, R.J., Morozov, A., Samuels, B.: The Adaptive Constraint Engine. In: Van Hentenryck, P. (ed.) CP 2002. LNCS, vol. 2470, pp. 525–542. Springer, Heidelberg (2002)
10. Grimes, D., Wallace, R.J.: Learning to identify global bottlenecks in constraint satisfaction search. In: Proceedings of the Twentieth International Florida Artificial Intelligence Research Society (FLAIRS) Conference, pp. 592–597. AAAI Press (2007)

11. Hamadi, Y., Monfroy, E., Saubion, F.: Special issue on autonomous search. *Constraint Programming Letters* 4 (2008)
12. Rossi, F., van Beek, P., Walsh, T.: *Handbook of Constraint Programming*. Elsevier (2006)
13. Soubeiga, E.: *Development and Application of Hyperheuristics to Personnel Scheduling*. PhD thesis, University of Nottingham School of Computer Science (2009)
14. Wallace, R.J., Grimes, D.: Experimental studies of variable selection strategies based on constraint weights. *J. Algorithms* 63(1-3), 114–129 (2008)

Evaluation of Normalization Techniques in Text Classification for Portuguese

Merley da Silva Conrado^{1,*}, Víctor Antonio Laguna Gutiérrez²,
and Solange Oliveira Rezende¹

¹ Sao Paulo University (USP),

P.O. Box 668, 13561-970, Sao Carlos-SP, Brazil,

² Pontifical Catholic University of Peru (PUCP),

P.O. Box 1761 Lima 32, Peru

{merleyc, solange}@icmc.usp.br,

victor.laguna@pucp.pe

Abstract. Text classification is an important task of Artificial Intelligence. Normally, this task uses large textual datasets whose representation is feasible because of normalization and selection techniques. In the literature, we can find three normalization techniques: stemming, lemmatization, and nominalization. Nevertheless, it is difficult to choose the most suitable technique for the text classification task. In this paper, we investigate this question experimentally by applying five different classifiers to four textual datasets in the Portuguese language. Additionally, the classification results are evaluated using unigrams, bigrams, and the combination of unigrams and bigrams. The results indicate that, in general, the number of terms obtained by each of the cases and the comprehensibility required in the results of the classification can be used as criteria to define the most suitable technique for the text classification task.

Keywords: Text classification, stemming, lemmatization, nominalization.

1 Introduction

Text Mining (TM) is a subarea of Artificial Intelligence defined as “a set of techniques and processes to discover innovative knowledge in the texts” [10]. According to Miner et al. [20], TM can be applied for seven main purposes: search and information retrieval, document clustering, document classification, web mining, information extraction, natural language processing, and concept extraction.

* This research was supported by FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo (Proc. No. 2009/16142-3 and 2011/19850-9) and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (Proc. No. 144629/2009-7), Brazil.

In TM, the text classification task aims to automatically identify the main topics in a document and link this document to one or more predefined categories [10,29]. Text classification is a very important task because it helps to keep large textual datasets organized, thus making possible to carry out other automatic and manual tasks more quickly.

In order to perform the classification, it is necessary to work with terms that conceptually represent the textual datasets for a better document classification. Additionally, Nuijian et al. [22] showed that the choice of word importance may increase the classification algorithm speed and save their resource used significantly. In this paper, terms (or features) are defined by a sequence of simple or compound words, which represent a unique concept. When terms are composed of only one word, they are called unigrams or simple terms, and when they are composed of more words, they are called n-grams (compound terms or combinations). Examples of simple terms are: *inteligencia* (intelligence), *artificial*, *processo* (process). Examples of compound terms are: *inteligencia_artificial* (artificial intelligence) and *processo_mineracao_textos* (text mining process).

Normalization techniques may help in the term extraction task, because it could represent the dataset as well as assist in reducing the high dimension presented in the attribute-value matrix. The use of these techniques can also improve the outcome of the information retrieval task. Through normalization, these techniques recover the searched word and other words with the same normalization, whose original spelling is different and concept is the same.

In related literature, the following normalization techniques for Brazilian Portuguese¹ are found: stemming [16], lemmatization [2], and nominalization [12]. Most of the text classification studies use the stemming [13,21,27] and some others use lemmatization [26] and nominalization [30]. Also, the effects of the using of these three techniques in information retrieval task were analyzed [12]. However, up to this date, no studies were found evaluating the use of these techniques in the text classification task. Because of that, until now it was not possible to answer the following question: Which technique is most suitable for use in the text classification task? Aiming to answer it, this paper provides an unprecedented evaluation of the use of these three techniques for normalization in the text classification task for Brazilian Portuguese. Furthermore, the classification results are compared when using only unigrams, only bigrams or unigrams with bigrams. Some noteworthy proposals have already studied the use of unigrams and bigrams in the text classification task [4,13].

2 Technique for Normalization

Normalization is the search for patterns that reduce the various forms of words presentation in the textual dataset and maintains the essential meaning at the same time.

¹ Portuguese spelling reform went into effect in Brazil on January 1, 2009. However, this paper makes a distinction of Brazilian Portuguese with Portuguese from others countries, because textual datasets presented here use the old spelling.

Stemming reduces words to their inflexional form and sometimes reduces the words to their derivatives. It means that stemming eliminates the prefixes and suffixes of words and turns the verbs to their infinitive form [17].

According to Aranha [3], stemming can be seen as *inflectional stemming*, which considers only the removal of verbal inflections, or *stemming to the root* which removes all the forms of prefixes and suffixes of terms. *Stemming to the root* is the most aggressive form of stemming. As algorithms for the stemming of Brazilian Portuguese we can mention Porter - Portuguese [2], PortugueseStemmer [23], Pegastemming [3], and STEMBR [1].

Lemmatization [2] aims to combine the variants of a simple term in a unique lemma. Lemma is considered as a set of words with the same root and the same class lexical-grammatical. Basically, the verbs must be transformed into their infinity form and the nouns and adjectives must be transformed into their singular masculine forms. For the Portuguese language, there exists a lemmatizer, named Nunes [8], besides other tools that morphologically tag words, such as MXPOST [25], TreeTagger [28], and the BRILL tagger [5]. It should be noticed that after the use of these tools, it is necessary to apply the lemmatization technique in tagged words.

During **Nominalization**, words begin to display a syntactic/semantic behavior similar to nouns [4]. For Portuguese, there exists the FORMA tool [12] that morphologically tags the words of the document. After the tagging process, it is possible to apply the CHAMA tool [12] in the tagging resulting words, which is responsible for the normalization of the words.

In Table 1 there is an example of the application of each technique for normalization. In this example, stopwords were removed and all the characters were changed to their lowercase form.

Table 1. Examples of the application of each normalization technique

Original:	“Pesquisas descrevem perfil de estudantes inteligentes.”
Stemming:	pesquis descr perfí estud intelig
Lemmatization:	pesquisa descrever perfil estudante inteligente
Nominalization:	pesquisa descrever perfil estudante inteligencia

3 Evaluation of Normalization Techniques in Text Classification

Aiming to know which normalization technique is most suitable for the text classification task, we evaluated and discussed the aspects of three techniques experimentally running five different classifiers on four textual datasets in Brazilian Portuguese.

² Snowball - <http://snowball.tartarus.org/index.php>

³ Pegastemming - <http://www.inf.pucrs.br/~gonzalez/ri/pesqdiss/analise.htm>

⁴ Traditional Grammar - <http://www.dacex.ct.utfpr.edu.br/paulo3.htm>

3.1 Datasets

For the experiments, we used four textual datasets in Brazilian Portuguese, known as CSTNews, IFM, NN, and CIMM. The CSTNews dataset [18] consists of newspaper articles, the IFM dataset⁵ (Millennium Factory Institute) has articles related to Production Engineering, the NN dataset⁶ (Nanoscience and Nanotechnology), as its name suggests, contains texts of Nanoscience and Nanotechnology, and finally, the CIMM dataset⁷ (Info Center Metal Mechanics) has documents related to metalworking.

Table 2 shows, for each of these datasets, the number of documents, the number of classes, and the number of documents belonging to the minority and majority classes.

For the CSTNews dataset, which has a total of 140 documents, only 131 documents were used. Among the dismissed documents, the *money* and *science* classes were removed, because they contained only 3 and 2 texts, respectively. For the IFM dataset, which contains a total of 603 documents, 27 documents were removed because they had damaged contents. Thus, for the IFM dataset, 576 documents were used. The CIMM dataset has a total of 3326 documents, but three repeated texts were removed. Therefore, we worked with 3323 texts.

Table 2. Description of the textual datasets used in the experiments

Datasets	CSTNews	IFM	NN	CIMM
# documents	131	576	1.057	3.323
# classes	4	5	5	3
Classes	Daily; Sport; World; Politics.	WP01; WP02; WP03; WP04; WP05.	Scientific; Scientific-Publicizing; Informative; Others; Technical and Administrative.	Dissertation; News; Thesis.
Text average per class	32.75	115.2	211.4	1107.671
# texts of minority class	24 (18.32%)	5 (0.87%)	6 (0.57%)	311 (9.36%)
# texts of majority class	40 (30.53%)	283 (49.13%)	433 (40.96%)	1976 (59.46%)

3.2 Evaluation Approach of Normalization Techniques in Text Classification

The approach used to evaluate normalization techniques in the text classification task (Figure 1) presents the following steps in order to perform the experiments. These steps were covered for each data set.

In the **1st step**, each dataset was treated in the following way: each word was transformed to its lowercase form; every number, accents, punctuation, and one-letter words were removed from the texts; and the words considered as stopwords

⁵ IFM - <http://www.ifm.org.br>

⁶ GETerm - <http://www.geterm.ufscar.br>

⁷ CIMM - <http://www.cimm.com.br>

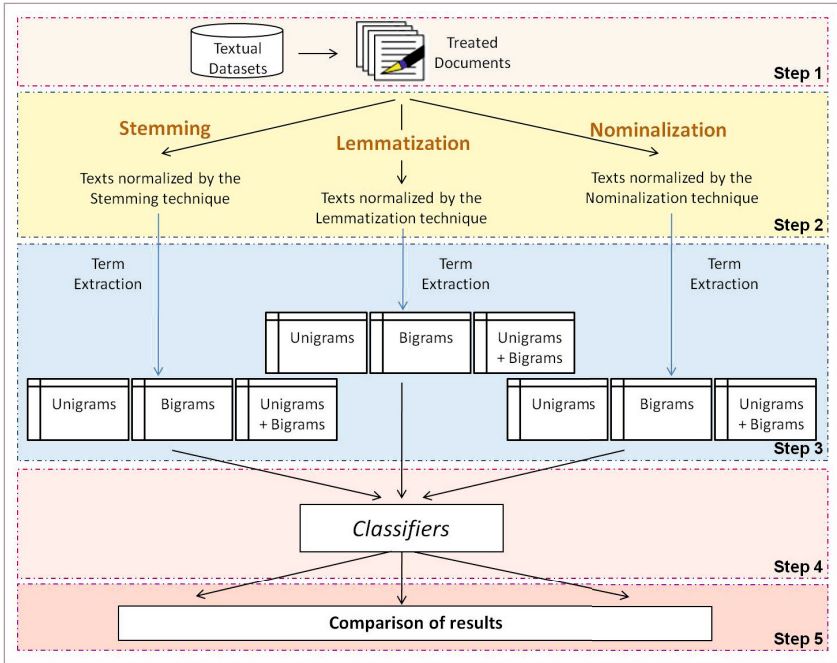


Fig. 1. Description of the preliminary experiment: normalization techniques in text classification

were also removed. The stoplist used in this experiment was obtained from the PRETEXT II tool [32], and the conjugations of the verb “to-be” were increased.

Some words corresponding to specific feature tags of the selected datasets were considered as stopwords and, therefore, were removed from the dataset documents, preventing the election of terms with such words. For the CIMM dataset, the following tags were removed: *doc*, *titulo*, *conteudo*, *resumo*, *autor*, *link*, *tese*, and *dissertacao*. For the IFM dataset, the following tags were removed: *doc*, *titulo*, *keywords*, *sp*, *atividade*, *descricao*, and *conteudo*.

In the **2nd step**, three techniques for normalization were applied. For the stemming technique, Porter - Portuguese algorithm available in the PRETEXT II tool [32] was used; for the lemmatization technique, the Nunes Lemmatizer [8] was applied; and for the nominalization technique, the FORMA and CHAMA tools [11] were used.

These techniques were applied to each dataset, with three sets of normalized texts, normalized with (i) stemming, (ii) lemmatization, and (iii) nominalization. Since most studies use stemming, it is considered our baseline. At the end, twelve term candidates were obtained for each of the four datasets used.

In the **3rd step**, several terms were extracted from the normalized text sets in each technique. Through the PRETEXT II tool, with the no-stemming option, each text set generated an unigram list, a bigram list, and a compound list of unigrams and bigrams.

In order to remove term candidates that do not represent the datasets well, the Document Frequency (DF) measure was used. In that way, only term candidates that appear at least in d_{min} and at most d_{max} documents in the datasets were considered, assuming that d_{min} and d_{max} were previously defined by the user.

As the datasets used in the experiments are unbalanced, in order not to interfere in the final results, the minimum cutoff value of DF was set at 1% of the number of documents from the minority class of each dataset. If the value obtained was less than two, the value two was used. The maximum DF cutoff was set at 90% of the number of documents from the majority class of each dataset. Thus, the values of DF minimum and maximum were, respectively, 2 and 22 for CSTNews dataset, 2 and 5 for IFM and NN datasets, and 3 and 280 for the CIMM dataset.

Besides the DF measure, bigrams were also submitted to the Log Likelihood Ratio method, available in the NSP package, considering $p_value = 0.005$. The application of this method aimed at keep the combination (grams) that were more than casual occurrences in the text documents.

After performing these statistics cuts, three new lists of terms were obtained: a list of unigrams, one of bigrams, and a unigrams together with bigrams list. The size of each list for each dataset used are shown in Table 3. For a better comparison of the reduction in numbers of terms when using each normalization techniques, the number of grams when considered each dataset as a whole is shown, i.e. without any treatment (*# initial terms*).

Table 3. Number of extracted terms using each normalization technique

Datasets	Grams	# Initial terms	# terms extracted by each technique		
			Stemming	Lemmatization	Nominalization
CSTNews	Unigrams	6.581	2.168	2.268	2.113
	Bigrams	21.147	2.708	2.972	2.804
	Unig. + Big.	27.728	4.876	5.240	4.917
IFM	Unigrams	47.818	11.486	13.865	13.623
	Bigrams	404.946	182.769	186.093	185.519
	Unig. + Big.	452.764	194.255	199.958	199.142
CIMM	Unigrams	43.425	8.833	9.553	8.631
	Bigrams	283.187	26.958	28.111	28.458
	Unig. + Big.	326.612	35.791	37.664	37.089
Nanoscience and Nanotechnology	Unigrams	68.421	12.797	15.487	15.052
	Bigrams	480.351	136.086	134.607	142.813
	Unig. + Big.	548.772	148.883	150.094	157.865

These terms listed were represented by feature-value matrices, considering just the term frequency.

In the **4th step**, these matrices were used as input to represent the documents in the textual classification task. Weka (*Waikato Environment Knowledge Analysis*) [14] was used to carry out the classification experiments. In order to generate

the classifiers, five classifiers typically used in the text classification tasks were chosen to compare the use of different normalization techniques. Within these five inducers, two have statistical basis, which are the *Naive Bayes Multinomial Classifier* [19] and the *Discriminative Multinomial Naive Bayes Classifier* (DMNBText) [33]; another one is based on the use of hyperplanes to divide the text into different classes, namely SVM (Support Vector Machines) with polynomial kernel classifier [15] and a decision tree based classifier, which is the J48 classifier [24]. The last has a base of decision tree rules, which is the conjunctive rule-based classifier (*Conjunctive Rules*). These classifiers are available and implemented in the Weka environment. To carry out the experiments, default settings were configured for each classifier and were run with *10-fold cross-validation*.

In the **5th step** the results obtained in the classification task for each normalization technique were compared. Aiming to verify whether there are statistically significant differences among the results obtained in classification, the accuracies of each classifier for each of the twelve sets of terms used (unigrams, bigrams, and unigrams + bigrams of each textual dataset) were compared by statistical tests. The employed tests were Friedman [9] (nonparametric) and SNK [31] (parametric), which provide, respectively, a *ranking* and a difference of means. Both tests were executed with 95% of confidence ($p\text{-value} = 0.05$).

3.3 Results Analysis

The results shown in Table 4 correspond to the accuracy and standard deviation rates obtained by the classifiers when using the three normalization techniques: stemming, lemmatization, nominalization. This table presents only the results about NN dataset. According to the SNK test, only for the NN dataset, a statistically significant positive difference was achieved when using the nominalization technique which outperforms stemming technique with only unigrams. As on the other datasets no statistically significant differences were obtained, and because of the limited space available in this paper, only the results of the NN dataset are shown (Table 4).

Two possible explanations for this behavior were found, which are related to the way these two techniques normalize the words. Nominalization technique

Table 4. Accuracy and standard deviation rates of the classifiers when used the Nanoscience and Nanotechnology dataset with unigrams

Classifiers	Unigrams		
	<i>Lemmatization</i>	<i>Stemming</i>	<i>Nominalization</i>
DMNBtext	61.57(5.24)	58.59(5.26)	61.41(4.66)
NaiveBayes	60.55(5.22)	59.07(5.30)	61.12(4.45)
SVM	53.73(4.73)	52.89(4.79)	55.77(4.39)
Conjunctive Rules	40.96(0.41)	40.96(0.41)	41.05(0.30)
J48	77.25(2.21)	77.37(2.65)	78.21(1.78)

reduce the words in a much more smooth way than the stemming, maintaining some characteristics of terms that can help to distinguish them. The second explanation is that in order to apply the nominalization technique, each should be searched in grammatical class pre-defined lists. With this search result, the word is reduced to a form that adds a name-like syntactic/semantic behavior. To apply the stemming technique, other rules, such as the word size, prefixes and suffixes of the words, are applied. Given the implementation of each technique, we can induce the explanation of the fact that the nominalization was significantly better than stemming for unigrams because the original word is more likely to remain the same (without normalization) for a specific domain, such as the field of Nanoscience and Nanotechnology.

In Table 5 are shown some examples of unigrams normalized by the stemming and nominalization techniques. The original unigrams *nanoamperes* and *nanoamperimetro* have been normalized by the stemming technique in two different unigrams, which were *nanoamp* and *nanoamperimetr*, respectively. However, these original unigrams correspond to the same subject and therefore should have been normalized into a single unigram, as done by the nominalization technique (*nanoamper*). The opposite happened with the original unigrams *nanocamadas(s)* and *nanocameras* and with unigrams *nanocave* and *nanocor*. These should have been normalized into different unigrams, but was only done in the nominalization technique. In these normalizations, the nominalization technique reduced these four unigrams by just removing their original plurals. This can be explained by the fact that they probably are not present in the unigrams predefined lists used by the nominalization technique, since these words are very specific in the Nanoscience and Nanotechnology field.

Table 5. Some examples of unigram normalization in the Nanoscience and Nanotechnology dataset

Unigrams examples		
<i>Originals</i>	<i>Stemmed</i>	<i>Nominalized</i>
nanoamperes	nanoamp	nanoamper
nanoamperimetro	nanoamperimetr	
nanocamada(s)	nanocam	nanocamada
nanocameras		nanocamera
nanocave	nanoc	nanocave
nanocor		nanocor

In relation to the use of different grams (only unigrams, only bigrams or unigrams with bigrams) a statistically positive difference was found when using unigrams + bigrams, which outperforms the unigrams alone for the NN dataset. As for the other datasets no statistically significant differences were found among them, only the results of the Nanoscience and Nanotechnology dataset are shown

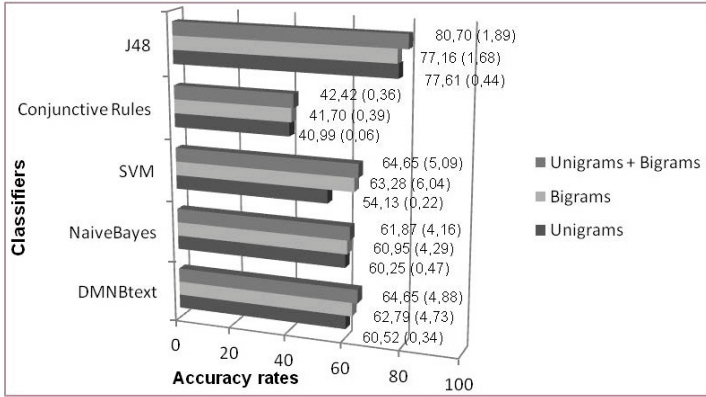


Fig. 2. Accuracy and standard deviation rates to the NN dataset with different n-grams

(Figure 2). In this table, the results shown correspond to the accuracy and standard deviation rates of the classifiers considering all the techniques together - when used only unigrams only bigrams and unigrams with bigrams.

This fact, for this dataset, can be explained because it belongs to a relatively new domain when compared to the other datasets. In addition, there are many compound terms (bigrams) present in this domain, such as *nano_reator*, *originalidade_nanocarbono* and *garrafas_plasticas*.

Another characteristic that we could observe is the complexities of the algorithms used to apply each normalization technique. In order to calculate the complexities, we have to analyze each algorithm.

For the stemming technique, we used the PRETEXT II tool algorithm [32], in which the authors created a data structure to store the words of the document and they used specific rules to stem the words in the Portuguese language. For the lemmatization technique, we developed an algorithm where each word of a document is replaced by its respective lemma, and this lemma is sought on a basis of canonical words of the Nunes Lemmatizer [8]. To calculate the complexity of the lemmatization algorithm, we assumed that the search complexity of the hash structure that was used in this algorithm is $\mathcal{O}(1)$. Finally, for the nominalization technique, we used the FORMA and CHAMA tools [11]. According to Gonzalez, to calculate the complexity of the lemmatization algorithm, the authors assumed that the tree data structure used in this algorithm is partially balanced and that the number of adjectives, adverbs, verbs, or nouns is at most equal to n , where n is the number of words in the document. The details of the algorithms can be seen in Conrado [6].

Considering the worst scenario, the complexity of these three algorithms to process a document belongs to the linear class $\mathcal{O}(n)$, where n is the number of

words in the document. As these complexities are the same, it is not possible to use this criteria to choose the most suitable technique for the text classification task.

4 Conclusion

In contrast to expectations, there are no significant difference in the classification results when using different normalization techniques to extract the terms that were used in the text classification task. The only exceptions were the results obtained with the NN dataset, which belongs to a new and particular domain. Anyway, it was necessary to make this experimental verification since, up to this date, no studies were found evaluating the use of these techniques in the text classification task for Brazilian Portuguese.

Regarding the use of unigrams, bigrams or unigrams with bigrams for classification, there are also no significant difference in the results obtained. The unique exception was with the textual dataset of Nanoscience and Nanotechnology. For this dataset, a statistically positive difference was found when using unigrams with bigrams, which outperforms the unigrams. These results are similar to the work of Bekkerman and Allan [4].

Thus, we conclude that, in general, for the textual datasets (except for NN dataset) other criteria could be used to aid in selecting which technique must be applied for normalization and how many grams to be used in the text classification task. An example of criterion that can be used is the number of terms that is desired. Nine among the twelve obtained termsets had the smallest numbers of terms. These termsets used the stemming (75% of cases), as shown earlier in Table 3. Regarding the use of grams, we suggest to use only the unigrams, since they had fewer terms and are computationally cheaper.

Or yet, the criteria can be chosen to analyze which normalization technique applies better to the final goal, as shown in Conrado et al. [7]. For example, when comprehensibility is needed of the results, the nominalization technique can be used. Secondly, lemmatization should be used, and finally the stemming. Nominalization is the technique that reduces words in the least aggressive way, followed by the lemmatization technique; therefore, the stemming is more aggressive in reducing the words. For this reason, the nominalization obtains the most comprehensible terms if compared with the other two techniques, followed by lemmatization and then stemming.

Since the three algorithm complexities of the normalization techniques are the same, it is not possible to use this criterion to choose the most suitable technique for the text classification task. But, as future work, we could analyze the run time of each technique. Furthermore, we intend to balance the number of terms extracted by each technique, i.e., carry out experiments with fixed numbers of terms extracted by the three techniques. For example, we will repeat the experiments of this article using only the first 50, 100, 150, and 200 terms extracted by each technique in order to verify which technique works better with

a fixed number of terms. Another interesting research would be to test other statistical measures besides DF to remove term candidates which would not represent the datasets as expected, for example, Term Frequency - Inverse Document Frequency (TF-IDF), Term Contribution (TC), C-Value, etc. Additionally, we intend to evaluate the use of these techniques in other Text Mining tasks, such as clustering and summarization of texts.

References

1. Alvares, R.V., Garcia, A.C.B., Ferraz, I.: STEMBR: A Stemming Algorithm for the Brazilian Portuguese Language. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 693–701. Springer, Heidelberg (2005)
2. Arampatzis, A., van der Weide, T., Koster, C., van Bommel, P.: Linguistically-motivated Information Retrieval, pp. 201–222. Marcel Dekker, NY (2000)
3. Aranha, C.N.: Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: sob o Enfoque da Inteligência Computacional. PhD thesis, Departamento de Engenharia Elétrica - PUC - Rio de Janeiro (2007)
4. Bekkerman, R., Allan, J.: Using bigrams in text categorization. Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst (2004)
5. Brill, E.: Transformation-based error-driven learning of natural language: A case study in part of speech tagging. *Computational Linguistics*, 543–565 (1995)
6. Conrado, M.S.: O efeito do uso de diferentes formas de geração de termos na compreensibilidade e representatividade dos termos em coleções textuais na Língua Portuguesa. Master's thesis, Instituto de Ciências Matemáticas e de Computação - USP, São Carlos, SP (2009)
7. Conrado, M.S., Marcacini, R.M., Moura, M.F., Rezende, S.O.: O efeito do uso de diferentes formas de geração de termos na compreensibilidade e representatividade dos termos em coleções textuais na Língua Portuguesa. In: *Proceedings of II Web and Text Intelligence - 7th Brazilian Symposium in Information and Human Language Technology*, São Carlos, SP (2009)
8. das Nunes, M.G.V.: The design of a lexicon for Brazilian Portuguese: Lessons learned and perspectives. In: *Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese*, Curitiba, pp. 61–70 (1996)
9. Demšar, J.: Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research* 7(1), 1–30 (2006)
10. Ebecken, N.F.F., Lopes, M.C.S., de Aragão, M.C.: Mineração de Textos. In: Rezende, S.O. (ed.) *Sistemas Inteligentes: Fundamentos e Aplicações*, 1st edn., Manole, ch. 13, pp. 337–364 (2003)
11. Gonzalez, M. A. I.: Termos e Relacionamentos em Evidência na Recuperação de Informação. PhD thesis, Instituto de Informática - UFRGS, Porto Alegre (2005)
12. Gonzalez, M.A.I., de Lima, V.L.S., de Lima, J.V.: Tools for Nominalization: An Alternative for Lexical Normalization. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) *PROPOR 2006*. LNCS (LNAI), vol. 3960, pp. 100–109. Springer, Heidelberg (2006)
13. Braga, Í.A., Monard, M.C., Matsubara, E.T.: Combining unigrams and bigrams in semi-supervised text classification. In: *14th Portuguese Conference on Artificial Intelligence - New Trends in Artificial Intelligence*, Aveiro, Portugal, pp. 489–500 (2009)

14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. In: Explorations of Special Interest Group on Knowledge Discovery and Data Mining, vol. 11, pp. 10–18 (2009)
15. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's smo algorithm for svm classifier design. *Neural Comput.* 13(3), 637–649 (2001)
16. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press (2008)
17. Manning, C.D., Raghavan, P., Schütze, H.: *Language models for information retrieval*. In: *An Introduction to Information Retrieval*, ch. 12. Cambridge University Press (2008)
18. Maziero, E.G., del Rosario Castro Jorge, M.L., Pardo, T.A.S.: Identifying multidocument relations. In: *Proceedings of 7th International Workshop on Natural Language Processing and Cognitive Science, Funchal/Madeira, Portugal*, vol. 1, pp. 60–69 (2010)
19. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: *AAAI Magazine - Workshop on 'Learning for Text Categorization*, pp. 1–8 (1998)
20. Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D.: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Science (2012)
21. Mitra, V., Wang, C.-J., Banerjee, S.: Text classification: A least square support vector machine approach. *Appl. Soft Comput.* 7(3), 908–914 (2007)
22. Nui pian, V., Meesad, P., Boonrawd, P.: Improve abstract data with feature selection for classification techniques. *Advanced Materials Research* 403-408, 3699–3703 (2011)
23. Oren go, V.M., Huyck, C.: A stemming algorithm for portuguese language. In: *Proceedings of Eighth Symposium on String Processing and Information Retrieval, Chile*, pp. 186–193 (2001)
24. Quinlan, J.R.: *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco (1993)
25. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pp. 491–497. University of Pennsylvania (1996)
26. Read, J., Webster, J., Fang, A.C.: In: *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Sendai, Japan*
27. Řehůřek, R., Sojka, P.: Automated Classification and Categorization of Mathematical Knowledge. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) *AISC 2008, Calculemus 2008, and MKM 2008*. LNCS (LNAI), vol. 5144, pp. 543–557. Springer, Heidelberg (2008)
28. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of International Conference on New Methods in Language Processing*, pp. 44–49 (1994)
29. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
30. Silic, A., Chauchat, J.-H., Basic, B.-D., Morin, A.: N-grams and morphological normalization in text classification: A comparison on a croatian-english parallel corpus. In: Neves, J., Santos, M.-F., Machado, J. (eds.) *13th Portuguese Conference on Artificial Intelligence, Guimaraes, Portugal*

31. Snedecor, G.W., Cochran, W.G.: Statistical Methods, 6th edn. Iowa State University Press, Ames (1967)
32. Soares, M.V., Prati, R.C., Monard, M.C.: PreText II: Descrição da reestruturação da ferramenta de pré-processamento de textos. Technical Report 333, Instituto de Ciências Matemáticas e de Computação - USP, São Carlos, SP (2008)
33. Su, J., Zhang, H., Ling, C.X., Matwin, S.: Discriminative parameter learning for bayesian networks. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1016–1023. ACM, New York (2008)

Extracting Definitions from Brazilian Legal Texts

Edilson Ferneda¹, Hércules Antonio do Prado^{1,2}, Augusto Herrmann Batista^{1,3},
and Marcello Sandi Pinheiro⁴

¹ Graduate Program on Knowledge and IT Management, Catholic University of Brasilia,
SGAN 916 Av. W5, 70790-160, Brasília, DF, Brazil

² Embrapa - Management and Strategy Secretariat,

Parque Estação Biológica - PqEB s/nº, 70770-90, Brasília, DF, Brazil

³ Logistics and Information Technology Secretariat, Ministry of Planning, Budget and
Management, Esplanada dos Ministérios - Bloco C, 70046-900, Brasília, DF, Brazil

⁴ Federal University of Rio de Janeiro / COPPE,

Cidade Universitária, Centro de Tecnologia, Bloco B, sala 101 - Ilha do Fundão,

Caixa Postal: 68506, 21941-972, Rio de Janeiro, RJ, Brazil

eferneda@pos.ucb.br, hercules@ucb.br,

augusto.herrmann@planejamento.gov.br,

msandipinheiro@yahoo.com.br

Abstract. In order to avoid ambiguity and to ensure, as far as possible, a strict interpretation of law, legal texts usually define the specific lexical terms used within their discourse by means of normative rules. With an often large amount of rules in effect in a given domain, extracting these definitions manually would be a costly undertaking. This paper presents an approach to cope with this problem based in a variation of an automated technique of natural language processing of Brazilian Portuguese texts. For the sake of generality, the proposed solution was developed to address the more general problem of building a glossary from domain specific texts that contain definitions amongst their content. This solution was applied to a corpus of texts on the telecommunications regulations domain and the results are reported. The usual pipeline of natural language processing has been followed: preprocessing, segmentation, and part-of-speech tagging. A set of feature extraction functions is specified and used along with reference glossary information on whether or not a text fragment is a definition, to train a SVM classifier. At last, the definitions are extracted from the texts and evaluated upon a testing corpus, which also contains the reference glossary annotations on definitions. The results are then discussed in light of other definition extraction techniques.

Keywords: Information extraction. Definition extraction. Natural Language Processing.

1 Introduction

As the information amount grows exponentially in organizations, especially unstructured text in natural language, it becomes ever more difficult to perform

Information Extraction (IE) by non-automated means [4]. Computational literature has produced plenty of approaches to mitigate this problem. Most notably, the fields of Computational Linguistics [15] and Text Mining [20] have contributed significantly to ease those difficulties by providing methodological and technological assets. Some IE efforts target specific problem domains in order to deal more adequately with their intrinsic characteristics. One such specific problem domain is Definition Extraction (DE).

Building domain specific glossaries can be useful, for instance: *(i)* to schematize knowledge externalization, *(ii)* to facilitate learning and grasping of concepts, and *(iii)* as a reference point for searching domain specific information. The glossary building process, however, is an expensive one, requiring extensive work from domain experts. Thus, automated or semi-automated processes for definition extraction, based upon computational algorithms for natural language processing (NLP), can significantly reduce the effort required for building a glossary from domain texts that have definitions amongst their discourses. In the last decade, many techniques have been developed aiming to mitigate this problem, with varied degrees of success.

Automated definition extraction has shown itself to be useful in many domain applications, such as automated dictionary construction, question answering systems, and ontology engineering [9].

This paper focusses on the automated definition extraction for the purpose of building a glossary. It presents the results of a definition extraction experiment applied to a corpus of Brazilian legal texts on telecommunications. The extracted definitions are then compared to a reference glossary of definitions manually extracted by domain experts.

In the following sections, a brief overview is presented on what are definitions, followed by a review of previous works on definition extraction. Then, the methodology used for text segmentation, part-of-speech (POS) tagging, the selection of attributes, training and classification tasks are described. Finally, the experiment results are revealed.

2 What Are Definitions?

According to Sager [36], *terminology*, as a theory, is a set of premises, arguments and conclusions required for explaining relationships among concepts and terms. Differently from lexicology, where the lexical unit is the starting point, terminology builds on concepts [44].

Concepts are mental constructs to which we assign labels [31] which, in turn, are known as *terms* [35]. The first discussions on definitions come from Philosophy. Aristotle studied the *genus et differentia* kind of definition. A *genus et differentia* definition explains a term (*definiendum*) by determining its type (*genus*) and one or more characteristics (*differentia*) that distinguish it from same type entities. This kind of definition can also be useful to express hyponym-hyperonym relations; in this case *genus* refers to the more general term and *definiendum* to the more specific one. The

hyponym-hyperonym relationship can be applied for building hierarchies or taxonomies, generating structures useful for constructing thesauri.

For example, consider the sentence “Cheese is a solid food made from milk of cow, goat, sheep, buffalo, or other mammals”; in this case, “cheese” is the *definiendum*, “food” is the *genus*, and the remaining is the *differentia*.

Shaw [41] categorizes the modalities of definitions as follows: (i) by etymology, based on the meaning of its components; (ii) by analysis, that consists in explaining the meaning of a term by describing its essential characteristics; (iii) by exclusion in which a term is described by what it is not; (iv) by example, by expressing an example of term use or function; (v) by analogy that consists in explaining the meaning of a term by comparing it to other terms, in some sense, with similar meaning; and (vi) by context by clarifying the term by reference to the words immediately preceding or following it. These modalities, referred by Shaw as methods and by Borg [9] as *definitional properties*, are not mutually exclusive and are often combined in the definitions by *genus et differentia*.

Not always a textual fragment that carries some meaning of a term is a definition. For example, the following sentence, a popular definition for the GNU free software tools, “GNU means 'GNU's not Unix” is a questionable one because it uses a recursive structure.

The legal and regulatory texts, by striving to be clear and objective for the sake of avoiding different interpretations, have a predominance of analytical definitions. Thus, this mode is the one considered in this work.

3 Previous Works on Definition Extraction

Research on DE stems mostly from the problem domain of question answering systems [7,21,29,37,38,39]. More specifically, in order to answer questions such as “what is a ...?”, those systems seek to answer such questions by providing a definition.

One of the first works to deal specifically with definition extraction, by proposing a pragmatic approach, was Klavans and Muresan [24]. It presented the DEFINDER system that is applied on texts about health targeted at lay people. The system is based on manually constructed patterns by using a tagger and a finite state grammar. The candidate definitions are filtered by certain rules, such as its syntactic structure, which may or may not reinforce the notion that the candidate is a definition.

Saggion [37] worked on the AQUAINT corpus, which is based on news texts from The New York Times, by extracting and collecting definitions for the purpose of further answering general questions. Upon manually identifying 36 patterns by inspecting the corpus, the proposed algorithm took part in the TREC QA 2003 question answering competition. It achieved better than average results and ranked among the 10 best, out of 25 total participants.

Definition extraction for the purpose of building a knowledge base for question answering systems was also tackled by Fernandes [21]. He proposed an approach based upon regular expressions and uses part-of-speech (POS) tagging and named

entity recognition techniques. Considering that the objective was to extract information useful for supporting a question answering system or, more specifically, questions such as “what is” or “who is”, not every one of the 19 patterns used are useful for the general DE problem. For instance, some of them are concerned with connectors such as “named” or “also called”, as well as a pattern that associates a person with an age.

Tanev et al [42] were some early users of POS tagging in support of definition extraction. Their objective was to develop an automated system to answer questions such as “what is Yakuza?” or “who is Jorge Amado?”. For that end, the system extracts definitions by using patterns in a specific grammar, whose elements can be the focus (*definiendum*), word POS tags or word lemmas. According to the authors, such patterns can capture a range of definitions such as “Yakuza is the Japanese mafia”, but not “the members of Yakuza are...”

Fahmi and Bouma [19] also considered DE in context of question answering systems. They used a corpus based on Dutch Wikipedia entries on the domain of health sciences. They identify some attributes which could be relevant to the experiment: (i) sentence position in the document, (ii) subject position in the sentence, (iii) syntactic properties and (iv) named entity classes. Machine learning (ML) methods are then used to determine which of those features bring best results. Some classifiers were trained – naïve Bayes, maximum entropy and support vector machines (SVM) – by using syntactic features such as POS tagging and subject position in the sentence, in order to classify whether or not sentences are definitions. The classifier to achieve best results was the maximum entropy based one, with 92.21% accuracy.

Westerhout and Monachesi [43] studied DE on Dutch e-Learning texts. Rule grammars were applied in order to capture a great number of definitions, favoring initially recall over precision. Then, ML techniques are applied to filter the results, improving precision. The results are slightly inferior to Fahmi e Bouma, reaching 88% accuracy.

Przepiórkowski et al [33] investigated DE on Slavic languages (Bulgarian, Czech and Polish). They studied POS-tagged texts concerning to e-Learning. By manually crafting recognition patterns, they achieved low precision results (22 to 23%) and also low recall (9 to 46%) and F-measure (0.339). The authors noted low agreement levels among definition annotators about whether or not a sentence is a definition. These annotations were used as a gold standard for evaluating the automated technique and , possibly, this was one of the reasons appointed for the low performance.

The earliest work found related to DE on Portuguese language texts was Pinto and Oliveira [32]. They sought to identify the most common syntactic structures in definitions. For this purpose, they used regular expressions for processing texts. Those were applied to the Corpógrafo, a corpus in Portuguese language [40] containing texts in the medical domain.

In one of the few studies dedicated to extract glossaries from texts in Portuguese, Del Gaudio and Branco [16] applied a rule based grammar on a corpus of 33 documents (like thesis, articles, and tutorials). They choose to extract three kinds of definitions: (i) those using the verb *ser* (“to be”), e.g., “*FTP é um protocolo para*

transferir arquivos pela internet” (“FTP is a protocol to transfer files in the Internet”), (ii) definitions using other verbs, e.g., “*Uma ontologia pode ser descrita como uma caracterização formal de objetos*” (“An ontology can be described as a formal definition for objects”), and (iii) definitions by using of punctuation marks, such as “*TCP/IP: um protocolo usado para a troca de informações entre computadores*” (“TCP/IP: a protocol used for the information interchange among computers”). They reached a precision of 14% and a recall of 86% in comparison to manually annotated testing set. The high recall is justified since the intention was recovering as many candidates to definitions as possible, to subsidize a further manual inspection.

On the other hand, most of the studied works found a low precision because, along with the valid definitions, there are linguistic constructions that seem to be definitions but are not. For example, “inflation was the hardest economical problem for our country in 1991” is not a definition for “inflation”. In order to improve the precision some authors, such as Fahmi and Bouma [19] and Westerhout and Monachesi [43], select definition candidates that are filtered by means of a supervised ML technique. In the same sense, Alarcón et al [1] developed a methodology that not only extracts definition patterns but also filters out non-definitions. They worked on a corpus in Spanish language that includes technical texts from many fields, annotated by means of part-of-speech tagging. The definitional contexts were classified as (i) simple verbal (e.g., to define, to mean) and (ii) compound verbal, that also includes a grammar particle such as a preposition or an adverb (e.g., to consist in). The filtering of non-definitions was also based on rules, defined after identification of non-definitional contexts. By this way, they improved significantly the precision: from 29% to 100% considering each definitional pattern. The recall ranged from 87% a 100% and the F-measure was not reported.

Unlike previous work that built rules from manual inspection of the most common syntactical forms of definitions, Borg et al [8] approached the problem by applying a genetic algorithm (GA) and a genetic programming (GP) technique. Six categories of definitions were considered. Methodologically, they innovated by evolving weights associated to functions applied to sentences. In a later work [9], they extended this proposal by combining both, GA and GP approaches. They evaluated the proposals per se and in combination. For GA alone, the best results reached an F-measure de 0.57, a precision of 70% and a recall of 56%. For the GP approach, the best results were 0.30, 25% and 36%, respectively. The combined process includes: (i) creation of an annotated training set; (ii) application of GP to learn simple characteristics useful to distinguish definitions from non-definitions; (iii) use of GA to approximate the weights for these rules; and (iv) application of these rules and weights in a tool for definitions classification.

Rigutini et al [34] presented a system for automated generation of crossword puzzles, including a component for definition extraction. They used, as a corpus, articles from the Italian Wikipedia. For POS tagging, they applied a support vector machine trained on Treebank, an Italian language annotated corpus [26].

Del Gaudio and Branco [17] approached the DE as a binary classification problem of textual periods as being or not a definition. They considered the definitions containing the verb *ser* (“to be”). The corpus LT4eL, which contains e-Learning texts

in many languages, including Portuguese, was used in the experiments. The 100 most frequent n -grams (with n varying from 1 to 4) were extracted from the POS tags, among 1.360 definitions in this corpus. The results were improved in comparison with their previous work [16], reaching an F-measure of 0.77.

Navigli and Velardi [30] proposed an approach based in word lattices for DE. The technique extracts not only definition but also hyperonym relations. Initially, some patterns are created based in some training phrases, keeping words which have very frequent use (e.g., “and”, “the”, usually called *stop words*) and replacing the remaining words by their respective part-of-speech tags. These patterns form the basis for building the word lattices that are used to extract the definitions. A set of annotated phrases from Wikipedia were used as corpus for training and testing, while a set of web pages from ukWaC [22] was chosen as a testing corpus. A precision of 98% was obtained on the Wikipedia corpus and 94% on the ukWaC corpus dataset. Recall was 60% for the Wikipedia testing set and over 75% for the ukWaC one.

4 Methodology

Development of this work benefited from the availability of open source tools (e.g., the Natural Language Toolkit – NLTK [25]), that do implement many of the techniques often used at most phases in the typical NLP pipeline, such as text segmentation and part-of-speech tagging of words.

Portuguese language legal texts were chosen for this work due to availability of manually annotated definitions (i.e. a glossary). Another reason for this choice is the fact that there have been a few research efforts in the NLP community to meet these conditions (Portuguese language and laws domain).

This work followed these steps: (i) pre-processing, (ii) segmentation of paragraphs, sentences, and texts elements like words and punctuation (also called tokenization), (iii) part-of-speech tagging, (iv) extraction of characteristics relevant to ED, (v) classifier training, (vi) classifier application, and (vii) classifier evaluation.

The methodology applied is based in the ones from Westerhout and Monachesi [43] and Del Gaudio and Branco [16] that privilege the recall over precision in order to find definition candidates. The candidates are chosen when the n feature extraction functions (FEF) are applied to form an n -dimensional vector for each sentence in the corpus. Those sentences with no positive values in at least one vector cell are immediately discarded in order to reduce the search space.

Unlike Del Gaudio e Branco [16], the definition candidates are not presented to specialists for manual inspection. Rather, they are filtered by means of an SVM classifier aiming at improving the precision, as done by Westerhout and Monachesi [43]. The FEF were defined by specialists, instead of using some (semi) automatic process (such as the GP used by Borg et al. [9]). If this approach was used, the classification of text and the calculation of the F-Measure (the fitness function) for the whole corpus at each individual generation would supposedly be very costly. The main reason for this is that the corpus processed has a substantially larger volume of sentences. So, the chosen course of action was to use only predefined FEF.

4.1 The Experimentation Corpus

The Brazilian Collection of Telecommunication Laws (BCTL) [6] was used as corpus for training and evaluation of the glossary generation process. The corpus has 1,940 reference documents that were used by specialists in order to manually extract the definitions included in the glossary. For each entry in the glossary, four information items are identified (see Figure 1), (i) the *definiendum*, (ii) an optional definition context, (iii) the *definiens*, and (iv) the source document. The glossary contains 2,097 definitions (*definiens*) for 1,757 terms (*definienda*). This difference happens due to the fact that some terms have more than one meaning. The glossary files were available in DocBook and PDF formats and the source documents in PDF. Those had been used by the specialists to manually extract the definitions and were also stored in the corpus. The existence of definitions pointing to documents not included in the corpus required two possible courses of action: either identifying the documents related to the definitions and retrieving the missing documents from an external source, or ignoring the orphan definitions. Due to the obvious cost involved in the first option, the second course of action was decided upon. For that end, a filter was applied to remove those definitions. After filtering, 1,797 definitions remained for 1,534 terms.

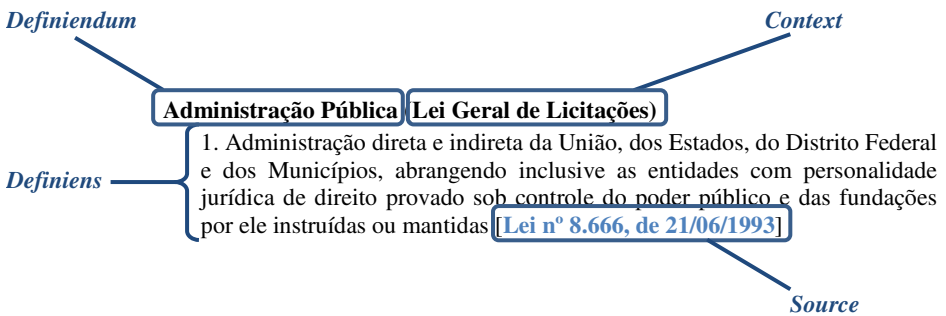


Fig. 1. Example of a glossary entry¹ in BCTL

Another difficulty was the existence of definitions related to multiple textual periods, a problem already reported by Przepiórkowski et al [33]. Such a difficulty was surpassed by taking only the first sentence and discarding any further ones. The rationale for this decision is the fact that, generally speaking, the information lost would be irrelevant, since usually the remaining periods refers to additional explanatory information.

¹ Translation:

Public Administration (General Law of Public Procurement)

1. Direct and indirect administration by the government, states, Federal District and municipalities, including those involving legal entities of private law under the control of the government and foundations instituted or maintained by it (Law No. 8,666 of 21/06/1993).

Moreover, many definitions in the glossary do not correspond strictly to the text in the documents. Sometimes different pronouns are used, modifications in the punctuations are done or even orthographic corrections are applied. In other cases, the *definiens* is complemented with other relevant information that, although not in the definition, occurs in the text. As the manual solution for these problems would also be laborious and expensive, a filter was again applied, now for removing those definitions whose *definiens* did not match exactly with the original text. After this second filtering, 1,549 definitions remained for 1,329 terms.

Notice that some definitions existing in the corpus were not included, for some reason, in the glossary (possibly for being out of scope or otherwise considered irrelevant for the glossary building task). Two examples of these definitions are *Contractor* (whoever is hired to provide a service for a part) and *Hirer* (the part which contracts a service) existing in the General Law of Public Procurement [11]. Definitions like these could be found by the extractor but are marked as non-definitions in the gold standard, making training and evaluation harder. Finally, agreement between annotators was not a consideration documented on this glossary's building process. The difference of opinions that usually happens among specialists is usually a factor that influences the evaluation outcome, as also noted by Przepiórkowski et al [33] and Borg [9].

4.2 Pre-processing

This phase starts by reading the DocBook XML file from the reference glossary and extracting its entries. This is done by separating each term, its context (if it is the case), definition text and source document reference.

Next, it was necessary to convert the documents that make up the corpus into a plain text format in order to enable its processing. So, each textual period was annotated, considering only those found in the glossary as belonging to the definitions class. This processing aims at gathering, in a single data structure, all reference definitions of the gold standard. This is the structure that would later be used for training and evaluation. Finally, the filtering mentioned in the previous section was applied.

4.3 Segmentation

The documents texts are then segmented by paragraph, by using regular expressions specific for detecting not only paragraph marks, but also the beginning of a new structured block of a legislation element. Many of the documents in the corpus are laws or norms that follow (albeit not strictly) the Complementary Law No. 95 [12], containing articles, items, etc. So, it was possible to determine a set of regular expressions to separate these blocks. Next, the *Punkt* algorithm, written by Kiss and Strunk [23] and implemented in NLTK, was applied to segment the corpus in periods and words. After that, a total of 6,120,832 tokens, including words and punctuation, became part of the corpus.

4.4 Part-of-Speech Tagging

The words in each sentence were POS tagged by applying a tagger trained in the Portuguese annotated MAC-MORPHO corpus [4]². This corpus was chosen as it is one of few in the Brazilian variant of the Portuguese language to contain part-of-speech annotations and to have sufficient amount of words to be used as training set for a tagger. This choice also defines a set of POS tags (e.g. “N” for noun, “V” for verb, and “PREP” for preposition). The corpus was divided into two segments; one containing 70% of the sentences intended to be used as training set for the tagger and the other with the remaining of the periods, for performance testing. Next, the construction of the tagger is described.

Many taggers available in NLTK were combined in sequence, chained in a back off tagger combination, in which the tagging task is relayed to the next tagger, whenever a tag cannot be determined. The set of taggers used were based in trigrams, bigrams, unigrams, affixes, and regular expressions, as shown in Figure 2. The tagger based in affixes was trained by considering the substrings corresponding to the three last characters of each word. The chaining order was chosen on the basis of experimentation with all possible permutations that reveals the order with the highest accuracy. In the following step, a Brill tagger [13], a transformation based supervised learning algorithm was applied to improve the tagger performance.

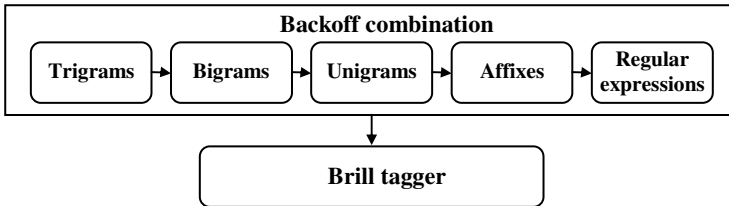


Fig. 2. POS tagger pipeline

This tagger was evaluated against the MAC-MORPHO corpus training set, reaching an accuracy rate of 90.44%. For the sake of comparison, a Hidden Markov Model based tagger was applied to the same corpus, obtaining only 57.56% of accuracy. Next, these results are discussed in face of other works related to POS tagging.

The same phenomenon of disagreement among different manual annotators, referring to DE and mentioned in session 3, was also observed in the case of POS tagging as reported by Marcus [27]. According to Marques and Lopes [28], the expected consistence level among tagging done by people is 98%. The best results found in automated POS tagging for the Portuguese language are 97.09%, found by

² The MAC-MORPHO corpus was built by NILC at the University of São Paulo, based upon news texts from the Folha de São Paulo newspaper. It is part-of-speech annotated with manually reviewed tags and contains 1,167,183 words. The corpus is available on the Lácio-Web catalog [5] at <http://www.nilc.icmc.usp.br/lacioweb/corpora.htm>.

Branco and Silva [10] with a tagger based in Brill transformations, although using a different corpus and set of tags. In this sense, Aluísio et al [5] reached 90.74% of accuracy in MAC-MORPHO corpus with a Brill tagger, similar to the one used in the present study.

The trained tagger was, then, applied to the BCTL corpus. As this corpus was not part-of-speech annotated, it was not possible to evaluate the accuracy degree of this process. However, this was the only feasible way to develop the study, taking into account that it was not possible to manually review these POS tags. Considering that the implemented tagger reaches results similar to other ones found, which also use MAC-MORPHO, sufficiently good results can be expected when applying it to the aim of DE in the BCTL. It should be noted, however, that a quantitative evaluation of the tagger performance on BCTL will not be possible without a manual review.

4.5 Feature Extractors

For a given textual period d , a Feature Extraction Function (FEF) returns a numeric value that quantifies some text feature of d . This feature can be, for example, the presence, absence, or combination of certain words or POS tags. The codomain of a FEF, considering the DE task, is binary, indicating whether or not a text is a definition.

The FEF are manually defined rules, based on the observation of some samples existing in the training corpus. Some of these are: (i) presence of verb *ser* (“to be”), in any conjugation form, followed by an article, (ii) presence of the punctuation symbol “:” or the “-” and its position in the sentence, (iii) presence of the expressions “*define-se como*” (“is defined as”), “*entende-se como*” (“is understood as”) or “*denomina-se*” (“is called”).

As shown by Alarcón et al [1], the consideration of negative characteristics can be useful for filtering non-relevant contexts. The following feature functions fulfill this purpose: (i) presence of the word *não* (“not”) before some form of verb *ser* (“to be”) in any of its conjugation forms and (ii) presence of word *tampouco* (“neither”). After training, these features can be expected to weight opposite to those found for affirmative features.

Regarding the use of definitional patterns and negative patterns filtering, this approach is similar to the one in Alarcón et al [1]. The adoption of some definitions categories using the verb *ser* (“to be”) and the use of other verbs and punctuation symbols follows the work of Del Gaudio e Branco [16].

4.6 Training

The corpus was divided into training (70%) and testing (30%) sets. Let l be the amount of sentences in the training set. The definition candidate sentences are associated to X_i , $i \in [1, l]$, n -dimensional vectors. These are formed by the results of n feature functions applied to each sentence P_i , and the binary class value to which it

belongs, meaning whether it is or not a definition. Thus, there are vectors $X_i = [f_1(P_i), \dots, f_n(P_i)]$ and the answers $D(P_i) = 1$ if P_i is a definition, or 0 otherwise, where $f_j(P)$ means, for instance, whether or not P contains the verb *ser* (“to be”) followed by an article, e.g. “*é um*” (“is a”), “*é o*” (“is the”), “*são os*” (“are the”). Those vectors and corresponding answers in the training set were used for training an SVM classifier. This project used the Orange free software machine learning tool [18], which in turn uses the libsvm library [14].

4.7 Classification

The trained classifier is then applied to the testing set. Consider $m < l$ the amount of sentences on the testing set. The SVM classifier outputs m binary values, one for each test sentence, which form an m -dimensional result vector $R(P_i)$. Those that have a value of 1 mean that the corresponding sentence was classified as a definition. Likewise, those with a value of zero attest that the respective sentence was not considered to be a definition by the classifier. Thus, a draft glossary is obtained by selecting the sentences from the test set which corresponding value in the classifier response is one.

At this stage only the testing set sentences were used, since the intention was to be able to evaluate quantitatively the results. However, if one's objective was to just build a glossary from natural language texts, and where no previous manual annotations are present (thus making any automated accuracy measurement effectually impossible), it would be interesting to apply the classifier to the whole corpus and then proceed to a manual review.

5 Results

The classifier result vector $R(P_i)$, which contains the binary values (definition or not a definition) obtained for each sentence P_i in the test corpus by applying the SVM classifier, is then compared to the respective gold standard vector $D(P_i)$, with the canonical values obtained from the reference glossary. The classification is considered correct in instances where the class determined by the classifier matches the reference class. Then precision, recall and F-measure are calculated.

The imbalanced dataset problem, as mentioned by Del Gaudio and Branco [17], was found in this experiment. There were 1,314 (2%) positive samples and 66,046 (98%) negative ones. Therefore, a random sampling of 1,314 negative samples was performed, in order to equalize sample quantities and improve classifier performance. This course of action was successful, obtaining 73,5% accuracy, 75,6% de precision, 69,6% de recall and 0.72 F-measure. The absolute and relative quantities of hits and misses for the positive and negative samples are shown on Table 1.

Table 1. Confusion matrix for the classifier results

Predicted Actual \	Negative	Positive
Negative	305 (38.7%)	89 (11.3%)
Positive	120 (15.2%)	275 (34.9%)

Upon inspecting samples of false negative results, it was not possible to pinpoint a reason for those samples not being included as definitions. Also, no obvious patterns were identified that could result in feature extraction tuning to improve recall. On the other hand, examining false positive samples revealed a few samples that might have been overlooked by the domain experts who initially built the glossary. Examples can be found on Table 2.

While these results cannot be directly compared to the approaches listed in the literature review, considering limitations due to different corpora and metrics used in each of them, it can be noticed that values are close to the ones obtained by Del Gaudio and Branco [17], which used a similar approach (0.67 in F-measure).

Nonetheless, it can be observed that the definitions extracted by the process can be quite useful in building a glossary, after some manual review process is conducted. This process would certainly be much less costly than a totally manual extraction of definitions from a large volume of texts.

Table 2. Examples of results generated

Extracted definitions	<ul style="list-style-type: none"> • <i>“Rede Externa: é o segmento de Rede de Telecomunicações suporte ao STFC, que se estende do PTR, inclusive, ao Distribuidor Geral de uma Estação Telefônica;”</i> "External Network: is the segment of Telecommunication Network which supports the PSTN and extends from the PTR, including to the General Distributor of a Telephone Station," • <i>“Unidade Operacional - UO é a unidade descentralizada, subordinada ao Escritório Regional que compõe a estrutura da Anatel;”</i> "Operating Unit - OU is a decentralized unit, subordinated to the Regional Office that composes the structure of Anatel;"
False negative	<ul style="list-style-type: none"> • <i>“Prestadora do SMP: entidade que detém autorização para prestar o SMP;”</i> "SMP Provider: entity that has authority to provide the SMP"
False positives	<ul style="list-style-type: none"> • <i>“Período de Coleta: mensal.”</i> "Collection Period: Monthly." • <i>“Termo de Autorização - ato administrativo vinculado que faculta a exploração de STFC, no regime privado, quando preenchidas as condições objetivas e subjetivas necessárias.”</i> "Authorization Form - binding administrative act which provides the STFC exploitation, in the private regime, when the objective and subjective necessary conditions are attended."

6 Conclusion and Future Efforts

During this work, a method was developed for extracting definitions in Portuguese language texts, by benefiting from the available free software libraries for NLP and for machine learning. Results compatible to similar studies were obtained.

As could be observed in literature, there are scant works on DE applied to Portuguese language texts, and none of them deals with the Brazilian dialect specifically, neither do they handle legal domain texts.

The main results obtained here, namely the definitions extracted by this processes, can be used as input for a manual review which would produce a glossary.

It has to be noticed, however, that it was not possible, during this work, to experiment with all the DE approaches found in literature. That would not be feasible in a single study. Rather, this can be seen as a starting point for further works that explore DE, some of which shall be suggested as follows.

This paper also did not consider the identification of component parts of definitions, as did Alarcón et al [2,3]. A future work might consider identifying those and other possible component parts of definitions, depending on their type [41]. Also it is not explored here the effects that noun phrase structures might have in identifying definitions or their component parts.

Considering the relative scarcity of POS-tagged Portuguese language texts, and the considerable size (more than 6.1 million tokens) of the BCTL corpus, a manual review of the automated POS tags generated as a byproduct of this work, would result in a valuable asset for NLP in this language.

A future work might also apply a similar approach to Brazilian Portuguese texts in domains other than legal texts, since none of the applied techniques are domain-specific.

References

1. Alarcón, R., Sierra, G., Bach, C.: Developing a Definitional Knowledge Extraction System. In: Proceedings of Third Language & Technology Conference, LTC 2007 (2007)
2. Alarcón, R., Sierra, G., Bach, C.: ECODE: A Definition Extraction System. In: Vetulani, Z., Uszkoreit, H. (eds.) LTC 2007. LNCS, vol. 5603, pp. 382–391. Springer, Heidelberg (2009)
3. Alarcón, R., Sierra, G., Bach, C.: Description and evaluation of a definition extraction system for Spanish language. In: Proceedings of the 1st Workshop on Definition Extraction, pp. 7–13. Association for Computational Linguistics, Borovets (2009)
4. Aluísio, S.M., Pinheiro, G., Finger, M., Nunes, M.G.V., Tagnin, S.E.: The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: Proceedings of Corpus Linguistics, Lancaster, UK, vol. 16, pp. 14–21 (2003)
5. Aluísio, S., Pelizzoni, J., Marchi, A.R., de Oliveira, L., Manenti, R., Marquifável, V.: An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.d.G.V. (eds.) PROPOR 2003. LNCS, vol. 2721, pp. 110–117. Springer, Heidelberg (2003)

6. Aranha, M.I., Lima, J.A.O.: *Coleção Brasileira de Direito das Telecomunicações, Grupos de Pesquisa*. v. 3. Brasília, Brazil (2009)
7. Blair-Goldensohn, S., McKeown, K.R., Schlaikjer, A.H.: Answering definitional questions: A hybrid approach. *New directions in question answering*. AAAI Press (2004)
8. Borg, C., Rosner, M., Pace, G.J.: Towards Automatic Extraction of Definitions. In: *Proceedings of the 5th Computer Science Annual Workshop, CSAW 2007* (2007)
9. Borg, C., Rosner, M., Pace, G.: Evolutionary algorithms for definition extraction. In: *Proceedings of the 1st Workshop on Definition Extraction*, pp. 26–32. Association for Computational Linguistics, Stroudsburg (2009)
10. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: *Proceedings of the 4th Language Resources and Evaluation Conference, LREC 2004, Lisbon, Portugal*, pp. 507–510 (2004)
11. BRASIL. Lei nº 8.666 (1993), <http://www3.dataprev.gov.br/sislex/paginas/42/1993/8666.html>
12. BRASIL. Lei Complementar nº 95 (1998), <http://www.lexml.gov.br/urn/urn:lex:br:federal:lei.complementar:1998-02-26;95>
13. Brill, E.: A simple rule-based part of speech tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing – ANLC*, pp. 152–155. Association for Computational Linguistics, Trento (1992)
14. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(27) (2011), <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>
15. Clark, A., Fox, C., Lappin, S. (Eds.): *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley and Sons (2010)
16. Del Gaudio, R., Branco, A.: Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach. In: Neves, J., Santos, M.F., Machado, J.M. (eds.) *EPIA 2007*. LNCS (LNAI), vol. 4874, pp. 659–670. Springer, Heidelberg (2007)
17. Del Gaudio, R., Branco, A.: Extraction of definitions in portuguese: An imbalanced data set problem. In: *Proceedings of Text Mining and Applications at EPIA* (2009)
18. Demšar, J., Zupan, B., Leban, G., Curk, T.: Orange: From Experimental Machine Learning to Interactive Data Mining. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) *PKDD 2004*. LNCS (LNAI), vol. 3202, pp. 537–539. Springer, Heidelberg (2004)
19. Fahmi, I., Bouma, G.: Learning to identify definitions using syntactic features. In: *Proceedings of the Workshop on Learning Structured Information in Natural Language Applications*, pp. 64–71. Association for Computational Linguistics, Trento (2006)
20. Feldman, R., Sanger, J.: *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press (2007)
21. Fernandes, A.D.: Answering definitional questions before they are asked. PhD Thesis. Massachusetts Institute of Technology, Cambridge, USA (2004)
22. Ferraresi, A., Zanchetta, E., Baroni, M., Bernardini, S.: Introducing and evaluating ukwac, a very large web-derived corpus of english. In: *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*, pp. 47–54. Marrakech, Marrocos (2008)

23. Kiss, T., Strunk, J.: Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32(4), 485–525 (2006)
24. Klavans, J.L., Muresan, S.: DEFINDER: Rule-based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text. In: *Proceedings of the AMIA Symposium*, pp. 1049–1049 (2000)
25. Loper, E., Bird, S.: NLTK: the Natural Language Toolkit. In: *Proceedings of the ACL 2002 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics – ETMTNLP*, vol. 1, pp. 63–70. Association for Computational Linguistics, Stroudsburg (2002)
26. Magnini, B., Cappelli, A., Tamburini, F.: Evaluation of natural language tools for italian: Evalita 2007. *Proceedings of the International Language Resources and Evaluation Conference, LREC 2008*, vol. 8, p. 2536-2543, 2008.
27. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the penn treebank. *Computational Linguistic* 19(2), 313–330 (1993)
28. Marques, N.C., Lopes, J.G.P.: A Neural Network Approach to Portuguese Part-of-Speech Tagging. In: Garcia, L.S. (ed.) *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado. CEFET-PR, Curitiba* (1996)
29. Miliaraki, S., Androutsopoulos, I.: Learning to identify single-snippet answers to definition questions. In: *Proceedings of the 20th International Conference on Computational Linguistics - COLING 2004*. Association for Computational Linguistics, Stroudsburg (2004)
30. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1318–1327 (2010)
31. Pearson, J.: *Terms in context*. John Benjamins Publishing Company (1998)
32. Pinto, A.S., Oliveira, D.: *Extracção de definições no Corpógrafo*. Faculdade de Letras da Universidade do Porto, Portugal (2004), <http://comum.rcaap.pt/bitstream/123456789/281/1/OliveiraPintoOut2004.pdf>
33. Przepiórkowski, A., Degórski, Ł., Wójtowicz, B.: Towards the automatic extraction of definitions in Slavic. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pp. 43–50. Association for Computational Linguistics, Prague (2007)
34. Rigutini, L., Diligenti, M., Maggini, M., Gori, M.: A Fully Automatic Crossword Generator. In: *Proceedings of the Seventh International Conference on Machine Learning and Applications*, pp. 362–367. IEEE Computer Society (2008)
35. Rondeau, G.: *Introduction à la Terminologie*, Québec, Gaëten Morin Editeur (1984)
36. Sager, J.C.: *A practical course in terminology processing*. J. Benjamins Pub. Co. (1990)
37. Saggion, H.: Identifying Definitions in Text Collections for Question Answering. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation* (2004)
38. Saggion, H.: Mining Profiles and Definitions with Natural Language Processing. In: Prado, H.A., Feneda, E. (Orgs.) *Emerging Technologies of Text Mining: Techniques and Applications*, IGI Global, Hershey (2008)
39. Sang, E.T.K., Bouma, G., De Rijke, M.: Developing offline strategies for answering medical questions. In: *Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains*, Pittsburgh, USA, pp. 41–45 (2005)
40. Sarmiento, L., Maia, B., Santos, D.: The Corpógrafo – a Web-based environment for corpora research. In: *Proceedings of the International Language Resources and Evaluation Conference, LREC 2004*, pp. 449–452 (2004)

41. Shaw, W.C.: *The Art of Debate*. Allyn and Bacon, New York (1922)
42. Tanev, H., Negri, M., Magnini, B., Kouylekov, M.: The DIOGENE Question Answering System at CLEF-2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004*. LNCS, vol. 3491, pp. 435–445. Springer, Heidelberg (2005)
43. Westerhout, E., Monachesi, P.: Extraction of Dutch definitory contexts for elearning purposes. In: *Proceedings of Computational Linguistics in the Netherlands, CLIN 2006* (2006)
44. Wüster, E.: Die allgemeine Terminologielehre—ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. *Linguistics* 12(119), 61–106 (1974)

A Heuristic Diversity Production Approach

Hamid Parvin, Hosein Alizadeh, Sajad Parvin, and Behzad Maleki

Nourabad Mamasani Branch, Islamic Azad University Nourabad Mamasani, Iran
hamidparvin@mamasaniiau.ac.ir, {halizadeh,s.parvin}@iust.ac.ir,
b.maleki@ut.ac.ir

Abstract. Multiple classifier systems (MCSs), or simply classifier ensembles, which combine the outputs of a set of base classifiers, have been recently emerged as a method to develop a more accurate classification system. There are two fundamental issues relating to constructing an ensemble of classifiers. The first one is how to construct a set of the base classifiers in such a way that their ensemble can be a successful one; and the second is how to combine a set of base classifiers. This paper deals with the first important issue of ensemble creation. In the paper, a new method for combining classifiers is proposed. The main idea is heuristic retraining of classifiers. Specifically, in the new method named Combinational Classifiers using Heuristic Retraining (CCHR) which proposes a new way for generating diversity in ensemble pool, a classifier is first run, then, focusing on the drawbacks of this base classifier, other classifiers are retrained heuristically. Experimental results show that the MCSs using the proposed method as the constructor of ensemble components outperform those using those using another method as the constructor of ensemble components in terms of testing accuracy.

Keywords: Classifier Fusion, Heuristic Retraining, Neural Network Ensemble.

1 Introduction

Nowadays, usage of recognition systems has found many applications in almost all fields [2], [4], [14] and [17-25]. Many researches are done to improve their performance. Most of these algorithms have provided good performance for specific problem, but they have not enough robustness for other problems. Because of the difficulty that these algorithms are faced to, recent researches are directed to the combinational methods that have more power, robustness, resistance, accuracy and generality. Although the accuracy of the classifier ensemble is not always better than the most accurate classifier in ensemble pool, its accuracy is never less than average accuracy of them [9]. Multiple Classifier Systems (MCSs) have often been outperforms a single classifier system in many pattern classification problems and many studies have shown it theoretically and experimentally. Roli and Kittler [28] articulate that the rationale behind the growing interest in multiple classifier systems (MCSs) is that the classical approach to design a pattern recognition system, which focuses on the search for the best individual classifier, has some serious drawbacks. The main drawback is that the best individual classifier for the classification task at

hand is very difficult to identify, unless deep prior knowledge is available for such a task [3]. In addition, Roli and Giacinto [9] express that it is not possible to exploit the complementary discriminatory information that other classifiers may encapsulate with only a single classifier. It is worth-noting that the motivations in favor of MCS strongly resemble those of a “hybrid” intelligent system. The obvious reason for this is that MCS can be regarded as a special-purpose hybrid intelligent system.

It is an ever-true sentence that "combining the diverse classifiers any of which performs better than a random results in a better classification performance". Generally in design of combinational classifier systems, the more diverse the results of the classifiers, the more appropriate final result. Diversity is always considered as a very important concept in classifier ensemble methodology. It is considered as the most effective factor in succeeding an ensemble. The diversity in an ensemble refers to the amount of differences in the outputs of its components (classifiers) in deciding for a given sample. Assume an example dataset with two classes. Indeed the diversity concept for an ensemble of two classifiers refers to the probability that they may produce two dissimilar results for an arbitrary input sample. The diversity concept for an ensemble of three classifiers refers to the probability that one of them produces dissimilar result from the two others for an arbitrary input sample. It is worthy to mention that the diversity can converge to 0.5 and 0.66 in the ensembles of two and three classifiers respectively. Although reaching the more diverse ensemble of classifiers is generally handful, it is harmful in boundary limit. It is very important dilemma in classifier ensemble field: the ensemble of accurate/diverse classifiers can be the best. It means that although the more diverse classifiers, the better ensemble, it is provided that the classifiers are better than random.

An Artificial Neural Network (ANN) is a model which is to be configured to be able to produce the desired set of outputs, given an arbitrary set of inputs. An ANN generally composed of two basic elements: (a) neurons and (b) connections. Indeed each ANN is a set of neurons with some connections between them. From another perspective an ANN contains two distinct views: (a) topology and (b) learning. The topology of an ANN is about the existence or nonexistence of a connection. The learning in an ANN is to determine the strengths of the topology connections. One of the most representatives of ANNs is MultiLayer Perceptron. Various methods of setting the strength of connections in an MLP exist. One way is to set the weights explicitly, using a prior knowledge. Another way is to 'train' the MLP, feeding it by teaching patterns and then letting it change its weights according to some learning rule. In this paper the MLP is used as one of the base classifiers.

Decision Tree (DT) is considered as one of the most versatile classifiers in the machine learning field. DT is considered as one of unstable classifiers. It means that it can converge to different solutions in successive trainings on same dataset with same initializations. It uses a tree-like graph or model of decisions. The kind of its knowledge representation is appropriate for experts to understand what it does [27].

The authors believe that Combinational methods usually result in the improvement of classification, because classifiers with different features and methodologies can cover drawbacks of each other. Kuncheva using Condorcet theorem has shown that combination of classifiers can usually operate better than single classifier. It means if more diverse classifiers are used in the ensemble, then error of them can considerably be reduced. Different categorizations of combinational classifier systems are

represented in [10], [28]. Valentini and Masouli divide methods of combining classifiers into two categories: generative methods, non-generative methods. In generative methods, a set of base classifiers are created by a set of base algorithms or by manipulating dataset [9]. This is done in order to reinforce diversity of base classifiers. Generally, all methods which aggregate the primary results of the fixed independent classifiers are non-generative.

Neural network ensembles as an example of combinational methods in classifiers are also becoming a hot spot in machine learning and data mining recently [26]. Many researchers have shown that simply combining the output of many neural networks can generate more accurate predictions than that of any of the individual networks. Theoretical and empirical works show that a good ensemble is one where the individual networks have both accuracy and diversity, namely the individual networks make their errors on difference parts of the input space [6], [8].

In this paper, a new method for combining classifiers is proposed. The main idea is heuristic retraining of classifiers. Specifically, in the new method named Combinational Classifiers using Heuristic Retraining (CCHR) which proposes a new way for generating diversity in ensemble pool, a classifier is first run, then, focusing on the drawbacks of this base classifier, other classifiers are retrained heuristically. Each of these classifiers looks at the data with its own attitude. The main concentration in the retrained classifiers is to leverage the error-prone data. So, retrained classifiers usually have different votes about the sample points which are close to boundaries and may be likely erroneous. Experiments show significant improvements in terms of accuracies of consensus classification. This study also investigates that focusing on which crucial data points can lead to more performance in base classifiers. Also, this study shows that adding the number of all “difficult” data points like boosting method, does not always cause a better performance. The experimental results show that the performance of the proposed algorithm outperforms some of the best methods in the literature. So empirically, the authors claim that forcing crucial data points to the training set as well as eliminating them from the training set can yield to the more accurate results, conditionally.

2 Background

In generative methods, diversity is usually made using two groups of methods. One group of these methods obtains diverse individuals by training classifiers on different training set, such as bagging [1], boosting [9], [30], cross validation [8] and using artificial training examples [13]. More details about these methods will be appeared in section 2.

Another group of methods for creating diversity employs different structures, different initial weighing, different parameters and different base classifiers to obtain ensemble individuals. For example, Rosen [29] adapted the training algorithm of the network by introducing a penalty term to encourage individual networks to be decorrelated. Liu and Yao [12] used negative correlation learning to generate negatively correlated individual neural network.

The third group is named selective approach group where the diverse components are selected from a number of trained accurate networks. For example, Opitz and

Shavlik [16] proposed a generic algorithm to search for a highly diverse set of accurate networks. Lazarevic and Obradoric [11] proposed a pruning algorithm to eliminate redundant classifiers. Navone et al. [15] proposed another selective algorithm based on bias/variance decomposition. GASEN proposed by Zhou et al. [31] and PSO based approach proposed by Fu et al. [5] also were introduced to select the ensemble components. In the rest of this paper, a new method to obtain diverse classifiers is proposed which uses manipulation of dataset structures.

Inspired from boosting method, in this paper a new sort of generative approaches is presented which creates new training sets from the original one. The base classifiers are trained focusing on the crucial and error prone data of the training set. This new approach which is called “Combination of Classifiers using Heuristic Retraining, CCHR” is described in section 2 in detail. In fact, the question of “how to create a number of diverse classifiers?” is answered in that section. Section 3 addresses the empirical studies in which we show the great accuracy and robustness of CCHR method for different datasets. Finally, section 4 discusses the concluding remarks.

2.1 Artificial Neural Network

A first wave of interest in ANN (also known as 'connectionist models' or 'parallel distributed processing') emerged after the introduction of simplified neurons by McCulloch and Pitts in 1943. These neurons were presented as models of biological neurons and as conceptual components for circuits that could perform computational tasks. Each unit of an ANN performs a relatively simple job: receive input from neighbors or external sources and use this to compute an output signal which is propagated to other units. Apart from this processing, a second task is the adjustment of the weights. The system is inherently parallel in the sense that many units can carry out their computations at the same time. Within neural systems it is useful to distinguish three types of units: input units (indicated by an index i) which receive data from outside the ANN, output units (indicated by an index o) which send data out of the ANN, and hidden units (indicated by an index h) whose input and output signals remain within the ANN. During operation, units can be updated either synchronously or asynchronously. With synchronous updating, all units update their activation simultaneously; with asynchronous updating, each unit has a (usually fixed) probability of updating its activation at a time t , and usually only one unit will be able to do this at a time. In some cases the latter model has some advantages.

An ANN has to be configured such that the application of a set of inputs produces the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to 'train' the ANN by feeding it teaching patterns and letting it change its weights according to some learning rule. For example, the weights are updated according to the gradient of the error function. For further study the reader must refer to an ANN book such as Haykin's book on theory of ANN [7].

2.2 Decision Tree Learning

DT as a machine learning tool uses a tree-like graph or model to operate deciding on a specific goal. Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

2.3 K-Nearest Neighbor Algorithm

k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbor.

As it is obvious, the k-NN classifier is a stable classifier. A stable classifier is the one converge to an identical classifier apart from its training initialization. It means the 2 consecutive trainings of the k-NN algorithm with identical k value, results in two classifiers with the same performance. This is not valid for the MLP and DT classifiers. We use 1-NN as a base classifier in the paper.

3 Proposed Method

The main idea of the proposed method is heuristically retraining of MLPs on different sets of data. In this method, base MLPs are trained on some possible permutations of 3 datasets named: TS, NS, and EPS. They are abbreviation for Train Set, Neighbor Set and Error-Prone Set, respectively. In the next step, the results of all these base classifiers are combined using simple average method.

3.1 Preparing Different Sets from the Main Dataset

Firstly, a simple Multi Layer Perceptron is trained on TS. Then, using this neural net (MLP), the data which may be misclassified are recognized. This work is done for different perspectives of training-test datasets. It means that it is tried to detect all error-prone data on TS. It can be implemented using leave-one-out technique and Cross-Validation.

In cross-validation which is also called the rotation method, an integer K (preferably a factor of N) is chosen and the dataset is randomly divided into K subsets of size N/K. Then, a classifier is trained on dataset-{i-th subset of the dataset} and evaluated using i-th subset. This procedure is repeated K times, choosing a different part for testing each time. When N=K, the method is called the leave-one-out or U-method.

Table 1. Different data combinations and reasons of their usages

Num	TS	Resultant Classifier
1	TS	Creation of base classifiers
2	TS+NS	Classification by complex boundaries with more concentration on crucial points and neighbor of errors (NS)
3	TS+EPS+NS	Classification by complex boundaries with more concentration on error prone(EPS) and crucial points(NS)
4	TS-EPS+NS	Classification by simple boundaries with more concentration on crucial points
5	TS+EPS	Classification by complex boundaries with more concentration on error prone points (EPS)
6	TS-EPS	Classification by very simple boundaries

In this paper, the dataset is divided into three partitions: training, evaluation and test sets. The leave-one-out technique is applied to train set for obtaining the Error-Prone Set, EPS. As it is mentioned, using leave-one-out technique an MLP on TS-{one of its data} is trained and evaluate whether that MLP misclassifies that out data or not. If it is misclassified we add it to EPS. As it is obvious, we run this work in number of items in training set. If training dataset is very large, the cross-validation technique can be used instead of leave-one-out technique, too.

In this study, the cross-validation technique is applied to {train set + validation set} for deriving the neighbor set, NS. Whereas the cross-validation is an iterative technique, the K-1 subset is considered to be as train set and the one subset as validation set, for each iteration. The errors in validation set are added to error set. In the next step, for each instance of error set, the nearest neighbor instance which belongs to the same label of that instance is found. This neighbor set is named NS.

```

NS: Neighbor Set,  $NS=\{\}$ ;
EPS: Error Prone Set,  $EPS=\{\}$ ;
Program CCHR
1.  $NS=FindNS()$ ;           //calculating NS
2.  $EPS=FindEPS()$ ;        // calculating EPS
3. Train 6 MLPs according to Table 1.
4. Combine the results using simple average.
End.

```

Fig. 1. The proposed CCHR algorithm

3.2 Creating an Ensemble of Diverse Classifiers

The EPS and NS are obtained from previous section. In this section some MLPs are trained based them. The more diverse and accurate base classifiers, the better results in final. So, some combinations as shown in Table 1 are used to create diversity in our ensemble. The used permutations and the reasons of their usage are shown in Table 1. Training of MLPs, using the combinations in Table 1, results in the classifiers that each of them concentrates on a special aspect of data. This can result in very good diversity in the ensemble.

In this paper, 6 MLPs are trained using different data according to Table 1. Their results are used in our classifier ensemble. Our proposed algorithm is shown in Fig 1.

3.3 Combining Classifiers

After creating diverse classifiers for our classifier ensemble, the next step is finding a method to fuse their results and make final decision. The part of making final decision is named combiner part. There are many different combiners. Combination method of base classifier decisions depend on their output type. Some traditional methods of classifier fusion which are based on soft/fuzzy outputs are as below:

Majority vote: assume that we have k classifiers. Classifier ensemble vote to class j if a little more than half of base classifiers vote to class j .

Simple average: the average of results of separate classifiers is calculated and then the class that has the most average value is selected as final decision.

Weighted average: it is like simple average except that a weight for each classifier is used for calculating that average.

In this paper, the simple average method is used to combine their results.

4 Experimental Results

This method is evaluated on two standard datasets: Wine and Iris respectively in Table 2 and 3. All the results presented are reported over 10 independent runs. Result of each of classifiers is reported on 30%, 50%, 70% and 30%, 50% of Iris and Wine as training set, respectively. Tables 2 and 3 show the method 5 that is trained on {TS+EPS} is relatively more robust than other methods. This method is concentrated on error-prone data.

Table 4 shows the result of performance of classification using our method and traditional methods comparatively.

Table 2. Average results on Iris dataset

Train set	Classifier number as Table 1						CCHR
	1	2	3	4	5	6	
70%	95.01	95.20	95.20	94.97	95.37	95.07	95.97
50%	95.95	95.75	95.87	95.89	96.24	95.78	96.60
30%	93.57	93.26	93.17	93.64	93.99	93.48	95.22

As it is obvious from Table 4, recognition ratio is improved considerably. Because of low number of features and records in Iris, the improvement is more significant on Wine dataset.

Table 3. Average results on Wine dataset

Train set	Classifier number as Table 1						CCHR
	1	2	3	4	5	6	
50%	91.58	91.64	92.66	91.98	93.77	91.29	96.74
30%	88.72	88.91	89.31	88.23	88.83	88.60	93.76

Table 4. CCHR vs. other methods

Classifier Type	Wine		Iris		
	50%	30%	70%	50%	30%
MLP	91.58	88.72	95.01	95.95	93.37
KNN	71.36	68.73	95.05	94.73	95.11
CCHR	96.74	93.76	95.97	96.60	95.22

Table 5. CCHR vs. other ensemble methods

	Wine			Iris		
	Train 30%	Train 50%	Train 70%	Train 30%	Train 50%	Train 70%
KNN	69.31	69.26	69.22	94.86	95.20	95.32
MLP	88.72	91.58	93.09	93.37	95.95	95.01
Simple Ensemble	92.70	94.05	95.41	94.77	96.00	95.03
Random Forest	88.32	93.37	95.56	91.52	94.67	96.22
Arc-X41	96.4	96.13	96.42	94.86	96.07	95.33
Arc-X42	95.52	95.73	96.22	95.33	96.20	96.07
CCHR	93.76	96.74	96.56	95.22	96.60	95.97

Table 5 shows the results of performance of classification accuracy of CCHR method and other traditional methods comparatively. These results are average of the ten independent runs of the algorithm. In this comparison, the parameter K in K-Nearest Neighbor algorithm, KNN, is set to one. Also, the average accuracy of KNN

method is reported over the 100 independent runs by randomly selecting a part of data as the training set, each time. To validate the CCHR method with harder benchmarks, an ensemble of simple MLPs is also implemented. These MLPs have the same structural parameters of the base MLPs of CCHR, i.e. two hidden layer with 10 and 5 neurons respectively in each of them. Like what is in the CCHR method, the voting method is chosen for combining their results.

The CCHR algorithm is compared with the two state of the art combination methods: random forest and boosting. Here, the ensemble size of the random forest is 21. The ensemble size for Arc-X₄₁ is 5 classifiers. While the ensemble size for Arc-X₄₂ is 11 classifiers.

5 Conclusion

This paper deals with the important issue of ensemble creation. In the paper, a new method for combining classifiers is proposed. The main idea is heuristic retraining of classifiers. Specifically, in the new method named Combinational Classifiers using Heuristic Retraining (CCHR) which proposes a new way for generating diversity in ensemble pool, a classifier is first run, then, focusing on the drawbacks of this base classifier, other classifiers are retrained heuristically. The main interesting conclusion of this paper is that emphasizing on the boundary data points, as boosting algorithm is not always very good. Although, boosting of the boundary data points in many cases is good, there are some cases of datasets where elimination of such points is better. The Monk's problem is one of such cases which deleting error-prone data leads to better results. Also, in data mining tasks which deal with huge data, the small size of ensemble is very interesting which is satisfied in the CCHR method as well.

References

1. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
2. Daryabari, M., Minaei-Bidgoli, B., Parvin, H.: Localizing Program Logical Errors Using Extraction of Knowledge from Invariants. In: Pardalos, P.M., Rebennack, S. (eds.) SEA 2011. LNCS, vol. 6630, pp. 124–135. Springer, Heidelberg (2011)
3. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*, 2nd edn. John Wiley & Sons, NY (2001)
4. Fouladgar, M.H., Minaei-Bidgoli, B., Parvin, H.: On Possibility of Conditional Invariant Detection 6881(2), 214–224 (2011)
5. Fu, Q., Hu, S.X., Zhao, S.Y.: A PSO-based approach for neural network ensemble. *Journal of Zhejiang University (Engineering Science)* 38(12), 1596–1600 (2004) (in Chinese)
6. Hansen, L.K., Salamon, P.: Neural network ensembles. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 12(10), 993–1001 (1990)
7. Haykin, S.: *Neural Networks, a comprehensive foundation*. Prentice Hall International (1999)
8. Krogh, A., Vedelsdy, J.: Neural Network Ensembles Cross Validation, and Active Learning. *Advances in Neural Information Processing Systems* 7, 231–238 (1995)
9. Kuncheva, L.I.: *Combining Pattern Classifiers, Methods and Algorithms*. Wiley, New York (2005)
10. Lam, L.: Classifier Combinations: Implementations and Theoretical Issues. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 77–86. Springer, Heidelberg (2000)

11. Lazarevic, A., Obradovic, Z.: Effective pruning of neural network classifier ensembles. In: Proc. International Joint Conference on Neural Networks, vol. 2, pp. 796–801 (2001)
12. Liu, Y., Yao, X.: Evolutionary ensembles with negative correlation learning. *IEEE Trans. Evolutionary Computation* 4(4), 380–387 (2000)
13. Melville, P., Mooney, R.: Constructing Diverse Classifier Ensembles Using Artificial Training Examples. In: Proc. of the IJCAI 2003, pp. 505–510 (2003)
14. Minaei-Bidgoli, B., Parvin, H., Alinejad-Rokny, H., Alizadeh, H., Punch, W.F.: Effects of resampling method and adaptation on clustering ensemble efficacy (2011) Online
15. Navone, H.D., Verdes, P.F., Granitto, P.M., Ceccatto, H.A.: Selecting Diverse Members of Neural Network Ensembles. In: Proc. 16th Brazilian Symposium on Neural Networks, pp. 255–260 (2000)
16. Opitz, D., Shavlik, J.: Actively searching for an effective neural network ensemble. *Connection Science* 8(3-4), 337–353 (1996)
17. Parvin, H., Minaei-Bidgoli, B.: Linkage Learning Based on Local Optima. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 163–172. Springer, Heidelberg (2011)
18. Parvin, H., Helmi, H., Minaei-Bidgoli, B., Alinejad-Rokny, H., Shirgahi, H.: Linkage Learning Based on Differences in Local Optimums of Building Blocks with One Optima. *International Journal of the Physical Sciences* 6(14), 3419–3425 (2011)
19. Parvin, H., Minaei-Bidgoli, B., Alizadeh, H.: A New Clustering Algorithm with the Convergence Proof. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS, vol. 6881, pp. 21–31. Springer, Heidelberg (2011)
20. Parvin, H., Minaei, B., Alizadeh, H., Beigi, A.: A Novel Classifier Ensemble Method Based on Class Weightening in Huge Dataset. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part II. LNCS, vol. 6676, pp. 144–150. Springer, Heidelberg (2011)
21. Parvin, H., Minaei-Bidgoli, B., Alizadeh, H.: Detection of Cancer Patients Using an Innovative Method for Learning at Imbalanced Datasets. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 376–381. Springer, Heidelberg (2011)
22. Parvin, H., Minaei-Bidgoli, B., Ghaffarian, H.: An Innovative Feature Selection Using Fuzzy Entropy. In: Liu, D. (ed.) ISNN 2011, Part III. LNCS, vol. 6677, pp. 576–585. Springer, Heidelberg (2011)
23. Parvin, H., Minaei, B., Parvin, S.: A Metric to Evaluate a Cluster by Eliminating Effect of Complement Cluster. In: Bach, J., Edelkamp, S. (eds.) KI 2011. LNCS, vol. 7006, pp. 246–254. Springer, Heidelberg (2011)
24. Parvin, H., Minaei-Bidgoli, B., Ghatei, S., Alinejad-Rokny, H.: An Innovative Combination of Particle Swarm Optimization, Learning Automaton and Great Deluge Algorithms for Dynamic Environments. *International Journal of the Physical Sciences* 6(22), 5121–5127 (2011)
25. Parvin, H., Minaei, B., Karshenas, H., Beigi, A.: A New N-gram Feature Extraction-Selection Method for Malicious Code. In: Dobnikar, A., Lotrič, U., Šter, B. (eds.) ICANNGA 2011, Part II. LNCS, vol. 6594, pp. 98–107. Springer, Heidelberg (2011)
26. Qiang, F., Shang-xu, H., Sheng-ying, Z.: Clustering-based selective neural network ensemble. *Journal of Zhejiang University Science* 6A(5), 387–392 (2005)
27. Qodmanan, H.R., Nasiri, M., Minaei-Bidgoli, B.: Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Systems with Applications* 38(1), 288–298 (2011)
28. Roli, F., Kittler, J. (eds.): MCS 2002. LNCS, vol. 2364. Springer, Heidelberg (2002)
29. Rosen, B.E.: Ensemble learning using decorrelated neural network. *Connection Science* 8(3-4), 373–384 (1996)
30. Schapire, R.E.: The strength of weak learn ability. *Machine Learning* 5(2), 1971–227 (1990)
31. Zhou, Z.H., Wu, J.X., Jiang, Y., Chen, S.F.: Genetic algorithm based selective neural network ensemble. In: Proc. 17th International Joint Conference on Artificial Intelligence, vol. 2, pp. 797–802 (2001)

Structuring Taxonomies from Texts: A Case-Study on Defining Soil Classes

Hércules Antonio do Prado^{1,2}, Edilson Ferneda¹, Francisco Carlos da Luz Rodrigues¹,
Éder Martins de Souza³, Osmar Abílio de Carvalho Jr.⁴,
and Alfredo José Barreto Luiz⁵

¹ Graduate Program on Knowledge and IT Management, Catholic University of Brasília, SGAN
916 Av. W5, 70.790-160 – Brasília, DF, Brazil

² Embrapa - Management and Strategy Secretariat,

Parque Estação Biológica - PqEB s/nº, 70.770-90 – Brasília, DF, Brazil

³ Embrapa - Centro de Pesquisa Agropecuária dos Cerrados – CPAC,
BR 020, Km 18, CxP 08223, 73.310-970 - Planaltina, DF, Brasil

⁴ University of Brasília – Human Sciences Institute, Departamento de Geografia,
Campus Universitario - ICC Ala Norte, Asa Norte, 70.910-900 – Brasília, DF, Brazil

⁵ Embrapa Environment, SP 340, Km 127,5, CxP 69, Tanquinho Velho,
13.820-000 – Jaguariúna, DF, Brazil

hercules@ucb.br, eferneda@pos.ucb.br,
carlosluzrodrigues@hotmail.com, eder@cpac.embrapa.br,
osmarjr@unb.br, alfredo@cnpma.embrapa.br

Abstract. Currently, most of the information digitally available is presented in textual form and it is largely acknowledged that, in many fields, the advance of knowledge may strongly benefit from this source of information. The treatment of this vast amount of texts by means of Text Mining (TM) techniques has produced interesting information in fields like Competitive Intelligence and Bibliometry that need to make sense from textual descriptions of facts. In this paper we approach the problem of taxonomy generation from texts, a common need from a large set of scientific disciplines. Taxonomy generation refers to building a hierarchical structure that organizes concepts in a knowledge domain. We applied TM techniques to help experts in Pedology in building taxonomy from redundant soils descriptions. The motto of the application is the fact that, in the early eighties, different organizations mapped and described equivalent classes of soils from Brazilian savannas, generating redundant descriptions with different class labels. There were produced 28 soil maps that covered 4,101 descriptions of soil classes. This profusion of redundant soil descriptions clearly represents a Babel Tower that makes difficult tasks like environment management and food production. The proposed process is based in clustering analysis and runs on the soil descriptions, performing a successive refinement of the abstractions found in soil descriptions. The method builds a frame that shows, for each cluster formed, the prototype (a representative word vector) and the soil descriptions related to that cluster. The results have been analyzed by a team of experts as input information to the laborious reasoning process involved in building concepts from the semantic relations among the soil descriptions. Without a help like the present process, the experts would

have to compare visually at least $4,101 \times 4,100 \times \dots \times 1$ soil descriptions to define the clusters, what is much more laborious.

Keywords: Text Mining, Taxonomies, Soil Science

1 Introduction

Most of the information currently available in digital form in the organizations is represented by texts. In face of the complexity involved in the modern economy [3], private companies, government, and research institutions are aware about the importance of relations that can emerge from this vast amount of information. Thus, a significant effort has been made to deal with these sets of texts in order to make sense from them.

Taxonomy generation refers to building a hierarchical structure that organizes concepts in a knowledge domain. For long, taxonomies have enabled people to deal with properties related to categories. For example, an animal taxonomy allows one to identify different kinds of animals and adopt specific relationship to each class.

In this paper, the problem of defining taxonomies from a set of texts is approached. The aiming is to provide insights to experts in a specific knowledge domain that can help them in understanding the underlying concepts. A real-world problem in the Pedology field is approached, focusing in the soil taxonomy of the Cerrado biome (as the Brazilian savannah is known). The problem arose when, in the early eighties, 28 regional institutions in Brazil had specified, separately, a map with soil classifications in the Cerrado biome (for its own region) using the Brazilian System of Soil Classification. Considering that soil types appear indistinctly in any region, the result of this *independent* classification is that many of these classes were differently described and labeled along the regions. As a matter of fact, for these classes, although using the same classification system, the regional experts had arranged the descriptors differently, beyond creating distinct labels for the specified classes. In this paper a method to help the soil experts in generating a satisfactory unified taxonomy from these already existing taxonomies is presented.

The method, based in clustering analysis, runs on the soil descriptions, performing a successive refinement of the abstractions found in this set of soil descriptions. A frame representation for each cluster is generated, showing the prototype (a representative word vector) and the soil descriptions within that cluster. It involves the cyclic application of the following steps: (i) generate configurations of clusters; (ii) interaction with the domain experts to analyze the configurations; and (iii) rebuild the configurations on the basis of the experts' analysis. A stop criterion is defined for each class as the moment when a set of original classes is considered as a unique class, under the knowledge of the involved specialists. It is expected that overcoming the difficulties brought by this profusion of redundant soil descriptions for the Cerrado biome will lead to improvements in the natural resources management.

2 Soil Classes in the Cerrado Biome

The Cerrado biome corresponds to 23% of Brazil area (Figure 1), taking 2 million km² of extension. The agricultural development, mainly in this region, is responsible for the economical achievements of Brazil in the last three decades. The natural resources management of this ecosystem requires a sound knowledge on soil properties in order to achieve and maintain a sustainable development of that biome.



Fig. 1. Brazilian Cerrado [4]

The Brazilian System of Soil Classification (SBCS) [2] is based on the Soil Taxonomy [7], a classification system developed by United States Department of Agriculture (USDA) between 1951 and 1975 that uses a set of taxonomic keys along with all words necessary for its aiming. This system accounted for the requirement, from the research community, that research communications related to a specific soil could be referred to similar soils (in terms of formation or origin) from other regions [1,5,6]. The taxonomic classes of soils are defined by properties that can be quantitatively evaluated. Some of these properties are: (i) depth, (ii) color, (iii) rupture-resistance class, (iv) texture, (v) structure (vi) cation-exchange activity, (vii) base saturation, (viii) mineralogy, (ix) amount of organic matter, and (x) amount of soluble salts [8]. See, for example, the description *textura argilosa* (for the *texture* property, meaning *clayey texture*) in the examples shown in Table 2. SBCS is a taxonomic system based in the American framework that includes six levels: (i) Order, (ii) Suborder, (iii) Great Group, (iv) Subgroup, (v) Family, and (vi) Series. These measures appear textually in the classes descriptions.

According to USDA [8], to use the soil taxonomy system, a user should classify a soil according to its properties by using the so-called “Key to Soil Orders”. The standard words prescribed by this classification system enable one to specify the classes, descriptions, respecting the six levels mentioned above. The amount of suborders, great groups, and subgroups by order, in the Brazilian System of Soil Classification, is shown in Table 1.

As previously mentioned, the concept of “keys” were applied differently to the same Cerrado soil classes, leading to an uncontrolled redundancy and turning harder the overall soil management process and the research in Pedology. Table 2 shows two equivalent soil classes generated in different institutions.

Table 1. Categories of soil in the Brazilian System of Soil Classification [2]

Order	Number of		
	Suborder	Great Group	Subgroup
01. Alissolos	2	5	12
02. Argissolos	4	10	85
03. Cambissolos	3	17	69
04. Chernossolos	4	10	31
05. Espodossolos	2	6	27
06. Gleissolos	4	15	53
07. Latossolos	4	22	90
08. Luvisolos	2	5	23
09. Neossolos	4	18	59
10. Nitossolos	2	7	22
11. Organossolos	4	9	42
12. Planossolos	3	10	44
13. Plintossolos	3	8	28
14. Verissolos	3	11	40
Total	14	153	625

Table 2. Two near equivalent classes generated with different labels

Labels in different maps	Classes descriptions
PVD7	PODZOLICO VERMELHO-AMARELO DISTROFICO argila de atividade baixa textura argilosa LATOSSOLO VERMELHO-AMARELO DISTROFICO textura argilosa e LATOSSOLO VERMELHO-ESCURO DISTROFICO textura argilosa floresta relevo plano e suave ondulado.
PE6	PODZOLICO VERMELHO-AMARELO EUTROFICO argila de atividade baixa textura argilosa PODZOLICO VERMELHO-AMARELO DISTROFICO argila de atividade baixa textura argilosa e LATOSSOLO VERMELHO-ESCURO DISTROFICO textura argilosa floresta relevo suave ondulado.

¹ We chose to keep in Portuguese the Brazilian names for the categories and the descriptions in order to make them coherent with the examples shown in the case study.

It can be observed that the descriptions are not totally equivalent, but can be aggregated into a more abstract level like, for example, “PODZOLICO VERMELHO-AMARELO DISTROFICO argila de atividade baixa textura argilosa e LATOSSOLO VERMELHO-ESCURO DISTROFICO textura argilosa floresta relevo plano suave ondulado”.

3 Material and Methods

3.1 Preparing the Data

By the time of the classes creation, experts defined two description schemes: the single units and the mapping units (exemplified in Figure 2). The first refers to a pure class description, i.e., a description that corresponds to a unique soil, while the latter includes two or more simple units, representing the occurrence of more than one soil class in a unit. It was found 1,069 single units and 3,032 mapping units. The possible types of mapping units are: (i) Classes with the symbol “+” indicating the join of two classes; (ii) Classes with the word *ambos* (in Portuguese, meaning “both”) indicating the description of two classes, composing a mapping unit; (iii) Classes with no association signal: may contain more than one soil class but no identification of composition is present.

Mapping Unit	Single Units
LATOSSOLO VERMELHO-AMARELO DISTRÓFICO A fraco e moderado textura média + LATOSSOLO VERMELHO-AMARELO eutrófico A fraco e moderado textura média relevo plano e suave ondulado	LATOSSOLO VERMELHO-AMARELO DISTRÓFICO A fraco e moderado textura média
	LATOSSOLO VERMELHO-AMARELO eutrófico A fraco e moderado textura média relevo plano e suave ondulado

Fig. 2. Examples of mapping and single units

The 1,069 single and the 3,032 mapping units were extracted (by means of a human effort for typing them) from the 28 soil maps available. The soil descriptions were named as the concatenation of the information source (the map produced by the institution that generated the classification) and the class name (see Figure 3). The distribution of soil classes, single units, and mapping units for each map can be seen in Table 3.

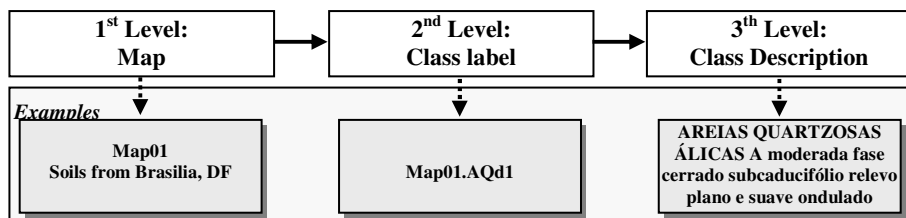


Fig. 3. Steps in data preparation

Table 3. Number of classes, single units, and mapping units in the soil maps

Map id	# classes	# single units	# mapping units	Map id	# classes	# single units	# mapping units
01	58	42	16	15	180	179	1
02	38	38	0	16	89	6	83
03	68	68	0	17	107	16	91
04	60	60	0	18	76	44	32
05	40	40	0	19	208	7	201
06	67	67	0	20	102	9	93
07	40	18	22	21	107	15	92
08	241	14	227	22	66	66	0
09	66	39	27	23	32	26	6
10	190	5	185	24	72	72	0
11	162	7	155	25	57	44	13
12	301	38	263	26	87	87	0
13	300	9	291	27	76	14	62
14	339	6	333	28	872	33	839

The proposed method takes both single and mapping units and generates a frame in which it is possible to identify single units contained within a mapping unit or even clusters of single units. In this study, text mining techniques were used to identify the relationships of repetition, equivalence, and composition among the descriptions. The pre-processing activity proceeded by standardizing the nomenclature of classes descriptions, adopting a basic lowercase letters representation for all texts.

After the separation and standardization of the classes descriptions, two final adjustments were necessary: (i) the elimination of graphical signals from Portuguese, since the TM tool did not recognize these signals (it was removed the accents, the cedilla, back ticks and other special characters), and (ii) the replacement of plural forms by singular ones. Since the TM tool does not use stemming, but pure words instead, it was necessary to adopt exactly the same form for the words variation (e.g. soil and soils had to be standardized to soil).

3.2 The Clustering Approach

We applied the Best-Star algorithm as implemented by Wives et al. [9] for grouping the texts. The algorithm works as follows. First, the more discriminant features are extracted from the texts by: (i) removing the stop words (prepositions, articles, numerals, etc.) and (ii) calculating the importance of each remaining word for each document, given by the absolute frequency of this word divided by total amount of words in the document. After that, a vector with the most important words (according to a user defined threshold) is created to represent the document. These vectors are used to calculate the similarity between each pair of documents. A symmetric matrix is built from these similarities, each cell containing the similarity between the texts corresponding to the row and the column of this cell. The overall process is shown in Figure 4:

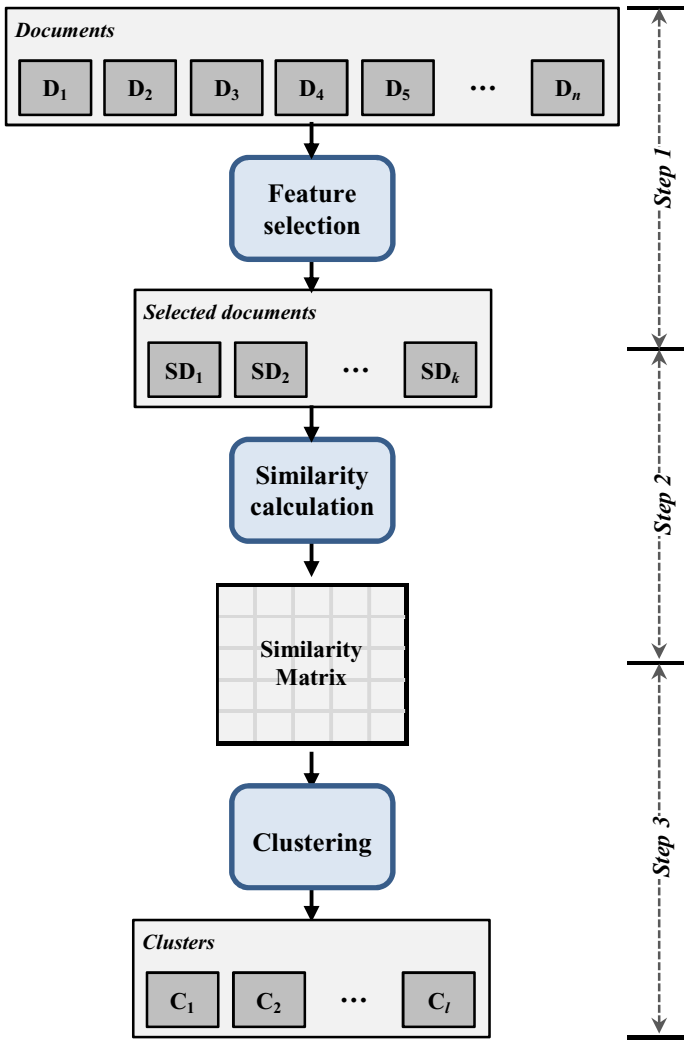


Fig. 4. The overall clustering process

- Step 1:** refers to the choice of the relevant terms for each document. The relevance of a term is given by the mentioned frequency calculation from which the most important terms for each document is obtained.
- Step 2:** here, the similarity is calculated for each pair of texts, saving the results in the corresponding matrix cells.
- Step 3:** corresponds properly to the core of Best-Star algorithm and is responsible for the formation of groups from the similarity matrix, considering the similarity parameter specified by the user. The algorithm compares each pair of texts and allocates the ones that satisfy the similarity parameter in the same cluster.

However, during the process, if a text previously allocated to a cluster happens to be more similar to other group, the algorithm reallocates it, assuring that each document will remain along with the most similar group.

3.3 Forming the Text Classes

The class formation process requires an expert analysis over the clusters frames, arranged as shown in Figure 5. Taking into account that each mapping unit possibly represents a cluster, we proceeded the clustering process into two general steps: (i) first, an exhaustive cyclic clustering process of the 1,069 single units and (ii) considering each mapping unit as a centroid for a definitive cluster, we grouped the clusters from single units around it. Many mapping units remained alone, with no single units connected. Also, many clusters from single units did not have a mapping unit to be connected to.

Composition of soil classes							
<ul style="list-style-type: none"> - LATOSSOLO VERMELHO-AMARELO DISTROFICO A fraco e moderado textura media - LATOSSOLO VERMELHO-AMARELO eutrofico A fraco e moderado textura media relevo plano e suave ondulado 							
map09lvd18	Corresponding label			Group	Source	Description	Weight
		R05C080	Cluster [23]		Map 09LVD17	LATOSSOLO VERMELHO-AMARELO DISTROFICO A fraco e moderado textura media relevo plano e suave ondulado	01% - E TEXTURA SUAVE RELEVO
					Map 15LVD4	LATOSSOLO VERMELHO-AMARELO distrofico e alico A fraco e moderado textura media relevo plano e suave ondulado.	
		R13C157	Cluster [86]		Map 16LED2	LATOSSOLO VERMELHO-ESCURO distrofico textura argilosa relevo plano e suave ondulado e LATOSSOLO VERMELHO-AMARELO distrofico textura argilosa.	01% - E LATOSSOLO TEXTURA DISTROFICO
					Map 03LVD2	LATOSSOLO VERMELHO-AMARELO distrofico textura argilosa e LATOSSOLO VERMELHO-ESCURO distrofico textura argilosa floresta relevo suave ondulado.	
		⋮					

Fig. 5. The frame representation of a cluster composition

The first step, generation of clusters from the single units, was repeated until the creation of new groups has stopped. In each cycle of this process, the clusters with the biggest diversity of classification sources (called *heterogeneous* clusters) were considered well defined single unit clusters. We considered these clusters as well defined because it is not expected to have redundancy among descriptions from the same source, but rather, more coherence. The descriptions in the clusters formed with descriptions from the same classification (called *homogeneous* clusters) source returned to a new cycle of clustering. The criteria to adjust the similarity parameter for the clustering algorithm were based on to the number of clusters formed. After analyzing the initial results and in accordance with the advice of experts, the recommended number of clusters in each cycle should range from 70 to 80. According to this criterion, the similarity parameter had to be tuned down for each cycle, after the tenth cycle. Also, after meeting the minimum similarity, the amount of clusters decayed until the unitary groups (a group with only one description) start forming.

The descriptions inside the heterogeneous clusters were stored into a container called “useful descriptions”, while the other ones were reintroduced into the clustering process. The database containing the useful descriptions can be used to generate new classes for the unified taxonomy, possibly along with a mapping unit.

4 Results and Discussion

After 25 clustering cycles, the possibility of forming new groups of heterogeneous clusters had been already exhausted. The results were stored in three repositories: (i) Database of Clusters, containing all well-defined clusters generated at each step; (ii) Database of Results, containing the list of clusters generated in a given step; (iii) Database of Reports, containing a report on the most relevant terms within a given cluster. For each group generated in the steps are registered terms of greater weight and relevance within the grouping.

It was formed 151 homogeneous clusters and, even relaxing the degree of similarity below 10%, the composition of the clusters did not vary and, therefore, there was no formation of new heterogeneous groups. This situation indicates that there are no repeated descriptions in the remained set of texts. At the 25th cycle, 441 descriptions remained not grouped (say, unitary homogeneous groups). Also, 628 heterogeneous clusters were generated. The graphical representation of clusters (Figure 5) allows the experts to figure out what descriptions must be considered to create new levels of aggregation. Notice that is not always required to have a mapping unit in the frame, but it happened in most case. Some frames have only single units and even some descriptions remained with no frame associated.

Departing from these frames, a judgment process involving the expertise of specialists is carried out. During this process, the descriptions may be rearranged according to the experts opinion. By this way, the creation of a unified map is in course.

5 Conclusions and Future Work

The main conclusion we can draw from our research comes from the observation of the relation between the used technique and the kind of text processed. Since we used a word-base TM technique over a well behaved set of texts, we generated effective structures that are helping experts in dealing with categorization of those texts. We consider the input set of texts as well-behaved due to the fact that, no matter the source of the descriptions, all of them were built on the same basic schema, the Brazilian System of Soil Classification. It assures that any expert that described a kind of soil used the same key words.

In practical terms, the completion of the work is going to take more time, since it requires a laborious task from experts to analyze each frame and propose a new class for the set of descriptions inside the frame. Although this great effort required, it is much smaller than the comparison of all pairs of description to figure out those that can be joined in the same class.

As a future work, it is expected to have the unification of the redundant soil descriptions plotted in a unique map.

References

1. Aandahl, A.R.: The first comprehensive soil classification system. *Journal of Soil and Water Conservation* 20, 243–246 (1965)
2. EMBRAPA. Centro Nacional de Pesquisa de Solos. Sistema Brasileiro de Classificação de Solos. Rio de Janeiro, Brazil (1999)
3. Prado, H.A., Oliveira, J.P.M., Ferneda, E., Wives, L.K., Silva, E.M., Loh, S.: Transforming Textual Patterns into Knowledge. In: Raisinghani, M. (ed.) *Business Intelligence in the Digital Economy*, pp. 207–227. IGI Global, Hershey (2004)
4. Sano, E.E., Rosa, R., Brito, J.L.S., Ferreira, L.G.: Mapeamento de cobertura vegetal do Bioma Cerrado: estratégias e resultados. *Embrapa Cerrados, Planaltina* (2007)
5. Simonson, R.W.: Soil classification in the United States. *Science* 137, 1027–1034 (1962)
6. Smith, G.D.: Objectives and basic assumptions of the new soil classification system. *Soil Science* 96, 6–16 (1963)
7. SOIL SURVEY STAFF. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. U.S. Soil Conservation Service Agricultural Handbook No. 436. U.S. Department of Agriculture (1975)
8. USDA - UNITED STATES DEPARTMENT OF AGRICULTURE, Keys to Soil Taxonomy. 11th edn. (2010), ftp://ftp-fc.sc.egov.usda.gov/NSSC/Soil_Taxonomy/keys/2010_Keys_to_Soil_Taxonomy.pdf
9. Wives, L.K., de Oliveira, J.P.M., Loh, S.: Conceptual Clustering of Textual Documents and Some Insights for Knowledge Discovery. In: Prado, H.A., Ferneda, E. (eds.) *Emerging Technologies of Text Mining: Techniques and Applications*, pp. 223–243. Idea Group, Hershey (2007)

Exploring Fuzzy Ontologies in Mining Generalized Association Rules

Rodrigo Moura Juvenil Ayres, Marcela Xavier Ribeiro,
and Marilde Terezinha Prado Santos

Department of Computer Science
Federal University of Sao Carlos - UFSCar
Sao Carlos, Brazil

{rodrigo_ayres,marilde,marcela}@dc.ufscar.br

Abstract. The most common use of fuzzy taxonomies in mining generalized association rules occurs in the pre-processing stage, through the concept of extended transaction. A related problem is that extended transactions lead to the generation of huge amount of candidates and rules. Beyond that, the inclusion of ancestors may to generate redundancy problems. Besides, it is possible to see that the works have only assumed the total relation between database items and taxonomy nodes. The total relation occurs when all structure items have an equivalent representative item in the dataset, and vice-versa. Furthermore, the works have been directing for the question of mining fuzzy rules, exploring linguistic terms, but few approaches have explored new steps of the mining process. In this sense, this paper proposes the extended FOntGAR algorithm, an algorithm for mining generalized association rules under all levels of fuzzy ontologies, where the relation between database items and ontology items do not need be total. In this work the generalization is done during the post-processing step.

Keywords: Key words: Generalized Association Rules, Fuzzy Ontologies, Fuzzy Taxonomies, Post-Processing, Redundancy.

1 Introduction

The mining association rules, introduced in [1], is one of the tasks in data mining. Traditional algorithms of association generate their rules based only on database items, producing an excessive amount of rules. In this sense, some algorithms use taxonomies for obtain generalized rules, in order to facilitate the user's analysis. The mining generalized association rules was introduced by [2] and [3].

According to [2], when taxonomy ancestors are inserted into database the transactions of the same are called extended transactions. Then, from such extended transactions, it is applied an algorithm for extract the final set of rules, which can be composed of traditional rules and generalized ones. This methodology can be advantageous, however, the inclusion of ancestors results the generation of many candidate itemsets, and it is extremely necessary the use measures for eliminate redundancies,

because the algorithm ends up generating redundant patterns. On the other hand, the work [4] says that the post-processing stage can produce few candidates and rules, becoming more advantageous. Moreover, the post-processing eliminates the need of to use measures for prune redundant rules, since the process is done based on the traditional patterns generated.

However, in many applications of the real world ontologies and taxonomies may not be crisp, but fuzzy [5], because these applications do not have classes of objects with pertinence criteria precisely defined [6]. In this context, Wei and Chen [5] introduced the use of fuzzy taxonomies. They considered the partial relationships possibly existing in taxonomy, where an item may partially belong to more than one parent item. For instance, tomato may partially belong to both fruit and vegetable with different degrees. However, Lee [7], highlights the existence of two different types of structures, fuzzy concept hierarchies and generalization hierarchies of fuzzy linguistic terms. In the first, a concept may have partial relationship with several generalized concepts, and the second is a structure in which upper level nodes represent more general fuzzy linguistic terms. Considering this, Wei and Chen thus defined a fuzzy taxonomic structure and considered the extended degrees of support, confidence and interest measures for mining generalized association rules.

Moreover, we want to emphasize that, regardless of the context (fuzzy or crisp), the works in mining generalized association rules simply ignore the fact of the relation between database and taxonomy items does not need be total. In the total relation, all database items have a corresponding one in the taxonomic structure, and vice-versa. On the other hand, assuming a partial relation, some dataset items may not be present in the structure leaves, and in the same way some leaf nodes may not be present in the database. In addition, the works have being directed to improve methods for obtain generalized fuzzy association rules, which are the ones composed by linguistic terms, but few of them have directed efforts for improve the mining generalized rules under fuzzy concept hierarchies.

Other significant problem is that the stage of the mining process has being little explored. Thus, this paper presents the extended FOntGAR algorithm, for mining generalized association rules using fuzzy ontologies where the relation database/ontology is not total, and the relationships of specialization/generalization varies in the interval $[0,1]$. The generalization is made during the post-processing stage, and can to occur in all levels of fuzzy ontologies. The paper is organized as follow: Section two shows some related works. Section three presents the extended FOntGAR algorithm. The section four presents the experiments, and the section five shows the conclusions.

2 Related Work

A generalized rule is an implication of the form $A \rightarrow B$, where $A \subset I$, $B \subset I$, $A \cap B = \emptyset$ and no item in B is an ancestor of any item in A . There are many works using crisp taxonomic structures, and such works explore the structures in different stages of the algorithm processing. The most common phases of use are the pre-processing

and the post-processing. In the pre-processing the generalized rules are obtained through extended transactions, which are generated before the pattern generation. On the other hand, in the post-processing the generalized rules are obtained after the generation of the traditional rules, through a sub-algorithm that uses some generalization methodology based on the patterns generated.

In [8], the mining is made using an efficient data structure. The goal is to use the structure for find rules between items in different levels of a taxonomy tree, under the assumption that the original frequent itemsets and association rules were generated in advance. Thus, the generalization occurs during the post-processing step. Also related to the post-processing, [4] proposed the GARPA algorithm. The algorithm, unlike what was proposed by [2], do not inserts ancestor items in the database transactions. The generalization was done using a method of replacement in the rules. This process obtains a final set of patterns composed by some ones that could not be generalized and generalized rules. Bay Vo and Bac Le [9] present a new algorithm for mining generalized association rules. This work is not included in the fuzzy context, but crisp. The authors develop an algorithm that scans the database one time only and uses a Tidset to compute the support of generalized itemset. A tree structure is developed for store the database used for mining frequent itemsets.

Most of the works using the fuzzy logic, in mining generalized rules, are mainly focused in to obtain generalized fuzzy association rules, which are the ones composed by fuzzy linguistic terms, such as young, tall, and others. In such approaches are used crisp taxonomies and the linguistic terms are generated based on fuzzy intervals, normally generated through clustering. Besides, these works are directed to explore quantitative or categorical attributes. In this context we can to point, for example, the works [10], [11], [12], [13] and [14].

In relation to the fuzzy concept hierarchies, according to Wei and Chen [5], the degree μ_{xy} which any node y belongs to its ancestor x can be derived based upon the notions of subclass, superclass and inheritance, and may be calculated using the max-min product combination. Specifically,

$$\mu_{xy} = \max_{\forall l: x \rightarrow y} (\min_{\forall e \text{ on } l} \mu_{le}) \tag{1}$$

Where $l: x \rightarrow y$ is one of the paths of attributes x and y , e on l is one of the edges on access l , μ_{le} is the degree on the edge e on l . If there is no access between x and y , $\mu_{xy} = 0$ [5].

The degree of the extended support (*Dsupport*) is calculated based on this μ_{xy} . If a is an attribute value in a certain transaction $t \in T$, T is the transaction set, and x is an attribute in certain itemset X , then, the degree μ_{xa} can be viewed as the one that the transaction $\{a\}$ supports x . Thus, the degree that t supports X may be obtained as follows:

$$\mu_{tX} = support_{tX} = \min_{\forall x \in X} (\max_{\forall a \in t} (\mu_{xa})) \tag{2}$$

Furthermore, an $\sum count$ operator is used to sum up all degrees that are associated with the transactions in T, in terms of how many transactions in T support X:

$$\sum_{\forall t \in T} \text{count}(\text{support}_{tX}) = \sum_{\forall t \in T} \text{count}(\mu_{tX}) \quad (3)$$

Thus, the support of a generalized association rule $X \rightarrow Y$, let $X \cup Y = Z \subseteq I$, can be obtained as follows, where $|T|$ is the total of transactions in the database:

$$\sum_{\forall t \in T} \text{count}(\mu_{tZ}) / |T| \quad (4)$$

Similarly, the confidence ($X \rightarrow Y$), called *Dconfidence*, can be obtained as follows:

$$\sum_{\forall t \in T} \text{count}(\mu_{tZ}) / \sum_{\forall t \in T} \text{count}(\mu_{tX}) \quad (5)$$

Angryk and Petry [16] investigate the application of fuzzy hierarchies to mining multi-level knowledge from large datasets via an attribute-oriented approach. Cai CH et al. [12] proposed a fuzzy multiple-level mining algorithm for extracting implicit knowledge from transactions stored as quantitative values. The proposed generalized fuzzy mining algorithm was based on Srikant and Agrawal's approach, and the items may be from any level of the given taxonomy. Chiu et al. [10] proposed a mining algorithm for discovering generalized fuzzy association rules from transaction database, based on the hierarchical relationships and cluster-based fuzzy sets tables. Differently than [5], they do not consider taxonomies with partial relationships, but they are interested in how separate the clusters in fuzzy sets, obtaining fuzzy association rules.

The work [17] introduces an algorithm to update the discovered frequent generalized itemsets. As well as both [5] and our work, the paper [17] also explores fuzzy taxonomies with partial membership degrees, however, the taxonomy of items cannot be kept unchanged, some items may be reclassified from one hierarchy tree to another for more suitable classification, for example. In [17] generalized rules are obtained like proposed in [2], where is made the use of extended transactions, as well as [5] and [18]. On the other hand, Some works, like [19] and [20] are directed to the semantic of the data mined. They use ontologies for extract associations of similarity existing between items of the database. These relations are represented in the leaves of ontology, and the authors say that the ontologies used are fuzzy; however, except by inclusion of relations of similarity, we can say they are crisp in essence, since the specialization/generalization degrees are constant 1, like crisp ontologies. The work [20] is an extension of [19], and the main differences are the introduction of a redundancy treatment and a step of generalizing non-frequent itemsets. However, both algorithms are limited, since generalizes at only one level of ontology (leaf nodes to parents).

Thus, as was demonstrated, there are few works exploring fuzzy concept taxonomies. Besides, the works are inserted in the line of mining generalized fuzzy association rules, which is a concept different, since the rules are obtained, most of the time, with the utilization of linguistic terms. Other important point is that the works consider only the total relation among database items and taxonomic structure items. Besides, it is possible to see a bias, which is the realization of the generalization process as presented in [2], exploring fuzzy taxonomies during the pre-processing stage, through extended transactions.

In this sense, we have two motivations for generalize during the post-processing: First: it is the reduction of the amount of rules generated, since previous work show that this is possible. Second motivation: it is the elimination of redundancy problems without the use of interest measures, which are, in general, very subjective. The redundancy problem is commonly derived from the use of extended transactions. However, when fuzzy taxonomies are used during the post-processing, the calculating of both support and confidence must be extended to the fuzzy context. In this sense, the quantity of database scans realized during the calculus must be checked, since it may affect the runtime of the algorithm.

3 The Extended FOntGAR Algorithm

The aim of the extended FOntGAR (*extended Fuzzy Ontology-based Generalized Association Rules Algorithm*) is to post-process a set of specialized association rules (AR) using fuzzy ontologies, in order to obtain a reduced non-redundant and more expressive set of generalized rules, considering the partial relation between database and ontology. Figure 3.1 illustrates all steps of the extended FOntGAR algorithm. The steps colored in grey are the main points of our algorithm.

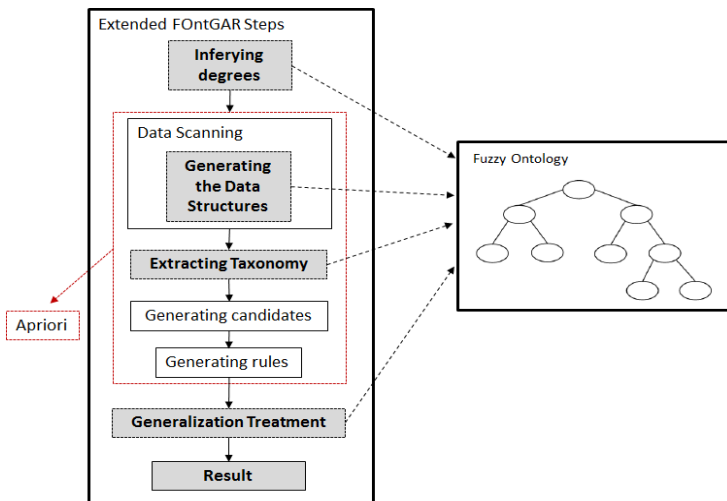


Fig. 3.1. Steps of the extended FOntGAR

3.1 Ideas Used in the Proposed Algorithm

The process of generating traditional association rules is based on Apriori [21], and as an mining association rule algorithm, it needs of an user-provided minimum support and minimum confidence parameters to run. Moreover, it needs of a minGen and a side parameters:

- *minsup*, which indicates the minimum support;
- *minconf*, which represents the minimum confidence;
- *minGen*, which represents the minimum quantity of descendants in different specialized rules;
- *side*, which represents the side to be generalized;

The *minsup*, *minconf* and *minGen* parameters are expressed by a real value in the interval $[0,1]$. The *side* parameter is expressed by a string value *left*, *right* or *lr*, indicating the generalization side. The generalization can be done on one side of the rule (antecedent or consequent) or both sides (*lr*: left and right side). While the left side indicates relations between classes of items and specialized items, the side *right* indicates relations between the specialized items and classes of items. The side *lr* indicates relations between classes.

The generalization is made through a sub-algorithm that uses a methodology of grouping and replacement in the rules. In this methodology, two or more rules are grouped in order to be replaced by a unique generalized rule. Several groups can be generated, and the grouping is done based on the parameter *side* and on the fuzzy ontology. In this case, when compared, two or more rules having identical parents, in the side of generalization, are grouped in a same group. It is important to say a group is generated only if two or more rules can be grouped, and as several groups may be generated, various generalized rules may be obtained. During the grouping, the ancestors analyzed are the immediate ones of items present on rules in question, which are the ancestor presents in the current level of generalization.

When the parameter *side* is *left* or *right*, rules having the same elements in the contrary side of generalization, and elements with identical parents in the generalization side are placed in a same group. For example, supposing ontology of bread and milk, where bread is a *breadA*, *breadB*, *breadC*, *breadD*, *breadE*, and milk is a *milka*, *milkB* *milkC*. Suppose the side was set with *right* and the traditional rules $milka \rightarrow breadA$, $milka \rightarrow breadB$, $milka \rightarrow breadC$. As all rules have the same elements in antecedent side (contrary side of generalization) and the parents of consequent items (side of generalization) is the same, these rules will be grouped together. When the parameter *side* is *lr*, rules having items with the same parents in both antecedent and consequent sides will be grouped together. For example, considering the traditional rules $milka \rightarrow breadA$, $milkB \rightarrow breadB$, $milkC \rightarrow breadC$, as all rules have the same parents in both antecedent and consequent side, these rules will be grouped together.

An important point is that generalized rules can be generated without the uses of all descendants of an ancestor. In this sense, to avoid an over-generalization, a set of specialized rules contained in a group can be substituted by a more general rule only if a *minGen* parameter [20] was satisfied. Consider the *minGen* value is 0.6 (60%), and the side is *lr*, the rule $milk \rightarrow bread$ will be generated even if there is no rule for each kind of bread and milk in the current group, but only if 60% of descendants of bread and milk are present in this set of rules. Assuming that a generalized rule contains summarized information and represents a general knowledge, the analysis of the rule $milk \rightarrow bread$ could be affected, since it does not cover all kinds of bread and milk, thus, the use of *minGen* could produce semantic loss. In this sense, in order to guide the user's comprehension, the algorithm show the items which have not participate in the generalization process. For example, suppose the item *breadE* is not

present in the specialized AR set, the generalized rule are shown as $\text{milk} \rightarrow \text{bread}$ (-breadE), indicating that the item breadE did not compose the generalization.

However, apart from such ideas, the algorithms supports situation where de relation between database items and ontology leaves are partial. In this sense, the uses of *mingen* was adapted to the new explored concept, for instance, assume that in the used ontology milk is represented by 5 (five) types of milk, milk_1 , milk_2 , milk_3 , milk_4 e milk_5 , and in the used dataset there are only 4 (four) types of milk, milk_1 , milk_2 , milk_3 , milk_4 . In this sense, the item milk_5 will never occur in the rules, because the mining is made in the data and not in the structure. Thus, when to verify if the minGen is satisfied must be considered that 100% of *milk* descendants are, milk_1 , milk_2 , milk_3 , milk_4 , without include milk_5 , because the same do not is part of the actual context. Thus, to facilitate the process of structure items identification that really are part of the database, assisting the minGen verification process, the algorithm extracts from the ontology a taxonomy where all items are present in the dataset. So, all times the minGen is checked the mentioned taxonomy is verified.

In this research, to represent fuzzy ontologies we follow the meta-ontology proposed in [22], which is an upper ontology as it represents fuzzy constructs to be inherited and/or instantiated by specific domain ontologies. Such ontology is based on OWL DL [23], a W3C recommendation supported by several reasoners and application programming interfaces used to develop ontology-based applications. Figure 3.2 illustrates how to model a fuzzy class, considering an example of domain ontology that inherits and instantiates fuzzy constructs of the upper ontology. Instances are colored in grey, fuzzy constructs are identified by the fuz: prefix and domain-specific ontology elements contain the veg: prefix. Instances of the fuz:FuzzyConceptMembership class are responsible for associating domain ontology individuals that have a membership degree $0 < \mu < 1$ to their correspondent fuzzy classes. In the Figure 3.2, the veg:tomato individual has a membership degree $\text{Fruit}(\text{tomato}) = 0.7$, expressing that veg:tomato belongs to the veg:Fruit class with a fuzzy degree.

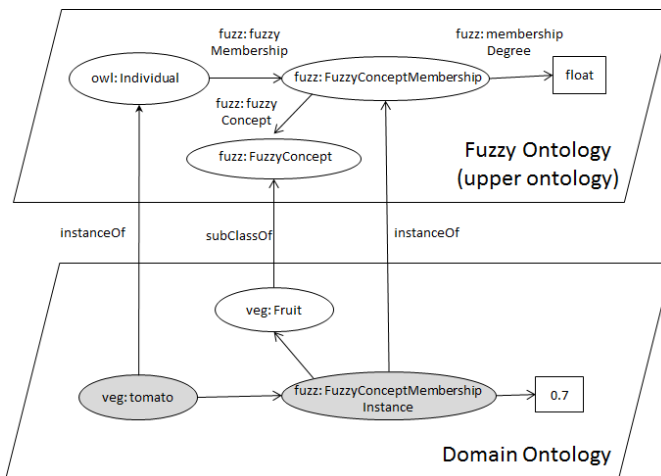


Fig. 3.2. The fuzzy meta-ontology

3.2 The Extended FOntGAR Step by Step

First, the ontology reasoner is used to infer the membership degrees of the leaves in relation to the ancestors, through the Equation 1. These degrees are stored in a data structure. The steps of data scanning, generating candidates and generating rules are done similarly to the Apriori. At end of generating rules we have a set of specialized rules, which will be used on the generalization treatment. However, during the first scanning in the database, the algorithm converts the same from the horizontal format to vertical format, including also the ontology ancestors in the conversion. This allows an immediate access to occurrences of an element. After the initial scan the algorithm uses the database items to extract the taxonomy where the relation between ontology items and dataset items is total. That structure is created only to facilitate the minGen verification. The Figure 3.3 illustrates a taxonomy extracted from ontology with more items.

Then, the generated rules and the side of generalization are passed for a groupingRules function, which is responsible to do the grouping treatment. Posteriorly, for each group generated, all rules in a group are represented by a more general rule, and for each group the minGen parameter is checked. Besides, it is verified if antecedent \cap consequent = 0 and if no consequent item is ancestor of any antecedent item. If such verifications are satisfied the calculus of support is made. If the general rule is not frequent, then the generalization do not occurs, but if the general rule is frequent, the rules of the corresponding group are replaced by the same, and it is inserted in the result. After that, if there are generalized rules, the same are used in the next level of generalization, and if this situation is true for all next levels, the generalization process will be done until one level below the ontology root. However, if there is no generalized rule at a certain level will be impossible to generalize in the next levels, and when this happens, the generalization process is concluded. Finally, after the generalization treatment, the algorithm enters its final stage, which is the results generation. In this step extended FOntGAR shows to the users all generalized rules

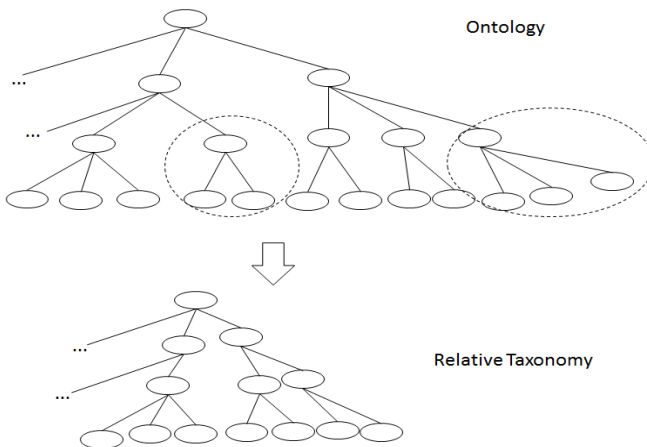


Fig. 3.3. Taxonomy extract from an ontology with more items in relation to the database

3.3 Calculating the Degrees of Support and Confidence

Considering the fuzzy taxonomy of Figure 3.4, $Fruit \rightarrow Meat$ is a generalized rule and {Fruit, Meat} is their itemset format. The support is calculated based on the sum of all degrees of transactions that support simultaneous occurrences of {Fruit, Meat}. However, {Fruit, Meat} is obtained and known only during the post-processing. Then, for obtain the degree of each transaction it would be necessary a new scanning in the database. As many generalized rules may be generated, the quantity of new scanning also may be huge, and depending on the quantity of rows of the database, the performance of the algorithm would be affected. In extended FOntGAR we use two data structures (Figure 3.4 and Figure 3.5) to allow the calculating of support avoiding additional scan. Such structures are composed by keys and values. In Figure 3.4, a key is an item of the database or ancestor of the ontology. Each key points a value, which is a vector storing the identifiers of the database transactions where the key appear.

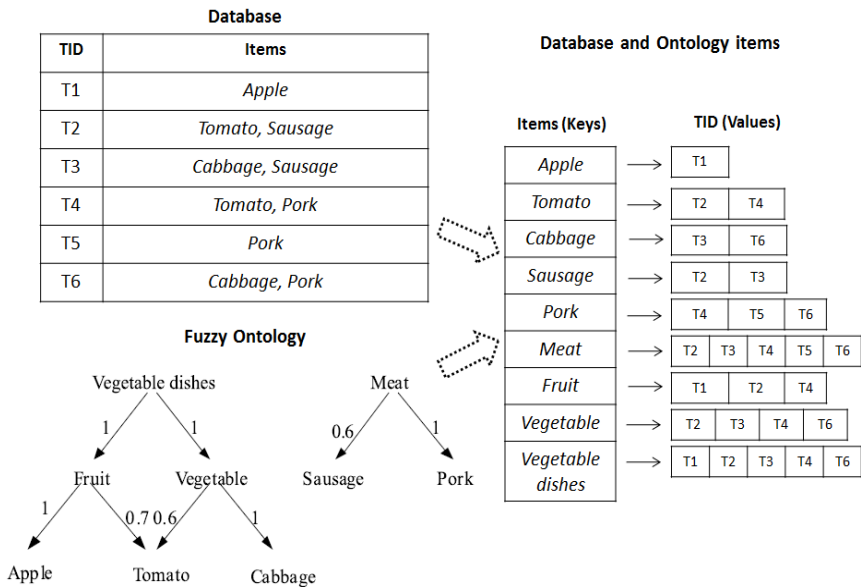


Fig. 3.4. Data structure used for store items and ancestors

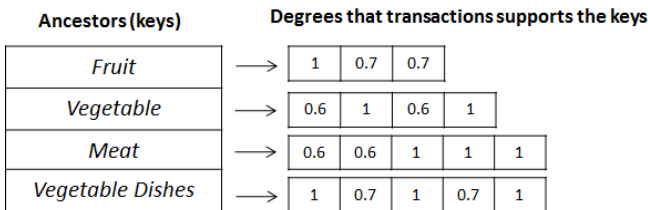


Fig. 3.5. Data structure used for store de degrees of transactions

The equation used in the calculus of support is derived from the Equation 2 (section two). So, if we partitioned the same in two subparts (Part 1 and Part 2), we have:

- **Part 1** = $\max_{\forall a \in t} (\mu_{xa})$
- **Part 2** = $\min_{\forall x \in X} (\mathbf{Part\ 1})$.

As said, we can have many generalized rules, but we don't know what will be generated. So, the itemset format of each may be any $X = \{x_1, \dots, x_n\}$, where X is the generalized rule, and x_1, \dots, x_n are items of the rule.

In this way, first we do the computation of Part 1, which is the degree that each transaction t supports an ancestor x . Based on the results of Equation 1, found at beginning of the algorithm, these degrees are calculated and stored in a data structure (Figure 3.5), where a key is the ancestor x (which will be present in generalized rules), and each key points a value, which is a vector storing the degrees mentioned. Thus, since the result of Part 2 correspond to min operator for the degrees related to any rule $\{x_1, \dots, x_n\}$, we use the stored degrees of x_1, \dots, x_n for calculating the Part 2, obtaining the support of any generalized rule.

An important point is that if $\mu_{tx} = 0$ the transaction does not supports x_n , then the degree μ_{tx} is not stored in the vector. Thus, each vector linked in a key of the Figure 3.4 has the same quantity of positions of the vector pointed by the same key of the Figure 3.5. Besides, in such vectors, the values of correspondent positions are related. For example, through Figure 3.4 we can see the key Fruit is present in three transactions, T1, T2 and T4. Then, from the Figure 4.5 we can infer the degree which T1, T2 and T3 support Fruit is 1, 0.7 and 0.7, in the same order.

Now, consider an example about how calculate the support of the rule *Fruit* → *Meat*: First, the algorithm uses the structure shown in the Figure 3.4 for verify the quantity of transactions in the intersection of values stored in vectors of these keys, since it represents all simultaneous occurrences of Fruit and Meat on the dataset transactions. Figure 3.6 illustrates this idea. In this case we have two occurrences of $\{Fruit, Meat\}$. Then, in relation to each key, the algorithm uses the positions of these transactions in Figure 3.4 to found the degree which each transaction supports these ancestors. Such degrees are present in the same positions of the vectors linked at Fruit and Meat on the Figure 3.5. In this case we have: Fruit: 0.7/T2, 0.7/T4; Meat: 0.6/T2, 1/T4, which are results of Part 1. Based on these degrees, we use Part 2 to calculate the μ_{tX} , where X is $\{Fruit, Meat\}$.

For T2 we have:

$$\mu_{tX} = \min_{\forall x \in X} (Part\ 1) = \min(0.7, 0.6) = 0.6$$

For T4 we have:

$$\mu_{tX} = \min_{\forall x \in X} (Part\ 1) = \min(0.7, 1) = 0.7$$

So, according to Equation 3, we have $0.6 + 0.7 = 1.3$. Furthermore, the Equation 4 is used to calculate the support, which is 0.21. Although we presented a specific example, the process applies to any rule.

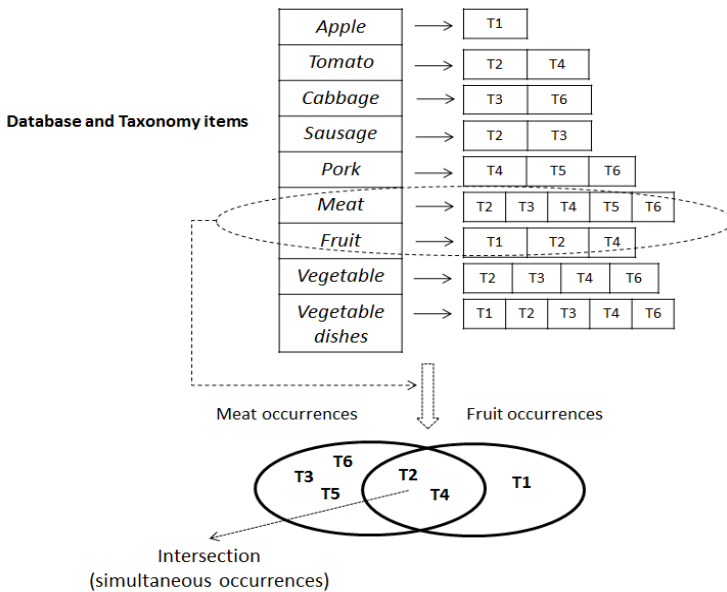


Fig. 3.6. Representation of idea used in the calculating of support

Several approaches using data structure similar to Figure 3.4, the work [9] cited previously is an example. Some works call it of database in a vertical format, since they use the database items. However we are also considering the inclusion of ancestor items. Is important to say that this structure allows an indexed access to data, unlike what happens in the database scans.

4 Experiments

This section shows some experiments performed to validate the algorithm. A real dataset was used. The dataset (DB-1) contains information about Years of study, Race or ethnicity and Sex, and was provided by Brazilian Institute of Geography and Statistics (IBGE). DB-1 contains 10000 transactions with 12 distinct items.

A fuzzy ontology, called Ont-1 ontology, was created. The Ont-1 was constructed contained one level of abstraction, except by the root, and the average value of specialization/generalization degrees was 0.8. The ontology was modeled in OWL (Web Ontology Language) and the Jena Framework was used to allow navigation through ontology concepts and relations. We also included the concept of partial relation between database items and ontology items.

In order to compare and illustrate the performance of extended FOnTGAR, the experiments were carried out with respect to a major aspect. With the DB-1, the GARPA algorithm [4] under a corresponding crisp taxonomy, NARFO [20] under a corresponding crisp ontology and extended FOnTGAR algorithm under the Ont-1 were run. The purpose was to show what the effect of fuzzy extensions could be. In this comparison, 2 experiments were conducted.

4.1 Comparisons

In such comparisons we performed 2 experiments with the real data and taxonomic structure mentioned above, changing a different parameter in each experiment. The experiments were done with default values of parameter, except for the one being varied. By default, $minsup = 0.02$, $minconf = 0.4$ and $mingen = 0.2$. The side of generalization was set to l_r in all algorithms.

Number of transactions

In the Figure 4.1, the vertical axis is the average of scanning time per transaction (in milliseconds) in relation to the first scan in the database. Here was compared the first scan on NARFO and the first scan on extended FOntGAR. We varied the number of transactions from 2000 to 10000. From Figure 4.1, it is possible see that the gap between extended FOntGAR, and NARFO show that the scanning with fuzzy ontologies is more time consuming than scanning with crisp ontologies. There are two reasons. First, the membership degree calculation demands more time. Second, the data structures generation contributes for increase the runtime. However, we can see that the gap tends keep stable with the increase of the number of transactions. This show that the computational complexity is linear with the number of transactions, which is the same as the crisp algorithm. The difference between the two curves turns to be constant.

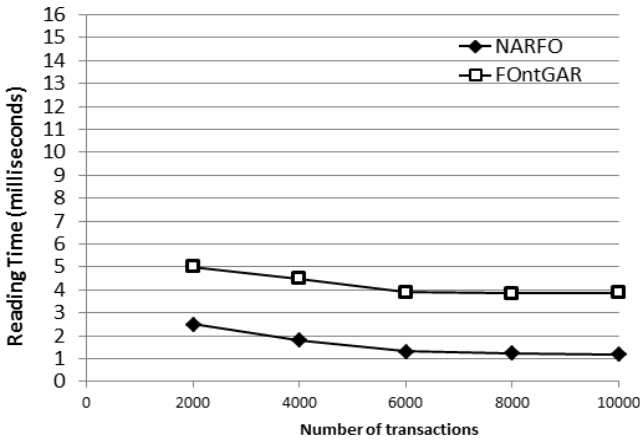


Fig. 4.1. Verification of the reading time per transaction

Minimum degree of support

In the Figure 4.2 we changed the minimum degree of support from 0.05% to 0.2%. The vertical axis is the total execution time in seconds. Notably, with the increase of $minsup$, the runtime of both extended FOntGAR and GARPA decreases. The reason is when the $minsup$ increases the amount of traditional rules decrease, and consequently a minor quantity of rules are post-processed. However, we can see that GARPA consumes more time than extended FOntGAR. The reason is that GARPA

demands more time during the calculating of support, because a new scan in the database is made for each generalized rule obtained. So, depending on the quantity of rules and rows of the dataset, the runtime can be very high. On the other hand, the data structures avoid the necessity of new scans in the database, decreasing the runtime.

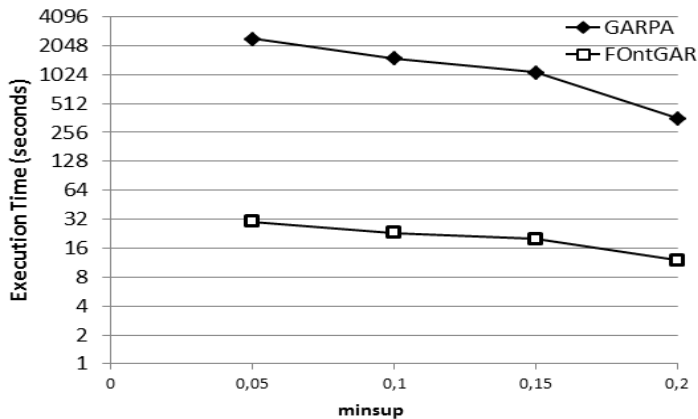


Fig. 4.2. Comparison in relation to the runtime, varying the minsup

5 Conclusions and Future Works

This paper proposes the extended FOntGAR algorithm, an algorithm for mining generalized association rules under all levels of fuzzy ontologies, including the concept of partial relation between database items and ontology nodes. The experiments show that when it is used fuzzy ontologies a complete scan in the database may be more time consuming, but the gap tends keep stable with the increase of the number of transactions. However, the total execution time of extended FOntGAR is faster than a crisp approach, and the main reason is the quantity of scans in the database, performed during the calculus of support. Thus, the data structures created in extended FOntGAR avoid the necessity of new scans during the calculating of support, allowing a more efficient access to data and decreasing the runtime.

This work presents several contributions. First and main, it is the introduction of the concept of partial relation between database items and taxonomy items. The literature shows that the most related works do not use that concept, and as introduced by [2], they also are focused only in to obtain fuzzy rules with linguistic terms. Thus, our algorithm makes an important improvement on the state of the art. Another important contribution is that extended FOntGAR generates non-redundant rules without use pruning measures, since the generalized rules are obtained based on the traditional rules, and the same, when generalized, are not included in the result. We presented our solution for the calculus of support on the post-processing stage, when using fuzzy taxonomic structures. It is also an important contribution, since solves the problem of scans in the database and improves the runtime of the algorithm.

For future works we are doing some improvements in the extended FOnTGAR algorithm. We are including user's preferences, and the introduction of context in fuzzy ontologies. Considering that all items of the database are indexed, we will to index only the items necessary for the calculus of support. So, the structures will be probably generated after the generalized rules generation.

Acknowledgments. This work has been supported by FAPESP (Sao Paulo State Research Foundation), CNPq (National Council for Scientific and Technological Development) and CAPES (Brazilian Federal Funding Agency for Graduate Education Improvement).

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases, Washington, DC, USA, pp. 207–216 (1993)
2. Srikant, R., Agrawal, R.: Mining Generalized Association Rules. Presented at the Proceedings of the 21th International Conference on Very Large Data Bases (1995)
3. Han, J., Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases. Presented at the 21 VLDB Conference, Zurich, Switzerland (1995)
4. Carvalho, V.O.D., Rezende, S.O., Castro, M.D.: Obtaining and evaluating generalized association rules. In: 9th International Conference on Enterprise Information Systems, ICEIS 2007, Funchal, Madeira, June 12-16, pp. 310–315 (2007)
5. Wei, Q., Chen, G.: Mining generalized association rules with fuzzy taxonomic structures. In: 18th International Conference of the North American Fuzzy Information Processing Society, NAFIPS 1999, pp. 477–481 (1999)
6. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)
7. Lee, J.-H., Lee-Kwang, H.: An Extension of Association Rules Using Fuzzy Sets. Presented at the 7th International Fuzzy Systems Association World Congress, Prague, Czech (1997)
8. Wu, C.-M., Huang, Y.-F.: Generalized association rule mining using an efficient data structure. *Expert Systems with Applications* 38, 7277–7290 (2011)
9. Vo, B., Le, B.: Fast Algorithm for Mining Generalized Association Rules. *International Journal of Database Theory and Application* 2, 1–12 (2009)
10. Hung-Pin, C., Yi-Tsung, T., Kun-Lin, H.: A Cluster-Based Method for Mining Generalized Fuzzy Association Rules. In: First International Conference on Innovative Computing, Information and Control, ICICIC 2006, pp. 519–522 (2006)
11. Mahmoudi, E.V., Sabetnia, E., Torshiz, M.N., Jalali, M., Tabrizi, G.T.: Multi-level Fuzzy Association Rules Mining via Determining Minimum Supports and Membership Functions. In: 2011 Second International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 55–61 (2011)
12. Cai, C.H., Fu, A.W.C., Cheng, C.H., Kwong, W.W.: Mining Association Rules with Weighted Items. In: International Database Engineering and Application Symposium, pp. 68–77 (1998)
13. Hong, T.P., Lin, K.Y., Wang, S.L.: Fuzzy data mining for interesting generalized association rules. *Fuzzy Sets and Systems* 138, 255–269 (2003)
14. Lee, Y.-C., Hong, T.-P., Wang, T.-C.: Multi-level fuzzy mining with multiple minimum supports. *Expert Systems with Applications: An International Journal* 34, 459–468 (2008)

15. Adomavicius, G., Tuzhilin, A.: Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery* 5, 33–58 (2001)
16. Angryk, R.A., Petry, F.E.: Mining Multi-Level Associations with Fuzzy Hierarchies. In: *The 14th IEEE International Conference on Fuzzy Systems*, Reno, NV, pp. 785–790 (2005)
17. Wen-Yang, L., Ming-Cheng, T., Ja-Hwung, S.: Updating generalized association rules with evolving fuzzy taxonomies. In: *2010 IEEE International Conference on Fuzzy Systems (FUZZ)*, pp. 1–6 (2010)
18. Chen, G., Wei, Q.: Fuzzy association rules and the extended mining algorithms. *Information Sciences - Informatics and Computer Science: An International Journal* 147, 201–228 (2002)
19. Escovar, E.L.G., Yaguinuma, C.A., Biajiz, M.: Using Fuzzy Ontologies to Extend Semantically Similar Data Mining. Presented at the 21 Brazilian Symposium on Databases, Florianópolis, Brazil (2006)
20. Miani, R.G., Yaguinuma, C.A., Santos, M.T.P., Biajiz, M.: NARFO Algorithm: Mining Non-redundant and Generalized Association Rules Based on Fuzzy Ontologies. In: Filipe, J., Cordeiro, J. (eds.) *ICEIS 2009. LNBIP*, vol. 24, pp. 415–426. Springer, Heidelberg (2009)
21. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. Presented at the Conference on Very Large Databases (VLDB), Santiago, Chile (1994)
22. Yaguinuma, C., Santos, M., Biajiz, M.: Fuzzy Meta-ontology for Representation of Imprecise Information in Ontologies. Presented at the II WOMSDE (2007)
23. Smith, M.K., Welt, C., McGuinness, D.L.: W3C Proposed Recommendation: OWL Web Ontology Language Guide (Dezembro 2, 2004)

BTA: Architecture for Reusable Business Tier Components with Access Control

Óscar Mortágua Pereira¹, Rui L. Aguiar¹, and Maribel Yasmina Santos²

¹ DETI, Instituto de Telecomunicações, University of Aveiro
3810-193 Aveiro, Portugal
{omp, ruilaa}@ua.pt

² Centro Algoritmi, University of Minho
4800-058 Guimarães, Portugal
maribel@dsi.uminho.pt

Abstract. Currently, business tiers for relational database applications are mostly built from software artifacts, among which Java Persistent API, Java Database Connectivity and LINQ are three representatives. Those software artifacts were mostly devised to address the impedance mismatch between the object-oriented and the relational paradigms. Key aspects as reusable business tier components and access control to data residing inside relational databases have not been addressed. To tackle the two aspects, this research proposes an architecture, referred to here as Business Tier Architecture (BTA), to develop reusable business tier components which enforce access control policies to data residing inside relational databases management systems. Besides BTA, this paper also presents a proof of concept based on Java and on Java Database Connectivity (JDBC).

Keywords: reuse, component, business tier, databases, access control.

1 Introduction

Object-oriented and relational paradigms are simply too different to bridge seamlessly, leading to a set of difficulties informally known as *impedance mismatch* [1]. Impedance mismatch derives from the diverse foundations of both paradigms and has been an open issue for more than 50 years [2]. To tackle impedance mismatch, several solutions have been devised, including Call-Level Interfaces (CLI), Embedded SQL, object-to-relational mapping techniques (O/RM), language extensions and persistent frameworks. These solutions are used to build business tiers aimed at dealing with and hiding all the complexity of the translation between the two paradigms. In spite of their key relevance to build business tiers, these solutions do not address two aspects: 1) reusability – they are not tailored to develop reusable business tiers components [3] and 2) security – they do not provide any access control mechanism to data residing inside relational database management systems (RDBMS). Next, a deeper analysis on both aspects is presented.

Reusability

Current solutions were mainly devised for tackling the impedance mismatch by providing a wide set of services to that end. Business tiers have not been foreseen as reusable components. Listing 1 resorts to JDBC [4], LINQ [5] and JPA [6] to prove it. The same CRUD expression is used (*Select top 10 o.clientId, SUM(o.value) from Orders o Group By o.clientId*) in the three cases. Components would enforce a clear separation between the development process of business tiers and the development process of application tiers. Listing 1 shows that programmers, for all the three cases, play two roles: business tier developer role and application tier developer role. They play the former role when they write CRUD expressions and source code to execute them (line 1-8; 12,13,15,16; 21-23,26-28). They play the latter role when they handle application data (line 5,6; 13-16; 22-28). Current solutions were not devised to avoid this tangling of roles.

```

// JDBC
1. String sql="Select top 10 o.clientId, SUM(o.value)...";
2. Statement st=conn.createStatement();
3. ResultSet rs=st.executeQuery(sql);
4. while (rs.next()){
5.     clientId=rs.getInt("clientId");
6.     value=rs.getFloat("value");
7.     // ... more code
8. }
9.
10.
11. //LINQ
12. String sql="Select top 10 o.clientId, SUM(o.value)...";
13. IEnumerable<Top10Orders> ord=db.ExecuteQuery<Top10Orders>(sql);
14. foreach (Top10Orders o in ord) {
15.     clientId=o.clientId;
16.     value=o.value;
17.     //... mode code
18. }
19.
20. // JPA
21. String sql="Select top 10 o.clientId, SUM(o.value)...";
22. Query qry=entityManager.createNativeQuery(sql);
23. List result=qry.getResultList();
24. for (int n=0; n<result.size(); n++) {
25.     Object ord=result.get(n);
26.     Object[] fields=(Object[]) ord;
27.     clientId=(Integer) fields[0];    // read clientId
28.     increment=(Double) fields[1];  // read value
29.     // ... more code
30. }

```

Listing 1. Examples with JDBC, LINQ and JPA

Security

Access control may be defined as being the enforced security policies which, for each database object schema (usually tables and views), control which users have access to them and the specific types of actions they are allowed to execute. For example, a user *User* may be allowed to issue *Select* and *Update* expressions on an object schema *OSchema* but not *Insert* and *Delete* expressions on the same object schema. Additionally, if necessary, it is possible to enforce a fine-grained access control at

column and at row levels. At the column level, the *User* may read a set of attributes but not another set from the same *OSchema*. At the row level, database administrators may resort to other mechanisms such as *triggers* and *Virtual Private Databases* [7]. In spite of being a key issue, current solutions do not address access control policies. Programmers of business tiers are free to write any CRUD expression. From Listing 1, we may see that programmers encode CRUD expressions inside strings (line 1,12,21) and order their execution (line 3,13,22) without any restriction. In reality, access control policies are mostly defined and enforced inside RDBMS. They are suited to control the access but not to control the information CRUD expressions they may get, this way opening a security gap. Listing 2 presents two CRUD expressions to justify this premise. Both CRUD expressions need exactly the same access control permissions: read columns *clientId* and *value*. The first CRUD expression basically retrieves raw data while the second CRUD expression retrieves critical statistical information. Access control mechanisms of RDBMS do not take into consideration the type of information being retrieved, this way opening a security gap. This situation may be overcome by controlling the CRUD expressions programmers of business tiers may use.

```

1. -- SQL 1
2. Select o.clientId, o.value
3.   from Orders o
4.    Where o.date between '2011-01-01' and '2011-12-31'
5.
6. -- SQL 2
7. Select top 10 o1.clientId, SUM(o1.value)-SUM(o2.value) as value
8.   from Orders o1, Orders o2
9.    where o1.clientId=o2.clientId and
10.         o1.date between '2011-01-01' and '2011-12-31' and
11.         o2.date between '2010-01-01' and '2011-12-31'
12.  group by o1.clientId

```

Listing 2. CRUD expressions to exemplify the security gap

To tackle these situations, this paper presents an architecture, herein referred to as Business Tier Architecture (BTA), to develop reusable business tier components that enforce access control mechanisms to data residing inside RDBMS. These issues are addressed by: 1) developing business tier components as independent software artifacts to support one or more business areas (accounting, warehousing, orders, etc.); 2) by deploying CRUD expressions at runtime to address specific users' needs and 3) in compliance to established security policies. Reusability is addressed by 1) and 2) and security is addressed by 3). Fig. 1 shows a block diagram of a database application built from a reusable business tier component derived from BTA. The component is completely decoupled from the application tier and CRUD expressions have been already deployed, at runtime, by a reliable entity. Reliable entity is any software artifact running under the supervision of database administrators or other personnel on their behalf.

Throughout this paper, all examples are based on Java, JDBC [4] and T-SQL [8] (SQL Server). The presented code may not execute properly, since we only show the relevant parts for the points under discussion.

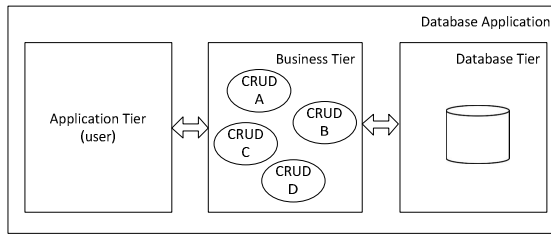


Fig. 1. Block diagram of a database application

This paper is structured as follows: section 2 presents the related work; section 3 presents the BTA; section 4 presents a proof of concept; section 5 presents a brief discussion and, finally, Section 6 presents the conclusion.

2 Related Work

Related work will be evaluated by grouping it into three categories: impedance mismatch approach, security approach and aspects approach.

Impedance mismatch approach

Several solutions have been devised to improve the development process of business tiers mainly for tackling the impedance mismatch. From those solutions, two categories have had a wide acceptance in the academic and commercial forums. Object-to-Relational mapping (O/RM) tools [9, 10] (LINQ [5], Hibernate [11], Java Persistent API (JPA) [6], Oracle TopLink [12]) and Low Level API (JDBC [4], ODBC [13], ADO.NET [14]). Other solutions, such as embedded SQL [15] (SQLJ [16]), have achieved some acceptance in the past. Others were proposed but without any general known acceptance: Safe Query Objects [17] and SQL DOM [18]. We will pick up one solution from each of the two main categories to assess them against the target of this research. O/RM tools and Low Level API have been already used in Listing 1 to support the premises about their lack of a component perspective and lack of security. Therefore, we will focus on some additional aspects about O/RM tools.

O/RM tools were devised to create in the object-oriented paradigm static representation models of relational database schemas. The static model is built in a first stage, eventually by a database administrator, and then programmers start the development process. The basic artifacts of the static representation are classes (entities), each one representing a database table. Through these entities, programmers may read data from tables, update data, insert new data and, finally, delete existing data, conveying the same result as the execution of native SQL statements. These four operations are intrinsic to entities not being possible to configure which should or which should not be made available for each table. To extend O/RM capabilities, they support several additional techniques. Among them the support for proprietary (HQL for Hibernate, JPQL for Java Persistent API, language extensions for LINQ) and native query languages is emphasized, as shown in Listing 1. All this easiness is against their capability of being used as components (it is not possible to separate the business tier and the application tier development processes) and against their

capability to cope with access control policies (there is no way to control the CRUD expressions being executed).

Low Level API were devised to feature high performance and ability to fully comply with SQL expressiveness. CRUD expressions are encoded inside strings before being executed on RDBMS. Low Level API provide a set of interfaces not only to execute any SQL statement, but also to manage local memory structures that keep data returned from Select statements. Through these local memory structures, programmers, similarly to entities of O/RM tools, may read data, insert new data, update and delete data. There is no way to avoid it. Once again, this easiness is against Low Level API to promote their capability of being used as components (it is not possible to separate the business tier and the application tier development processes) and against their capability on providing access control policies (there is no way to control the CRUD expressions to be executed).

Security approach

SELINKS [19] is a programming language in the type of LINQ and Ruby on Rails which extends Links [20]. SELINKS is focused on dealing with access control and, thus, it does not address the componentization aspect, conveying a similar behavior as LINQ does. Regarding access control, security policies are coded as user-defined functions on RDBMS. Through a type system named Fable, it is assured that sensitive data is never accessed directly without first consulting the appropriate policy enforcement function. RDBMS run policy functions to check, at runtime, which type of actions users are granted to perform, not addressing this way the security gap here presented. Moreover, queries are only statically checked, regarding the enforcement of the policies to be checked at runtime. At development and compile time, developers are not aware about the impact of the enforced policies but only aware of the existence of access control policies. This means that developers only after running the CRUD expressions have feedback about the successful or the unsuccessful execution. With BTA, programmers cannot order the execution of CRUD expressions that are not aligned with the enforced access control policies.

Jif [21] extends Java with support for information access control and also for information flow control. Jif enforces access control by providing labels to be inserted in-line with the source code of the host programming language to express security policies. Thus, regarding componentization, Jif does not enforce any architecture to that end. Regarding the implemented access control policies, they do not address the key issue of the security gap because they are mainly focused on the access to information and not to what may be done with the information. An advantage is that the access control mechanisms are enforced at compile time and at runtime. A disadvantage is that, at development time, programmers will only be aware of inconsistencies after running the Jif compiler, which is not as efficient as the strategy followed by BTA.

Rizvi et al. [22] approach relies on views to filter contents of tables and simultaneously to infer and check at runtime the appropriate authorization to execute any CRUD expression. The process is transparent and CRUD expressions are rejected if they do not have the appropriate authorization. This would overcome the security gap but authors have not assured the completeness of the inference rules. Besides, there are some additional disadvantages: 1) the inference rules are complex and time

consuming; 2) security enforcement is transparent, so users do not know that their queries are run against views; 3) programmers cannot statically check the correctness of queries which means they are not aware of access control policies at development and at compile time. Regarding the reusability, it is not addressed because the enforcement is completely implemented by views in the RDBMS.

Differential-privacy [23] has had significant attention from the research community. It is mainly focused on preserving privacy from statistical databases. It really does not directly address the points here under discussion. The interesting aspect is Frank McSherry's [24] approach to address differential-privacy: PINQ - a LINQ extension. The key aspect is that the privacy guarantees are provided by PINQ itself not requiring any expertise to enforce privacy policies. PINQ provides the integrated declarative language (SQL like, from LINQ) and simultaneously provides native support for differential-privacy for the queries being written.

Aspects approach

Aspect-oriented programming [25] (AOP) community considers persistence as a crosscutting concern [26]. Several works have been presented but none addresses the points here under consideration. The following works are emphasized: [27] is focused on separating scattered and tangled code in advanced transaction management; [26] addresses persistence relying on AspectJ; [28] presents AO4Sql as an aspect-oriented extension for SQL aimed at addressing logging, profiling and runtime schema evolution. It would be interesting to see an aspect-oriented approach for the points herein under discussion, mainly for the access control.

3 Business Tier Architecture (BTA)

BTA objective is to promote the development of reusable business tier components that enforce access control mechanisms to data residing inside RDBMS, herein known as Business Tier Components (BTC). To achieve this goal, BTA provides three key interfaces: IAdm through which reliable entities manage, at runtime, the set of CRUD expressions to be made available; IUser through which users may access to BTC internal functionalities such as the execution of CRUD expressions and, finally, BTI (Business Tier Interface), responsible for providing services to support one or more business areas and, therefore, to manage the execution of CRUD expressions. Fig. 2 presents a general BTC instance with the 3 main interfaces. In a), the BTC has no CRUD expressions yet. In b), CRUD expressions (CE1, CE2,...,CEn) have been deployed and are managed by BTI.

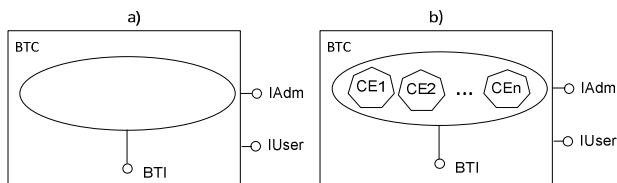


Fig. 2. General BTC instance without a) and with b) CRUD expressions

Before entering the detailed presentation, two common relational concepts are introduced: relation is a data structure returned from a relational database through a Select statement; tuple is a row of one relation.

3.1 Interfaces

Each BTC implements three main interfaces:

- BTI: this interface manages the execution of CRUD expressions;
- IAdm: manages the set of CRUD expressions available in each BTC;
- IUser: this interface is used by application tiers to instantiate artifacts aimed at managing the execution of CRUD expressions.

Business Tier Interfaces

BTI are used by application tiers to deal with the execution of CRUD expressions. Application tiers need to ask for the execution of queries and need also to access to the returned data (Select statement). CRUD expressions have four distinct types: Insert (Create), Select (Read), Update (Update) and Delete (Delete) expressions. Each CRUD expression type demands specific services and therefore specific interfaces. To tackle this situation, BTI comprises four interfaces, herein known as CRUD Interfaces, each one aimed at managing one type of CRUD expression: ISelect, IUpdate, IInsert and IDelete for Select, Update, Insert and Delete expressions, respectively. On composing these CRUD Interfaces, it came clear that the methods could be organized in smaller interfaces, herein known as Service Interfaces, each one representing coherent and disjoint functionalities. For example: 1) Update and Delete expressions return an *integer* indicating the number of affected rows; 2) local memory structures need methods to provide scrolling on their tuples; 3) every CRUD expression needs a method to trigger its execution. Thus, in order to organize the presentation of all CRUD Interfaces the presentation begins with the Service Interfaces from which the final CRUD Interfaces are built.

Service Interfaces

Each Service Interface represents a coherent functionality needed to manage the execution of one or more types of CRUD expressions. Service Interfaces are presented in Fig. 3 and then individually explained.

IExecute is associated with all types of CRUD expressions. It provides a single method to invoke the execution of CRUD expression. This method may be invoked as often as necessary with the same or with different values for its arguments.

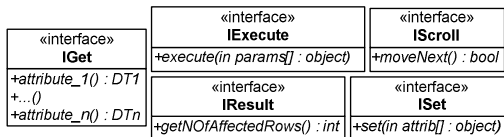


Fig. 3. Service Interfaces

`execute`: this method executes the underlying CRUD expression.

returns: void.

params: comprise all runtime parameters used in conditions inside the underlying CRUD expressions. In order to generalize its usage it may be implemented as an array of objects as follows:

```
void execute(Object[] params) {
    ...
    for (int n=0;n<params.length;n++)
        ps.setObject(n+1,params[n]);
    ...
}
```

`ps.setObject` is used to set runtime parameters of CRUD expressions as the ones below presented for *id* and *value*:

```
Select *
From Table
Where id=? and value>?
```

IResult is associated to Update and Delete CRUD expressions and is used to collect the number of effected rows as consequence of their execution.

`getNAffectedRows`: this method returns the number of affected rows as the consequence of the query execution: Update or Delete.

returns: the number of affected rows.

IScroll is used only with Select expressions. It may comprise other scrollable methods depending on the particular implementation. We have defined the only one that is mandatory.

`moveNext`: this method moves the cursor one tuple forward and points to the next available tuple.

returns: Boolean (*true* – there is next tuple, *false* – otherwise).

ISet interface is used to set the runtime values for the attributes used on Insert and Update expressions.

`Set`: this method sets the runtime values for the attributes used on Insert and Update statements. Although each CRUD expression may have its particular attributes, this method has an independent signature.

returns: void.

attrib: *attrib* comprises all the values to be used for the attributes. It is a single object and it may have several different implementations. The previous presented case for *execute* method may also be directly used in this case.

IGet interface comprises one getter method for each individual attribute of the BTI. There are as many methods as the possible different attributes of all returned relations. Each method's signature is strictly related to a returned attribute. The presented interface suggests that attributes *attrib_1*, ..., *attrib_n* are returned from the host database and their data type in the host programming language are *DT1*, ..., *DTn*, respectively. The following queries are supported by the IGet interface, provided their SQL data types are in accordance with the ones defined in the IGet interface.

```
Select attrib_1, attrib_3, attrib_7 from Student ...
Select attrib_10, attrib_1, attrib_30 from Course ...
```

Fig. 4. shows a case for a general IGet Service Interface. It comprises a set of methods, one for each attribute (attrib 1, attrib 2, ..., attrib n), to address the needs for one or more business areas. Each Select expression uses a sub-set of those methods.

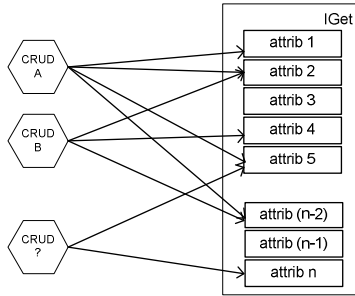


Fig. 4. General IGet Service Interface

CRUD Interfaces

CRUD Interfaces are presented in Fig. 5. They are four, one for each type of CRUD expression: Select, Insert, Update and Delete. Next, we will thoroughly explain the composition of each CRUD Interface.

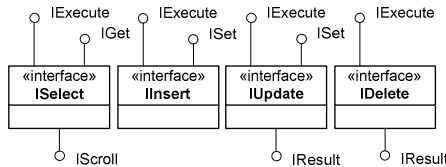


Fig. 5. CRUD Interfaces

ISelect: Select expressions are managed by the ISelect interface which comprises three Service Interfaces: 1) IExecute to execute Select statements and to set the runtime values of their parameters, 2) IGet to read the attributes of the returned relation and 3) IScroll to scroll on the returned relation.

IInsert: Insert expressions are managed by the IInsert interface which comprises two Service Interfaces: 1) IExecute to execute Insert statements and set the runtime values of their parameters and 2) ISet to set runtime values for the attribute list.

IUpdate: Update expressions are managed by the IUpdate interface which comprises three Service Interfaces: 1) IExecute to execute Update statements and set the runtime values of their runtime parameters, 2) ISet to set the runtime values of the attribute list to be updated and 3) IResult to get the number of updated rows in the host table as consequence of the query execution.

IDelete: Delete expressions are managed by the **IDelete** interface which comprises two Service Interfaces: 1) **IExecute** to execute Delete statements and set the runtime values of their parameters and 2) **IResult** to get the number of deleted rows in the host table as consequence of the query execution.

Each CRUD Interface is implemented by a class herein known as Business Tier Entity (BTE): **BTE_S**, **BTE_I**, **BTE_U** and **BTE_D** for **ISelect**, **IInsert**, **IUpdate** and **IDelete**, respectively. Each BTE type manages all CRUD expression of its type. **IInsert**, **IUpdate** and **IDelete** are generalizable and usable on all BTC, leading to a comfortable situation where **BTE_I**, **BTE_U** and **BTE_D** are reused by all BTC. Thereby, only **BTE_S** needs to be rewritten for each BTC release. This derives from the fact that **ISelect** comprises the **IGet** Service Interface which is responsible for the customization of each BTC release, regarding the attributes to be supported by the business areas being addressed.

IAdm

IAdm interface is used by reliable entities to manage the set of CRUD expressions to be made available in each running BTC instance. Through this interface, CRUD expressions may be inserted, updated and removed at runtime from the internal pool of BTC. Fig. 6 presents a simplified **IAdm** class diagram.

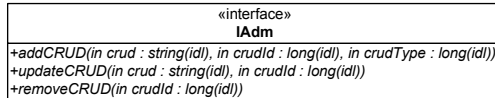


Fig. 6. Simplified **IAdm** class diagram

- addCRUD:** this method inserts a new CRUD expression into the pool.
 returns: void.
 crud: CRUD expression to be kept.
 crudId: unique Id.
 crudType: type of CRUD expression.
- updateCRUD:** this method updates a CRUD expression.
 Return: void.
 crud: new CRUD expression to be kept.
 crudId: id of CRUD expression to be updated.
- removeCrud:** this method removes a CRUD expression from the pool.
 Returns: void.
 crudId: CRUD expression to be removed.

Other methods are also necessary such as those to define the connection with the host database server. Those methods are out of scope of this paper.

IUser interface is used by application tier developers to instantiate BTE, herein known as Business Tier Workers (BTW). Fig. 7 presents the **IUser** class diagram. There is one method for each type of CRUD expression.

createBTW_S: this method creates a Business Worker of type ISelect.
 returns: ISelect.
 crudId: CRUD expression identification.

createBTW_I: this method creates a Business Worker of type IInsert.
 returns: IInsert.
 crudId: CRUD expression identification.

createBTW_U: this method creates a Business Worker of type IUpdate.
 returns: IUpdate.
 crudId: CRUD expression identification.

createBTW_D: this method creates a Business Worker of type IDelete.
 returns: IDelete.
 crudId: CRUD expression identification.

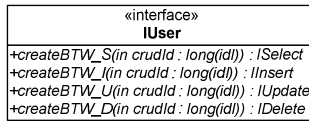


Fig. 7. Simplified IUser class diagram

3.2 Architecture

Key blocks of BTA have been presented in a scattered way: CRUD Interfaces, IAdm and IUser. In the next step, it is necessary to aggregate them in a way to address the two main objectives of BTA: component reutilization and access control. There are several approaches to achieve those objectives.

Fig. 8 presents a simplified class diagram for a possible BTA:

- IBTC is an interface that aggregates IAdm and IUser;
- the entry point is the public static method *login* (with authentication) which returns IBTC interface (IAdm for reliable entities and IUser for users);

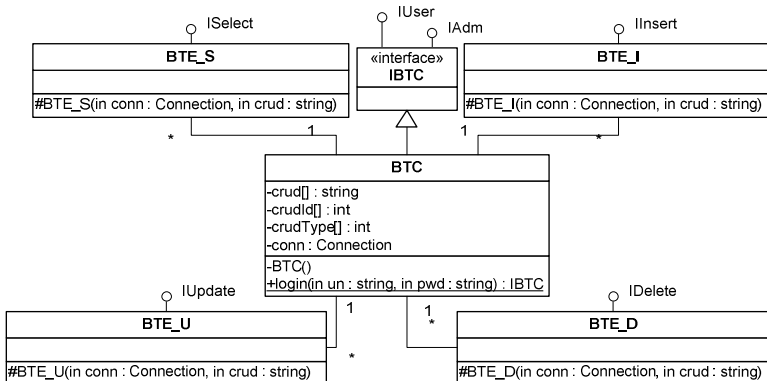


Fig. 8. Class diagram for BTA

- BTC implements IBTC;
- one BTE for each CRUD Interface;
- each BTE receives, at instantiation time, the CRUD expression to be managed and a connection object to the host database.

4 Proof of Concept

A proof of concept has been devised based on Java and JDBC for SQL Server 2008. To promote a more dynamic use of BTC, the configuration process was put inside an external component, herein known as CONFIG. The start up and running sequences are as follows:

- users try to login using the static method *login* from BTC;
- BTC creates an instance of itself;
- BTC instance loads and instantiates CONFIG at runtime using reflection;
- If user authentication is valid,
 - BTC instance is configured by CONFIG with the CRUD expressions to be made available to that user;
 - BTC returns IUser interface;
- Otherwise:
 - a null reference is returned.

Fig. 9 partially shows a BTE_S implementation. At instantiation time, the CRUD expression is compiled (line 22) using a *preparedStatement* [29]. Before being executed (line 28), all the parameters are set (line 26-27) for the CRUD expression. From now on, attributes may be read (line 32,36) and programmers may scroll (line 40) on the local memory structure (ResultSet).

Fig. 10 shows a BTC configuration process where each user gets permission to execute a pre-defined set of CRUD expressions. In this case, the user had already been authenticated and assigned to a group sharing the same set of CRUD expressions. Group identification is used to collect all CRUD expressions associated with that group (line 44) and to send them to the BTC instance (line 45-48).

Fig. 11 partially shows the use of a BTC instance from the programmers' point of view. He tries to login (line 29) and to get access to the IUser interface. In the meanwhile, as previously explained, CONFIG is instantiated by BTC and the configuration process of BTC is triggered. If authentication is correct (line 30), a BTW is instantiated (line 31) to manage the CRUD expression identified by *orderId* (*Select * from Orders where orderId=?*). If *orderId* is not a valid CRUD expression identification, an exception will be raised (not shown). CRUD expression is executed on invoking the method *execute* with an argument (line 32). If a tuple is returned (line 33), attributes are read (line 34-36).


```

16 class BTE_S implements ISelect {
17
18     private PreparedStatement ps;
19     private ResultSet rs;
20
21     BTE_S(Connection conn, String sql) throws SQLException {
22         ps=conn.prepareStatement(sql);
23     }
24     @Override
25     public void execute(Object[] params) throws SQLException {
26         for (int n=0; n<params.length; n++)
27             ps.setObject(n+1, params[n]);
28         rs=ps.executeQuery();
29     }
30     @Override
31     public int orderId() throws SQLException {
32         return rs.getInt("orderId");
33     }
34     @Override
35     public float value() throws SQLException {
36         return rs.getFloat("value");
37     }
38     @Override
39     public boolean moveNext() throws SQLException {
40         return rs.next();
41     }

```

Fig. 9. BTE_S example

```

42 void config(IAdm adm, String group) {
43     if ( userCRUD.containsKey(group) ) {
44         List<CRUD> user=userCRUD.get(group);
45         for ( int n=0; n<user.size();n++) {
46             CRUD crud=user.get(n);
47             adm.addCRUD(crud.crud(), crud.crudId(), crud.crudType());
48         }
49     }
50 }

```

Fig. 10. Configuration process of a BTC

```

29 IUser user = BTC.login(username, password);
30 if ( user != null ) {
31     ISelect s=user.createBTW_S(orderById);
32     s.execute(new Object[] {orderId});
33     if ( s.moveNext() ) {
34         orderId=s.orderId();
35         clientId=s.clientId();
36         value=s.value();
37         // .. more code
38     }

```

Fig. 11. BTC from programmers' perspective

5 Discussion

BTA and a proof of concept have been presented. There are five aspects that deserve some additional brief discussion: applicability, life-cycle, reusability, access control policies and achievements of BTA.

Applicability: the presented case study uses JDBC, leading to an easy implementation. This easiness is extended to other software artifacts suited to handle CRUD expressions, such as ODBC and ADO.NET. This has been confirmed by a component built with ADO.NET. Things are not so obvious with O/RM tools. O/RM tools are mostly oriented to handle database tables as entity classes. Nevertheless, CRUD expressions may also be handled by O/RM but that is not the focus of O/RM. Therefore, in our opinion, BTC should be implemented using Low Level Interfaces instead of O/RM.

Life-cycle: all BTC releases share most of the source code. The only exception is the BTE_S which is responsible for the IGet Service Interface implementation. If IGet is carefully planned for the business areas to be addressed, maintenance of BTC is confined to the set of CRUD expressions. CRUD expressions are edited and maintained outside BTC and are deployed to BTC at runtime. This approach allows the definition of new CRUD expressions and the updating of existent CRUD expressions without the need to proceed with any maintenance of BTC. Moreover, users may be added, removed and their pool of CRUD expressions may also be updated. Thus, BTA promotes BTC to evolve and adapt to new situations without requiring any maintenance in its core component.

Reusability: moving the configuration process to an external component promotes the reusability of BTC. Each running instance may manage a different set of CRUD expressions this way allowing a clear adaptation to each user profile.

Access control policies: access control policies have not been addressed. The access control mechanisms here presented may derive from any of the known access control policies: discretionary [30], mandatory [31, 32] and role based [33, 34].

Achievements: BTA has two main goals: reusability of business tier components and access control to data residing inside RDBMS. Reusability has been achieved by BTA by promoting the building of BTC as separated and independent software artifacts that are configurable and adaptable at runtime for each user – each user (identified by username and password) gets a customized set of CRUD expressions he may use. Access control has been achieved by defining for each user which CRUD expressions are in compliance to access control policies, and logically which actions and which information, he may execute in data residing inside RDBMS.

6 Conclusion

In this paper BTA has been presented. It addresses access control policies to data residing inside RDBMS and also reusability of business tier components. These objectives have been achieved:

- Reusability of business tier components:
 - the development process of BTC is completely decoupled from the development process of application tiers;
 - BTC are configured at runtime, set of CRUD expressions to be supported, promoting and leveraging their capability for being reused;
- Access control:
 - each user of BTCs is authenticated by username and password;

- each user may only use a set of CRUD expressions, deployed at runtime, supervised by reliable entities.

A proof of concept has been presented based on Java. BTC may be easily devised based on other Low Level Interfaces such as ODBC and ADO.NET. In spite of being possible to be used, O/RM are not the elected options to build BTC.

It is expected that the outcome of this research may contribute to open new perspectives to devise reusable business tiers components, implementing access control policies to data residing inside RDBMS.

References

1. David, M.: Representing database programs as objects. In: Bancilhon, F., Buneman, P. (eds.) *Advances in Database Programming Languages*, pp. 377–386. ACM, N.Y. (1990)
2. Cook, W., Ibrahim, A.: Integrating programming languages and databases: what is the problem? (May 2011), <http://www.odjms.org/experts.aspx#article10>
3. Heineman, G.T., Councill, W.T.: *Component-Based Software Engineering: Putting the Pieces Together*, 1st edn. Addison-Wesley (2001)
4. Parsian, M.: *JDBC Recipes: A Problem-Solution Approach*. Apress, NY (2005)
5. Erik, M., Brian, B., Gavin, B.: LINQ: Reconciling Object, Relations and XML in the .NET framework. In: *ACM SIGMOD International Conference on Management of Data*. ACM, Chicago (2006)
6. Yang, D.: *Java Persistence with JPA2010*. Outskirts Press
7. Oracle. Oracle9i Database Administrator's Guide, Release 2 (9.2). (December 2011), http://docs.oracle.com/cd/B10501_01/server.920/a96521/toc.html
8. Sack, J.: *SQL Server 2008 - Transact-SQL Recipes*. In: Gennick, J. (ed.). Apress (2008)
9. Keller, W.: *Mapping Objects to Tables - A Pattern Language*. In: *European Conference on Pattern Languages of Programming Conference (EuroPLOP)*, Irsse, Germany (1997)
10. Lammel, R., Meijer, E.: *Mappings Make data Processing Go 'Round: An Inter-paradigmatic Mapping Tutorial*. In: *Generative and Transformation Techniques in Software Engineering*. Springer, Braga (2006)
11. Christian, B., Gavin, K.: *Hibernate in Action*. Manning Publications Co. (2004)
12. Oracle. Oracle TopLink (October 2011), <http://www.oracle.com/technetwork/middleware/toplink/overview/index.html>
13. Microsoft. Microsoft Open Database Connectivity (October 2011), [http://msdn.microsoft.com/en-us/library/ms710252\(vS.85\).aspx](http://msdn.microsoft.com/en-us/library/ms710252(vS.85).aspx)
14. Mead, G., Boehm, A.: *ADO.NET 4 Database Programming with C# 2010*. Mike Murach & Associates, Inc., USA (2011)
15. Moore, J.W.: The ANSI binding of SQL to ADA. *Ada Letters* XI(5), 47–61 (1991)
16. Eisenberg, A., Melton, J.: Part 1: SQL Routines using the Java (TM) Programming Language. In: *International Committee for Information Technology American National Standard for Information for Technology Database Languages, SQLJ 1999* (1999)
17. William, R.C., Siddhartha, R.: Safe query objects: statically typed objects as remotely executable queries. In: *27th International Conference on Software Engineering*. ACM, St. Louis (2005)

18. Russell, A.M., Ingolf, H.K.: SQL DOM: compile time checking of dynamic SQL statements. In: 27th International Conference on Software Engineering. ACM, St. Louis (2005)
19. Corcoran, B.J., Swamy, N., Hicks, M.: Cross-tier, Label-based Security Enforcement for Web Applications. In: Proceedings of the 35th SIGMOD International Conference on Management of Data, pp. 269–282. ACM, Providence (2009)
20. Cooper, E., Lindley, S., Yallop, J.: Links: Web Programming Without Tiers. In: de Boer, F.S., Bonsangue, M.M., Graf, S., de Roever, W.-P. (eds.) FMCO 2006. LNCS, vol. 4709, pp. 266–296. Springer, Heidelberg (2007)
21. Zhang, D., et al. Jif: Java + information flow (December 2011), <http://www.cs.cornell.edu/jif/>
22. Rizvi, S., et al.: Extending Query Rewriting Techniques for Fine-grained Access Control. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 551–562. ACM, Paris (2004)
23. Dwork, C.: Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D.-Z., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008)
24. McSherry, F.: Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis. *Commun. ACM* 53(9), 89–97 (2010)
25. Gregor Kiczales, J.L., Mendhekar, A., Maeda, C., Videira, C.L., Loingtier, J.-M., Irwin, J.: Aspect-Oriented Programming. In: ECOOP, Jyvaskyla, Finland (1997)
26. Laddad, R.: AspectJ in Action: Practical Aspect-Oriented Programming. Manning Publications, Greenwich (2003)
27. Fabry, J., D’Hondt, T.: KALA: Kernel Aspect Language for Advanced Transactions. In: Proceedings of the 2006 ACM Symposium on Applied Computing, pp. 1615–1620. ACM, Dijon (2006)
28. Dinkelaker, T.: AO4SQL: Towards an Aspect-Oriented Extension for SQL. In: Proceedings of the 8th Workshop on Reflection, AOP and Meta-Data for Software Evolution (RAMSE 2011), Zurich, Switzerland (2011)
29. Oracle. Interface PreparedStatement (December 2011), <http://download.oracle.com/javase/6/docs/api/java/sql/PreparedStatement.html>
30. Sandhu, R.S., Samarati, P.: Access Control: Principle and Practice. *IEEE Communications Magazine* 32(9), 40–48 (1994)
31. Jajodia, S., Sandhu, R.: Toward a Multilevel Secure Relational Data Model. In: Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data, pp. 50–59. ACM, Denver (1991)
32. Lunt, T.F., et al.: The SeaView Security Model. *IEEE Transactions on Software Engineering* 16(6), 593–607 (1990)
33. Sandhu, R., Ferraiolo, D., Kuhn, R.: The NIST Model for Role-based Access Control: Towards a Unified Standard. In: Proceedings of the fifth ACM Workshop on Role-based Access Control, pp. 47–63. ACM, Berlin (2000)
34. Barker, S., Stuckey, P.J.: Flexible Access Control Policy Specification with Constraint Logic Programming. *ACM Transactions on Information and System Security* 6(4), 501–546 (2003)

Analysing the PDDL Language for Argumentation-Based Negotiation Planning

Ariel Monteserin, Luis Berdún, and Analía A. Amandi

ISISTAN Research Institute

CONICET - Universidad Nacional del Centro de la Pcia. Bs. As.
Campus Universitario, Paraje Arroyo Seco, Tandil, Argentina
{amontese, lberdun, amandi}@exa.unicen.edu.ar

Abstract. In this paper, we explore the usage of the well-known Planning Domain Description Language (PDDL) to model the argumentation-based negotiation planning. Particularly, we analyse how to define a problem planning for argumentation plan generation by using the PDDL language. Finally, we present a case study to illustrate our work, and analyse the pros and cons of this approach.

Keywords: PDDL, Argumentation-based negotiation, Planning.

1 Introduction

In a previous work, we have presented a novel approach for argumentation-based negotiation planning (Monteserin and Amandi, 2011). In multi-agent environments, negotiation is a fundamental tool to reach agreements among agents with conflictive goals.

In argumentation-based negotiation approaches (Kraus et al., 1998; Rahwan et al., 2003; Ramchurn et al., 2003; Amgoud et al., 2007; Geipel and Weiss, 2007), agents are allowed to exchange some additional information as arguments, besides the information uttered on the proposals. Thus, in the context of the negotiation, an argument is seen like a piece of information that supports a proposal and allows an agent (a) to justify its position of negotiation, or (b) to influence the position of negotiation of other agents.

In every conflictive situation where it is needed to negotiate, in real life as well as in a system composed of multiple agents, the ability to plan the course of action that it will be executed to resolve the conflict, allows the negotiator to anticipate the problems that it could find during the interaction, and also, to analyze anticipated solutions of the conflict in order to avoid or minimize its problematic effects. It is worth noticing that this anticipation is also useful to evaluate in advance several plans in order to choose the most profitable to the negotiator.

We have shown that planning aids in this task (Monteserin and Amandi, 2011). Planning algorithms are able to provide a plan of action that, when it is executed on the specified initial state, it allows the agent to achieve an expected final state. Then,

an agent can model the argumentation process as a planning problem and obtaining so, an argumentation-based negotiation plan. In this kind of plan the actions represent the arguments that the agent will use during the argumentation process, in order to persuade its opponents, and so, to reach a profitable agreement.

It is worth to remark that the arguments generated by the agent must be determined taking into account its mental state, which is composed of beliefs, goals, and preferences over type of arguments, among others. For this reason, it is necessary that the planning algorithm and the planning problem description support preferences.

In this paper, we explore the usage of the well-known PDDL standard (Ghallab et al., 1998; Gerevini and Long, 2005) to model the argumentation-based negotiation planning. Particularly, we analyse how to define a problem planning for argumentation plan generation by using the PDDL language. Finally, we present a case study to illustrate our proposal.

The paper is organized in the following way. Section 2 shows the argumentation process as a valid problem to be solved with planning. Section 3 describes the PDDL language. Section 4 shows how the argumentation problem is modelled by using PDDL. Section 5 analyses the usage of PDDL preferences to generate argumentation plans. In Section 6, a case study is presented. Section 7 discusses the pros and cons about the usage of PDDL in argumentation-based negotiation planning. Finally, in Section 8, concluding remarks are presented.

2 Defining an Argumentation Process as a Planning Problem

In this section, we explain how the argumentation process may be modelled as a planning problem in order to obtain an *argumentation plan* (Monteserin and Amandi, 2011). We define an *argumentation plan* as a partial order sequence of arguments that allows the agent to reach an expected agreement when it is uttered in a specified conflictive situation. In simpler terms, an argumentation plan will determine how the agent must perform the argumentation process during a given negotiation. That is, for instance, to establish the set of arguments and the order in which they should be presented to the counterparts in order to achieve an agreement.

For this, we will describe the main characteristic of an argumentation process, some mechanisms of a planning algorithm (those relevant to our proposal), and define conceptually a planning problem. Then, we will match every component of the argumentation process with its corresponding one in a planning problem. With the purpose of facilitating the understanding, in Figure 1 a graphical representation of this idea is showed.

First, we start depicting the argumentation-based negotiation scenario. When an agent detects a conflict, it can access to information about this conflict and the context of the conflict before the negotiation begins. By definition, the conflict that will generate the negotiation is rooted in the conflictive interest that the involved agents have, and these conflictive interest are represented, for example, in the mental states of such agents. Thus, the information that the agent can access before starting the negotiation process includes: self-information (agent's mental state, such as beliefs,

preferences and goals), information about its opponents (in realistic situations agents have only incomplete information about their opponents, because agents have some private information about its state that is unavailable to the other agents), and information about the conflict context (relevant knowledge about conflict and its resolution and historic information about past negotiations).

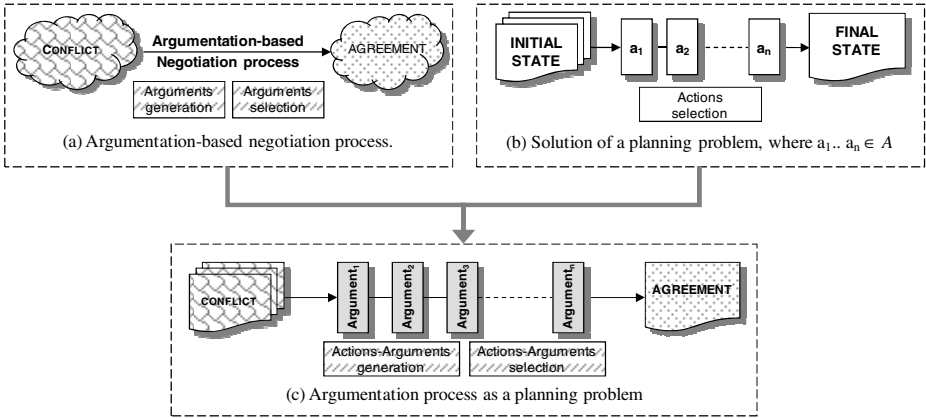


Fig. 1. Graphical representation of an argumentation process as a planning problem

Henceforth, we will call *negotiation information* to the available information to be used by the agent during the negotiation.

As we introduced earlier, agents can exchange arguments in order to justify their proposals, to persuade their opponent, and to reach an expected agreement. In contrast to agents without this argumentative ability, an argumentative agent, in addition to evaluating and generating proposals, must be able to (a) evaluate incoming arguments and update its mental state as a result; (b) generate candidate outgoing arguments; and (c) select an argument from the set of candidate arguments (Ashri et al., 2003). Thus, we can say the argumentation process is composed of the evaluation of outgoing arguments and the generation and selection of incoming arguments. To achieve this, the agent starts the argumentation process, and takes every decision concerned with this process on the basis of *negotiation information*. In other words, this information is part of the input of the evaluation, generation and selection of arguments. Also, we observe that the agent uses the argumentation process, combined with the proposal evaluation and generation, to reach an agreement in order to resolves an initial conflictive situation (See Figure 1.a).

In this work, we focus on incoming arguments; that is the generation and selection of arguments:

- *Arguments generation* is related to the generation of candidate arguments to present to a counterpart. For this end, rules for creation of arguments are defined (e.g. Kraus et al., 1998; Ramchurn et al., 2003). Such rules specify conditions for the argument generation. So, if the condition is satisfied in the negotiation context, the argument may be generated and it becomes a candidate argument.

- *Arguments selection* is concerned with selecting the argument that should be uttered to a counterpart from the set of candidate arguments generated by the arguments generation process. Once candidate arguments have been generated, then the arguments selection mechanism must apply some policy, in accordance with the agent's mental state, to select the best argument. Policies are varied, they can order all arguments by their severity and select first the weakest (Krauss et al., 1998); take into account the trust or reputation of the counterpart (Ramchurn et al., 2003); among other.

On the other hand, a planning algorithm is used to find a plan of action. In our approach, a *plan* is a partial order sequence of actions which, when it is executed in any world satisfying the initial state description, will achieve a desired final state (Weld, 1994). Conceptually, a planning problem is defined as a tree-tuple $\langle i, f, A \rangle$, where:

- i (initial state) is a complete description of the world, in which the plan will be executed.
- f (final state) describes the agent's goals. In other words, it describes where the agent wants to arrive by executing the plan.
- A (actions) is the set of available actions to build a plan. For each action its precondition and effects are defined. Thus, so that an action can be added to the plan, its preconditions must be satisfied, and it is assumed that, with its later execution, the world will be modified by the effects.

These pieces of information are the input of a planning algorithm. Internally, the planner will search for a plan, and it will use its input for that. There is one mechanism in a planning algorithm that is important for our proposal: the *action selection* mechanism. This mechanism chooses which action will be added to the plan in a particular iteration of the algorithm.

Now, we are in condition of explaining how the argumentation process may be modelled as a planning problem. In this direction, we outline how each input of the planning problem must be defined in order to generate argumentation plans:

- *Initial state*: the conflict is the beginning point of argumentation, and it is described in the negotiation information. To put it differently, the initial state describes the world where the conflict takes place.
- *Final state*: the agent's goal in a conflictive situation is to reach an expected agreement; therefore, this is the argumentation process' goal too. Thus, the final state represents the expected agreement, which is a proposal generated by the agent. Hence, the obtained argumentation plan will support the expected agreement in the same way as an argument supports a proposal.
- *Actions*: as we described above, a rule for argument generation defines the condition to create an argument, so we can think that the argument is the effect of the rule. For that reason, we can define actions to generate arguments with the same rules patterns, where the action preconditions are the condition to generate an argument and the action effect represents the argument. Furthermore, we define actions that outline the possible opponent's responses in order to represent the effects causes by argument acceptances.

Moreover, we must emulate in the planner both argument generation and selection mechanisms. For that, we describe how both mechanisms are present in the planner:

- *Argument generation*: since the rules to generate arguments can be seen as actions in a plan, when the planner establishes what actions might be added to the plan, checking its preconditions and its effects in view of the current state and the expected final state of the world, implicitly, it will be generating the candidate arguments.
- *Argument selection*: for the same reason explained above, the action selection mechanism of the planner emulates the argument selection. But, an important consideration to take into account is that the action selection mechanism in a traditional planning algorithm is implemented as a nondeterministic function, and does not consider the preferences stored in the agent’s mental state. In contrast, the selection of arguments is made on the basis of these preferences. Therefore, we need to define PDDL preferences to emulate this mechanism.

All in all, the argumentation in a negotiation process can be modelled as a planning problem representing the main characteristic of this process as inputs and mechanisms of a planning algorithm.

3 The PDDL Language

The Planning Domain Description Language (PDDL) was first proposed by Drew McDermott for the First International Planning Competition in 1998 (Ghallab et al., 1998). The language was based on Lisp syntax, using a structure based on the widely used variants of STRIPS notations. Establishing a common standard language has had a similar impact on planning research as the introduction of standards in other areas of research: it opens the route to stronger collaboration, exchange of tools, techniques and problems and provides a platform for comparative evaluation of approaches (Geverini et al., 2009).

PDDL has been extended in several stages, in order to capture more expressive variants. The significance and impact of these changes is described below, in Section 3.1. Next, Section 3.2 introduces the use of soft constraints and preferences in the PDDL language.

3.1 A Brief Review of PDDL

This section contains a short overview of PDDL. The key details of syntax and semantic of PDDL can be found in Fox and Long (2003) and Hoffmann and Edelkamp (2005).

PDDL allows actions to be described in terms of pre- and post-conditions. The expressive levels of the language are associated with tags that are used to label domain files: the addition of a tag to a domain file indicates that the domain may use the corresponding syntax layer of the language. Preconditions can be simple conjunctions of atoms (or literals, if negative preconditions are allowed), or even

arbitrary formulae (if quantification and ADL are allowed). Postconditions can contain, add and delete effects and may use conditional effects, if allowed, and also quantification.

To give an example of the language the one presented by Ghallab et al. (1998). This example considers the problem of transportation of objects using a briefcase whose effects involve both universal quantification (all objects are moved) and conditional effects (if they are inside the briefcase when it is moved). The domain is described in terms of three action schemata. PDDL encapsulates these schemata by defining the domain and listing its requirements.

```
(define (domain briefcase-world)
(:requirements :strips :equality :typing :conditional-effects)
(:types location object)
(:constants (B - object))
(:predicates (at ?x - object ?l - location)
(in ?x ?y - object))
...)
```

A domain's set of requirements allow a planner to quickly tell if it is likely to be able to handle the domain. For example, this version of the briefcase world requires conditional effects, a keyword (symbol starting with a colon) used in a *requirements* field is called a *requirement flag*; the domain is said to declare a *requirement for that flag*. All domains include a few built-in types, such as *object* (any object), and *number*. Most domains define further *types*, such as *location and object* in this domain. A *constant* is a symbol that will have the same meaning in all problems in this domain. In this case, *B* (the briefcase) is such a constant.

Inside the scope of a domain declaration, one specifies the action schemata for the domain.

```
(:action mov-b
:parameters (?m ?l - location)
:precondition (and (at B ?m) (not (= ?m ?l)))
:effect (and (at b ?l) (not (at B ?m))
(forall (?z)
(when (and (in ?z) (not (= ?z B)))
(and (at ?z ?l) (not (at ?z ?m)))))))
```

This specifies that the briefcase can be moved from *location ?m* to *location ?l*, where the symbols starting with question marks denote variables. The preconditions dictate that the briefcase must initially be in the starting location for the action to be legal and that it is illegal to try to move the briefcase to the place where it is initially. The effect equation says that the briefcase moves to its destination, is no longer where it started, and everything inside the briefcase is likewise moved.

PDDL2.1 (Fox and Long, 2003) extended the language to include number-valued fluents (with a corresponding *requirements* tag). A variant of these was included in the original PDDL specification, but had not been adopted. Two other important extensions were added in PDDL2.1, both relying on the use of numbers: plan metrics, which can be used to specify the way in which plans are to be evaluated in a specific problem instance, and durative actions. Durative actions are actions that execute over an interval of time. PDDL2.1 plan describes a trajectory of states, where states are valuations on the propositional and metric variables of the problem. The initial state is as specified for the planning problem. Transitions are caused by happenings, which are the collections of instantaneous actions that occur at the same time points.

PDDL2.2 (Edelkamp and Hoffmann, 2004) extended the language still further, adding axioms, which allow derived propositions to be inferred from the satisfaction of logical formulae in a state, and timed initial literals, which specify effects that are triggered at predetermined times during the execution of the plan.

3.2 Soft Constraints and Preferences

For the Fifth Planning Competition, PDDL was extended to include two important new features (Geverini and Long, 2005; Geverini and Long, 2006). The first is the ability to express goals that apply not only to the final state of the trajectory of states visited by a plan, but also to the intermediate states. These goals take the form of trajectory constraints, familiar from work on temporal logics. The second extension is the ability to express soft constraints, or preferences. Both of these extensions to the language are motivated by the desire to see planning bridge the gap between research and application.

A soft constraint is a condition on the trajectory generated by a plan that the user would prefer to see satisfied, but is prepared to accept might not be satisfied because of the cost involved, or because of conflicts with other constraints or goals. In PDDL3.0 are introduced quantitative preferences. An example of the expressions we wish to capture is the following: “We prefer that every fragile package is insured while it is loaded in a vehicle”.

```
(:constraints
  (and (forall (?p - package)
        (preference P1 (always (implies (and (fragile ?p) (loaded ?p))
        (insured ?p)))))) ...))
```

This example illustrates the power of combining preferences and trajectory constraints (Geverini et al., 2009).

4 Definition of the Argumentation Plan Problem Using PDDL

The negotiation information, such as the conflict context, the possible agreements, the amount of implicated agents and the agents’ mental states, changes from one negotiation to another. However, some characteristics of the negotiation process do not change; like for example, the types of argument the agent can use, the rules to generate them and its preconditions and effects. Thus, the actions, which generate the arguments, will be the same in each negotiation, whereas the initial and final states should be defined for each one. In the following sections, we show how the general predicates that are part of the negotiation language, how the initial and final states are built, and the actions for argument generation are defined by using PDDL.

4.1 Negotiation Language

As defined by Monteserin and Amandi (2011), the agent utilises a simple negotiation language L to define the planning problem. This language is composed of predicates that represent the information that the agent has in its mental state and information

about the negotiation context. According to the PDDL syntax, we first define the types that will be used during the problem definition:

- `agent`: it represents an agent that is participating in the negotiations.
- `action`: it represents an action that can be performed by an agent.
- `goal`: it represents a goal that can be pursued by an agent.
- `level`: it represents a level that some agent or negotiation attribute can take.
- `iam`: it is a subtype of `agent`. It represents the agent that is executing the planner.

The basic predicates of L expressed in PDDL are the following:

- `(bel-impl ?x ?p ?q)`: it represents the fact that agent $?x$ believes that performing action $?p$, goal $?q$ can be fulfilled.
- `(bel-fulf ?a ?z ?q ?r)`: it represents the fact that agent $?x$ believes that in the past $?z$ fulfilled goal $?g$ by performing action $?a$.
- `(isgoal ?x, ?g)`: agent $?x$ pursues goal $?g$. Agent $?x$ has $?g$ in its goals.
- `(prefer ?x ?g1 ?g2)`: $?g1$ and $?g2$ are $?x$'s goals, and $?x$ prefers to fulfil $?g1$ over $?g2$.
- `(cando ?x ?a)`: agent $?x$ can perform action $?a$. It means that agent $?x$ has the resources to perform action $?a$, or that another agent has committed to perform it.
- `(do ?x, ?a)`: agent $?x$ will perform action $?a$.
- `(pastpromise ?x ?y ?p)`: $?x$ promised execute action $?p$ to $?y$, but has not fulfilled it yet.
- `(wasgoal ?x ?g)`: $?x$ pursued goal $?g$ in the past.
- `(did ?x ?a)`: $?x$ performed action $?a$ in the past.
- `(conf ?x ?l)`: the trust in $?x$ reach a level $?l$, where $?l$ can be *low*, *medium* or *high*.
`(opposite ?g1 ?g2)`: goal $?g1$ is opposite to goal $?g2$.

We assume that the negotiation information is expressed in L .

4.2 Initial and Final States

As mentioned before, to generate argumentation plans the initial state i must describe the world where the conflict takes place. This is the negotiation information, where the information about the conflict is represented. Therefore, the initial state will be defined as the negotiation information, and since this information changes from one negotiation to another, the initial state will also vary according to the negotiation problem.

On the other hand, the final state f represents the state of agreement where the agent wants to arrive through the argumentation. Consequently, the predicates that form this state will depend on the kind of agreement that the agent can reach. In this work, we have considered agreements about task execution, but it is also possible another one. For example, if the expected agreement is that agent $ag1$ accepts to perform action `alpha`, the final state should include `(do ag1 alpha)`. However, if our agent only wants to persuade $ag1$ that performing action `alpha`, goal $g1$ can be fulfilled, the final state might be expressed as `(bel-impl ag1 alpha g1)`.

In addition, according to the PDDL specification, it is possible to have some variables on the final state, in order to instantiate them conveniently. For example, the final state may be composed of `(exists (?x - agent) (do ?x alpha))`, where `?x` will be instantiated with an opponent that the agent can persuade by executing `alpha`. So, after executing the planner, the agent will obtain the arguments that should utter and the opponent who it should negotiate with.

4.3 Actions

In order to represent the argument generation in the planning, the actions of the plan represent the arguments that the agent can utter to other counterparts. Before defining these actions, we briefly introduce the argument types that the agent can generate in the argumentation-based negotiation context. Three general argument types are defined in the literature about argumentation-based negotiation: *appeals*, *rewards* and *threats* (Rahwan et al., 2003). *Appeals* are used to justify a proposal; *rewards* to promise a future recompense; and *threats* to warn of negative effects if the counterpart does not accept a proposal.

Next, for each argument type, we will present the actions to generate it, according to the axioms defined in the framework of Kraus et al. (1998). We distinguish between two general structures of actions: *create-argument* actions and *accept-argument* actions. The first depict the argument generation such as in rules, whereas the second represent the counterpart's acceptance of the argument. Thereby, we define several *create-argument* actions to generate the same argument type, but we must only define one *accept-argument* action, whose precondition is the set of arguments, to reflect the argument effect in the current state, when the effect of each argument is the same. As PDDL does not support different structure definitions for a same predicate, we must use a disjunctive precondition (`or`) to group all the arguments with the same effects in an individual *accept-argument* action.

Appeals. Varying the premises of the appeals, we can define several of them: to past promise, counterexample, to prevailing practice, to self-interest and trivial appeals.

Moreover, we separate the appeals in two parts. The first subgroup includes appeals so that our opponents perform a given action. These are:

– Appeal to prevailing practice

```
(:action create-prev-pract-appeal
  ;Description: it proposes an action execution using as justification
  historic information about a third agent. Agent ?y believes that if it
  performs ?a, it should not fulfill ?g (?h), but agent ?x has historic
  evidence (wasgoal and did) that deny this belief.
  ;parameters (?x - iam ?y ?z - agent ?a - action ?g ?h - goal)
  ;precondition (and (not(= ?x ?y)) (not(= ?x ?z)) (not(= ?z ?y)) (isgoal
  ?y ?g) (bel-impl ?y ?a ?h) (wasgoal ?z ?g) (did ?z ?a) (opposite ?g ?h))
  ;effect (appeal-prev-pract ?x ?y ?z ?a ?g) )
```

– Counterexample

```
(:action create-counterexample-appeal
  ;Description: it is similar to the previous appeal, but the historic
  information is about its opponent.
  ;parameters (?x - iam ?y - agent ?a - action ?b - action ?g - goal ?h - goal)
```

```

:precondition (and (not(= ?x ?y)) (not(= ?a ?b)) (isgoal ?y ?g) (did ?y
?b) (bel-impl ?y ?b ?h) (opposite ?g ?h))
:effect (appeal-counterexample ?x ?y ?a ?b) )

```

– Appeal to past promise

```

(:action create-past-prom-appeal
;Description: it is used to generate appeals to past promises that the
opponent did not fulfill.
:parameters (?x - iam ?y - agent ?a - action)
:precondition (and (not(= ?x ?y)) (pastpromise ?y, ?x ?a))
:effect (appeal-past-prom ?x ?y ?a) )

```

– Appeal to self-interest

```

(:action create-self-interest-appeal
;Description: it allows the agent to generate appeals to self-interest
where the proposed action implies a profit to ?y.
:parameters (?x - iam ?y - agent ?a - action ?g - goal)
:precondition (and (not(= ?x ?y)) (isgoal ?y ?g) (bel-impl ?x ?a ?g)
(cando ?y ?a))
:effect (appeal-si ?x ?y ?a ?g) )

```

As mentioned above, we need to define an *accept-argument* action whose preconditions include all the appeals with the same effects:

```

(:action accept-appeal
;Description: it represents the acceptance of the appeal by opponent. The
effects include: the agent's compromise to perform ?a, and the capacity
of agent ?x to count on that execution (cando ?x ?a)).
:parameters (?x - iam ?y ?z - agent ?a ?b - action ?g - goal)
:precondition (or (appeal-counterexample ?x ?y ?a ?b) (appeal-past-prom
?x ?y ?a) (appeal-si ?x ?y ?a ?g) (appeal-prev-pract ?x ?y ?z ?a ?g))
:effect (and (do ?y ?a) (cando ?x ?a)) )

```

The second subgroup includes appeals in order that own opponent believes a given belief:

– Appeal to prevailing practice (justification)

```

(:action create-prev-pract-j-appeal
;Description: it uses historic information (bel-fulf ?x ?z ?g ?a) about a
third agent ?z as justification.
:parameters (?x - iam ?y ?z - agent ?a - action ?g - goal)
:precondition (and (not(= ?x ?y)) (not(= ?x ?z)) (not(= ?z ?y)) (bel-fulf
?x ?z ?g ?a))
:effect (jappeal-prev-pract ?x ?y ?a ?g ?z) )

```

– Counterexample (justification)

```

(:action: create-counterexample-jappeal
;Description: it is similar to the previous appeal, but the historic
information is about ?y.
:parameters (?x - iam ?y ?z - agent ?a ?b - action ?g - goal)
:precondition (and (not(= ?x ?y)) (bel-fulf ?x ?y ?g ?a))
:effect (jappeal-counterexample ?x ?y ?g ?a) )

```

– Trivial appeal for *bel-impl* belief (justification)

```

(:action create-trivial-impl-jappeal
;Description: it is the simplest justification. ?x leads ?y to believe a
self belief "bel-impl".
:parameters (?x - iam ?y - agent ?a - action ?g - goal)
:precondition (and (not(= ?x ?y)) (bel-impl ?x ?a ?g))
:effect (jappeal-trivial-impl ?x ?y ?a ?g) )

```

Notice that we need to define one trivial appeal for each kind of belief defined in L . For this reason, we must also define one *accept-argument* action for each kind of belief, due to the fact that the belief is the effect of such actions.

- Appeal acceptance

```
(:action accept-j-appeal-impl
;Description: it represents the acceptance of the appeal by the opponent.
Therefore, ?y believes the belief "bel-impl".
:parameters (?x ?y ?z - agent ?a - action ?g - goal)
:precondition (and (jappeal-prev-pract ?x ?y ?a ?g ?z) (jappeal-
counterexample ?x ?y ?g ?a) (jappeal-trivial-impl ?x ?y ?a ?g))
:effect (bel-impl ?y ?a ?g) )
```

Rewards. Different actions related to the promise of future rewards can be defined. We show a general action to generate this kind of argument below:

- Simple reward

```
(:action create-reward
;Description: ?x proposes to ?y the execution of ?p in exchange for
execution of ?r.
:parameters (?x - iam ?y - agent ?p ?r - action ?g - goal)
:precondition (and (not(= ?x ?y)) (isgoal ?y ?g) (bel-impl ?y ?r ?g)
(cando ?y ?p) (cando ?x ?r))
:effect (reward ?x ?y ?p ?r) )
```

- Reward acceptance

```
(:action accept-reward
;Description: it represents the acceptance of rewards. Consequently, ?y
undertakes to execute ?p in exchange for the execution of ?r by ?x. As in
the acceptance of appeals, ?x obtains the ability to execute ?p.
:parameters (?x ?y - agent ?p ?r - action)
:precondition (reward ?x ?y ?p ?r)
:effect (and (do ?y ?p) (do ?x ?r) (cando ?x ?p)) )
```

Threats. We define the most general threats, but others may be defined varying its preconditions.

- Simple threat

```
(:action create-threat
;Description: ?x threatens ?y in the following way: if ?y does not
perform ?p, ?x will perform ?t, because ?t contradicts a goal ?g
preferred by ?y.
:parameters (?x - iam ?y - agent ?p ?t - action ?g ?h ?i ?j - goal)
:precondition (and (not(= ?x ?y)) (isgoal ?y ?g) (isgoal ?y ?h)
(cando ?x ?t) (cando ?y ?p) (prefer ?y ?g ?h) (bel-impl ?x ?t ?i)
(opposite ?g ?i) (bel-impl ?x ?p ?j) (opposite ?h ?j))
:effect (threat ?x ?y ?p ?t) )
```

- Threat acceptance

```
(:action accept-threat
;Description: this represents the acceptance of a threat. ?x will not
perform ?t if ?y performs ?p.
:parameters (?x - iam ?y - agent ?p ?t - action)
:precondition (threat ?x ?y ?p ?t)
:effect (and (do ?y ?p) (not (do ?x ?t)) (cando ?x ?p)) )
```

Note that other actions may be defined varying preconditions and effects.

5 PDDL Preferences for Argumentation Plan Generation

Rahwan et al. (2003) consider the argument selection like the essence of the strategy in argumentation-based negotiation. This mechanism consists of selecting the next

argument to be uttered from the set of candidate arguments, which might be uttered. The agent decides which argument to select on the basis of, for instance, the type of argument and the opponents that it must persuade contemplating, to this end, the argument strength (Kraus et al., 1998) or the trust in its opponents (Ramchurn et al., 2003). Thus, if there are two different candidate arguments, the agent will select one by applying some policy to determine which one should be uttered.

In this context, to generate argumentation plans, the planning algorithm must take into account the argument selection policy of the agent. In our work, we represent this policy into the agent's mental state as preferences over actions and goals. So, if the agent prefers to utter appeals instead of threats, in the agent's mental state the *create-argument* actions for appeals will have a higher preference level than the *create-argument* actions for threats. Moreover, these preferences can change in accordance with other factors, such as the opponent's trust. For example, as in the work of Ramchurn et al. (2003), when the opponent's trust is low, the agent can prefer to use a strong argument (e.g. a threat), whereas when the trust is high, to use a weak argument (e.g. an appeal).

Some proposals from the Fifth International Planning Competition IPC-5 (Baier et al., 2007; Baier et al., 2006; Edelkamp et al., 2006) were useful to address our work. These algorithms consider constraints and preferences during the planning process based on an extension of the language PDDL (Gerevini and Long, 2006). PDDL3 preferences are highly expressive. However, they are solely state centric, identifying preferred states along the plan trajectory (Sohrabi et al., 2009).

For example, it is possible to specify a preference about an effect of an action, e.g. a preference about the state in which the predicate (`do ?x ?a`) is true, but not about the way in which this effect is achieved. Consequently, it is not possible specify preference among different actions with the same effect, at least in a direct way.

An alternative to define this kind of preferences is to indicate an effect that represents the action executed. Thus, we add a predicate without parameters to the effect of each *create-argument* action, which represents the argument generated. For example, in the action `create-prev-pract-appeal`, we add a predicate `appealPP`, and add the following preference in the problem definition: `(preference p-appealpp (not (appealPP)))`. Finally, we add a plan metric to penalize the preference violation: `(* 20 (is-violated p-appealpp))`. With this metric, we indicate that if the appeal to prevailing practice is added to the argumentation plan the penalization cost is 20. Following this idea, we can emulate the argument selection mechanism described by Kraus et al. (1998), where the preference over arguments is given by the argument strength, taking into account the appeals as the weakest argument, and threats as the strongest argument. To do this, we define the following preferences: `(preference p-appealpp (not (appealPP))) (preference p-appealc (not (appealC))) (preference p-appealpstp (not (appealPstP))) (preference p-appealsi (not (appealSI))) (preference p-jappealpp (not (jappealPP))) (preference p-jappealc (not (jappealC))) (preference p-jappealtr (not (jappealTr))) (preference p-rewardp (not (rewardP))) (preference p-threatp (not (threatP)))`.

Also, we define the following metrics: `(:metric minimize (+ (* 20 (is-violated p-appealpp)) (* 20 (is-violated p-appealc)) (* 20 (is-violated p-appealpstp)) (* 20 (is-violated p-appealsi)) (* 20 (is-violated p-jappealpp)) (* 20 (is-violated p-jappealc)) (* 30 (is-violated p-jappealtr)) (* 40 (is-violated p-rewardp)) (* 60 (is-violated p-threatp)))`

Thus, argumentation plans without threat will be preferable to plans with this kind of argument.

Moreover, we can add preferences to the precondition of a *create-argument* action, in order to define the context in which the argument should be generated. For example, as mentioned above, when the opponent's trust is low, the agent can prefer to use a threat (Ramchurn et al., 2003). To represent this priority, we add `(preference p-trust (exists (?l - level) (and (conf ?y ?l) (= ?l low))))` to the precondition of the *create-threat* action. Then, if the `?y`'s trust is low the preference will not be violated and there will not be penalization.

It is worth to notice that the agent can modify the existing preferences and add more general ones (e.g. by agents), as long as it keeps obtaining useful information from its opponents.

6 Case Study

In this section, we show a case study about the utilization of PDDL planning for the generation of argumentation plans. For this, we use the planning algorithm MIPS-XXL introduced by Edelkamp et al. (2006).

Given the agents `ag1`, `ag2`, `ag3` and `ag4` on a competitive environment, `ag1` should reach an agreement with some opponent in order to perform the action `a5`, since this action is needed to fulfilling one of its goals. Agent `ag1` knows that `ag2`, `ag3` and `ag4` are agents; `a1`, `a2`, `a3`, `a4`, `a5` and `a6` are actions; and `g1`, `g2` and `g3` are goals. Additionally, `ag1` knows the following negotiation information: `(isgoal ag1 g1) (cando ag1 a1) (cando ag1 a6) (bel-impl ag1 a5 g1) (bel-fulf ag1 ag4 g2 a4) (bel-impl ag1 a4 g2) (isgoal ag2 g2) (cando ag2 a2) (cando ag2 a5) (isgoal ag3 g3) (cando ag3 a3) (cando ag3 a4) (bel-impl ag3 a1 g3)`; as well as its preferences about the arguments (as we described in Section 5). Thus, this information composes the initial state of the planning problem, and the final state is composed of the predicate `(exists (?x - agent) (do ?x alpha))`, which represents the agreement that `ag1` expects to reach, due to the need of performing `a5` in order to fulfil `g1`. Finally, the actions of the planner have been defined in Section 4.3.

The resultant argumentation plan is the following: `0: (create-prev-pract-j-appeal ag1 ag2 ag4 a4 g2). 1: (create-reward ag1 ag3 a4 a1 g3). 2: (accept-j-appeal-impl ag1 ag2 ag4 a4 g2). 3: (accept-reward ag1 ag3 a4 a1). 4: (create-reward ag1 ag2 a5 a4 g2). 5: (accept-reward ag1 ag2 a5 a4).`

Observing the plan, we can see the actions that represent the arguments, which `ag1` should utter during the argumentation process. Firstly, `ag1` should use an appeal to prevailing practice in order to lead `ag2` to believe that the execution of `a4` implies to reach the goal `g2`. And secondly, `ag1` should persuade `ag3`, using a reward, to carry out the execution of `a4` in order to offer it to `ag2` in exchange of `a5`.

The preferences of PDDL provide a useful versatility to planning. For example, an alternative plan could be built to solve the previous problem. In this alternative plan, the appeal to prevailing practice is replaced by the trivial appeal, since ag_1 also knows that if action a_4 is performed, goal g_2 is fulfilled ($bel\text{-}impl\ ag_1\ a_4\ g_2$). However, this plan is not preferable since the preferences indicate that the penalization of using a trivial appeal is bigger than the penalization of using an appeal to prevailing practice. Consequently, if ag_1 decrease the penalization of the metric $p\text{-}jappeal_{tr}$ to 10, the planner will prefer the trivial appeal instead of the previous one.

7 Discussion: Pros and Cons of Using PDDL for Argumentation-Based Negotiation Planning

After analysing the PDDL language to model argumentation-based negotiation planning, we can distinguish some pros and cons of this approach. One of the advantages of using PDDL is that is a well-known standard, which has been tested in complex environments. Consequently, we count with a wide range of planners that can be used by the agent that need to plan its argumentation. However, few planners implement the PDDL3.0 features (plan constraints and preferences) completely. One of these planners is MIPS-XXL, which was applied in the case study.

Despite the good expressivity of PDDL, some predicates related to the agent's mental state and argument locutions cannot be declared, at least without loss clarity. This is because PDDL does support neither nested predicates nor different structure definitions for the same predicate. For this reason, we need to define different belief structures ($bel\text{-}impl$ and $bel\text{-}fulf$), and consequently, different *create-argument* actions, such is the case of the trivial appeals. Similarly, we have to define a different predicate for each type of appeal and use a disjunctive precondition to join them in a same *accept-argument*.

In regards to the preferences definition, it is possible specify a preference about an effect of an action, but not about the way in which this effect can be achieved. Consequently, it is not possible specify preference among different actions with the same effect, at least in a direct way. An option to define this kind of preferences is indicate an effect that represents the action execution. For these reason, we have to add a predicate without parameters to the effect of each *create-argument* action, to represent the generated argument.

In summary, PDDL gives all the alternatives to generate argumentation plans for argumentation-based negotiation. However, certain peculiarities of the language lead to the necessity of defining additional predicates and actions that add confusion to the problem definition.

8 Conclusions

We have analysed how to define a problem planning for argumentation plan generation by using the PDDL language. We explained how the argumentation

process, during the negotiation, may be modelled as a planning problem using PDDL in order to obtain an *argumentation plan*.

To illustrate this analysis, we have presented a case study that showed how an argumentation problem can be defined using PDDL and how an argumentation plan can be generated by using a PDDL planner.

Finally, taking into account the pros and cons discussed above, we conclude that the standardization of the argumentation-based negotiation planning using PDDL allows us to be easily assimilated by the community and adapted to different domains.

Acknowledgements. This work has been partially funded by ANPCYT through PICT project N° 1231/10.

References

- Amgoud, L., Dimopolous, Y., Moraitis, P.: A Unified and General Framework for Argumentation-based Negotiation. In: Proc. 6th Int. Joint Conf. on Autonomous Agents and Multi-Agents Sys-tems, AAMAS 2007, Honolulu, Hawaii, pp. 1–8 (2007)
- Ashri, R., Rahwan, I., Luck, M.: Architectures for Negotiating Agents. In: Mařík, V., Müller, J.P., Pěchouček, M. (eds.) CEEMAS 2003. LNCS (LNAI), vol. 2691, p. 136. Springer, Heidelberg (2003)
- Baier, J., Bacchus, F., McIlraith, S.: A Heuristic Search Approach to Planning With Temporally Ex-tended Preferences. In: 20th Int. Joint Conf. on Artificial Intelligence, pp. 1808–1815 (2007)
- Baier, J., Hussell, J., Bacchus, F., McIlraith, S.: Planning with Temporally Extended Preferences by Heuristic Search. In: 5th Int. Planning Competition Booklet, pp. 20–22 (2006)
- Edelkamp, S., Hoffmann, J.: PDDL2.2: The Language for the Classic Part of the 4th International Planning Competition, Tech. Rep. 195, Institut für Informatik, Freiburg, Germany (2004)
- Edelkamp, S., Jabbar, S., Naizih, M.: Large-scale Optimal PDDL3 Planning with MIPS-XXL. In: 5th Int. Planning Competition Booklet, pp. 28–30 (2006)
- Fox, M., Long, D.: PDDL2.1: An Extension to PDDL for Expressing Temporal Planning Domains. *Journal of Artificial Intelligence Research* 20, 61–124 (2003)
- Geipel, M.M., Weiss, G.: A Generic Framework for Argumentation-Based Negotiation. In: Klusch, M., Hindriks, K.V., Papazoglou, M.P., Sterling, L. (eds.) CIA 2007. LNCS (LNAI), vol. 4676, pp. 209–223. Springer, Heidelberg (2007)
- Ghallab M., Howe A., Knoblock C., McDermott D., Ram A., Veloso M., Weld D., Wilkins D.: PDDL – The Planning Domain Definition Language. Tech. Rep. CVC TR98-003/DCS TR-1165, Yale Center for Computational Vision and Control (1998)
- Gerevini, A., Long, D.: Plan Constraints and Preferences in PDDL3. Tech. Rep. 2005-08-07, Department of Electronics for Automation, University of Brescia, Italy (2005)
- Gerevini, A., Long, D.: Preferences and Soft Constraints in PDDL3. In: ICAPS 2006 Workshop on Preferences and Soft Constraints in Planning, pp. 46–53 (2006)
- Gerevini, A., Haslum, P., Long, D., Saetti, A., Dimopoulos, Y.: Deterministic Planning in the Fifth International Planning Competition: PDDL3 and Experimental Evaluation of the Planners. *Artificial Intelligence* 173(5-6), 619–668 (2009)

- Hoffmann, J., Edelkamp, S.: The Deterministic Part of IPC-4: An overview. *Journal of Artificial Intelligence Research* 24, 519–579 (2005)
- Kraus, S., Sycara, K., Evenchik, A.: Reaching Agreements through Argumentation: a Logical Model and Implementation. *Artificial Intelligence* 104(1-2), 1–69 (1998)
- Monteserin, A., Amandi, A.: Argumentation-based Negotiation Planning for Autonomous Agents. *Decis. Support Syst.* 51(3), 532–548 (2011)
- Rahwan, I., Ramchurn, S., Jennings, N., McBurney, P., Parsons, S., Sonenberg, L.: Argumentation-based Negotiation. *The Knowledge Engineering Review* 18(4), 343–375 (2003)
- Ramchurn, S., Jennings, N., Sierra, C.: Persuasive Negotiation for Autonomous Agents: A Rhetorical Approach. In: Reed, C., Grasso, F., Carenini, G. (eds.) *IJCAI Workshop on Computational Models of Natural Argument*, pp. 9–17 (2003)
- Sohrabi, S., Baier, J.A., McIlraith, S.A.: HTN Planning with Preferences. In: *21st Int. Joint Conf. on Artificial Intelligence*, pp. 1790–1797 (2009)
- Weld, D.: An Introduction to Least Commitment Planning. *AI Magazine* 15, 27–61 (1994)

Predicting Potential Responders in Twitter: A Query Routing Algorithm

Cleyton Caetano de Souza¹, Jonathas José de Magalhães¹,
Evandro Barros de Costa², and Joseana Macêdo Fechine¹

¹ Department of Systems and Computing
Federal University of Campina Grande
Campina Grande-PB, Brazil

² Computing Institute
Federal University of Alagoas
Maceió-AL, Brazil

Abstract. A phenomenon not so recent is the substantial increase in popularity and use of online social networks. With that has emerged a new way to find information online: the social query, which consists of posting a question in a social network and wait for responses from close friends. Usually, a question is posted to be visible to everyone, but we believe that this is not the best way: there will be the possibility of receiving several responses (including wrong), keep receiving answers where there is no need, do not receive answers, etc. The query router problem consists of finding the most able individual in the personal social network of the questioner. This work presents an algorithm to Routing Questions in Twitter. The model was validated through its predict capacity and the results shows that its recommendations match in half cases only when combined with a technique to enrich the information present in the question.

Keywords: Social Query, Routing Algorithm, Social Network, Twitter.

1 Introduction

A phenomenon not so recent, but perceived with greater intensity in last two years, is the substantial increase in popularity and use of online social networks. The Facebook¹, for instance, is the most visited website actually, surpassing even Google² [23]. On these virtual environment, people with common characteristics meet each other and discuss topics of mutual interest. The social networks was designed initially only to allow interaction between people. They have evolved and today are used for different purposes, e.g., for the distribution of games or the announcement of products between users.

Another interest phenomenon is a new way to find information online which born in Community Question & Answer Sites (Q&A Sites) and was extended to

¹ <http://facebook.com>

² <http://google.com.br>

usual social networks: the social query, which consists of posting a question in a social network and wait for responses from close friends [15].

In social networks like Twitter³, for example, the social query, appears as a valid alternative, in some cases, to find answers online [19]. Usually, a question is posted to be visible to everyone. However, we believe that this is not the best way. After posting a question that will be visible to everyone there are some possible scenarios, e.g., receiving several responses, including wrong or continue to receive responses when no longer needed. Moreover, there is the possibility of potential responders may never see the question, thus they will never answer it.

One solution to this problem is the routing of questions, which consists of identifying what is the user connected to the questioner is more able to provide the correct answer and direct the question just for him (query router) [3]. Such recommendation is characterized by the intimacy implicit between the questioner and the holder of knowledge; because they are direct connect in the social network [19].

However, decide to whom to direct the question is not a trivial task. When you choose the wrong individual can spend a long time to get a response, you can get one wrong answer or the chosen one can simply ignore the question. Thus, the technical problem of this research consists in identify the individual from the personal social network of the questioner who is better able to respond correctly and timely the question and direct it only for him/her. In this respect, this work proposes an algorithm for Routing Questions, which can be understood as a technique of recommendation that identifies the better able individual using various criteria (knowledge, social aspects, availability, etc.). We use the Weight Product Model (WPM), a strategy for making decisions with multiple criteria, to qualify the aptitude of all candidates [14].

The Routing algorithm proposed was designed to work in the Twitter, but it can be easily adapted to any context of other social networks. Previous work applies Routing algorithm in small social network developed only for Q&A and to offer support to the algorithms. The differential of our work is (1) we start proposing a model in a pre-existent and very popular social network and (2) we lead with the problem as a problem of decision making with multiple criteria, instead as a probabilistic problem like some most part of previous work.

Another pertinent question is how evaluating a Query Routing algorithm? There is no an evaluation technique common to the entire academic community and the majority of researches that deals with personalized recommendation present a qualitative evaluation, which difficult the comparison among the different algorithms [9]. However, assuming that the initial goal of a recommendation technique is to predict something that will interest someone [12], we decide to evaluate the proposal on the perspective of their predictive ability. Does the proposed Routing algorithm make recommendations that will reflect the events of the real world? Does the proposed Routing algorithm can predict who will answer questions posted openly on Twitter? With this purpose was validated the following hypotheses:

³ <http://twitter.com>

- $H_{0,1}$: The proposed Routing algorithm cannot predict the events of the real world at least 50% of trials;
- $H_{a,1}$: The proposed Routing algorithm can predict the events of the real world at least 50% of trials.

To evaluate the validity of these hypotheses, we designed an experiment where some posted questions on Twitter were monitored and recorded who answered them. Then, the necessary information was passed to the algorithm and a list of recommendations with the most able was requested. The list of recommendation was compared with the users who really answer the question. Next, we calculate the hit rate (recall) obtained, aiming to study the validity of the cited research hypotheses. Importantly, the objective of the proposed algorithm is not predict who will respond, but indicate who respondents are more likely. We have assumed that who answer the question automatically must be considered one of the fittest and therefore possibly the Routing algorithm should indicate him/her as a recommendation.

In Twitter, the users can post message with until 140 characters, i.e., the questions that users have posted are subject to this limitation. Manipulate short questions is a hard task, because there is a need for terms that characterize well the topic of the question [6]. For this reason, the same experiment was performed using a synonymy expansion in question before it be passed to the algorithm. Thus, it is expected to obtain a recall rate equal to or greater than the recall rate of the first method, since the expansion terms have the ability to prioritize the most relevant results, according with Ramalho and Robin [17].

To evaluate the effectiveness of the proposed Routing algorithm combined with the synonymy expansion, we consider the following hypotheses:

- $H_{0,2}$: The proposed Routing algorithm combined with the synonymy expansion in question cannot predict the events of the real world at least 50% of trials;
- $H_{a,2}$: The proposed Routing algorithm combined with the synonymy expansion in question can predict the events of the real world at least 50% of trials.

Finally, to evaluate if the combination of Routing algorithm proposed with synonymy expansion produces a recall rate higher than the simple Routing algorithm, we consider the following hypotheses:

- $H_{0,3}$: The combination of the Routing algorithm with the synonymy expansion do not produces a recall rate higher than the same technique without expansion;
- $H_{a,3}$: The combination of the Routing algorithm with the synonymy expansion produces a recall rate higher than the same technique without expansion.

The study was conducted with nine persons who publish twenty nine questions and involved the processing of a graph composed for 1201 users, 131.962 messages

and 2.047.305 links between users. The analysis over the recall rate indicated that the Routing algorithm combined with the synonymy expansion reached the level expected (50%), but the simple technique did not reach, i.e., the hypotheses $H_{0,1}$ and $H_{a,2}$ and was accepted. Furthermore, the recall rate of both techniques were compared and the obtained conclusion is that the technique with synonymy expansion present results statically better than the simple technique (without expansion), confirming the hypothesis $H_{a,3}$. In our opinion, these results make clear the need for methods that improve the analysis of the content of the question.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work concerning previous studies that deal with social networks focused on finding resources on the Web. Section 3 presents the social network Twitter and discusses about the reasons that led us to choose it as context of this research. Section 4 describes the model, the characteristics that should be considered when we discuss about the capability of a candidate and how we measure these. Section 5 presents the Weight Product Model (WPM), as a model adopted in our approach for finding the best candidates. Section 6 describes the experiment and results used for validating our proposal; and, finally, Section 7 offers conclusions and discusses some future work.

2 Related Work

We decided to split this section into four parts. The first Subsection (2.1) presents researches that deal with the act of publishing questions on the web. The second Subsection (2.2) presents researches about Expertise Finding using multiple criteria. The third Subsection (2.3) presents studies that specifically address the query router problem either in conventional social networks (like Twitter or Facebook) or CQA Sites. Finally, the Subsection (2.4) presents the main differences between the previous work and our proposal.

2.1 About the Act of Publish Question in the Web

The fact is that search engines are not always the best way to find information on Web. To some needs of information are better solved by people, for instance, personal questions, recommendations, opinions, advices and high contextualized questions [6]. An alternative to this kind of questions are the CQA Sites as AnswerBag⁴ and the Yahoo! Answers⁵, which consists in virtual communities where users post and answer questions voluntarily. After posting a question, the user waits for answers of others users, who usually are unknown to him. But, people prefer ask questions to their close friends in social networks than to unknown persons in CQA Sites [18].

Regarding to the explicit action of publishing a question on social networks, Morris, Teevan and Panovich [15] present important statistics that confirm this

⁴ <http://www.answerbag.com>

⁵ <http://answers.yahoo.com>

as a viable strategy to get answers online. In their study case, 93.5% of users received answer to their questions after post them and these responses, in 90.1% of cases, were provided within one day. Paul, Hong and Chi [16] conduct a similar study using only Twitter, they conclude that, in this specific context, only a few part of questions posted receive answers (18.7%) and that the fact of receive or do not is intrinsically connected to the amount of followers of the questioner. However, questions posted in Twitter normally be answered quickly, in their study 67% of the responses come within the range of 30 minutes and 95% within the range of ten hours. We believe these findings are result mainly of the features of Twitter: when a user post questions to all followers, only a portion of his/her followers will view and a smaller portion will respond (as will be detailed in Section 3). Thus, users with more followers are more likely to get answers, because there is a larger viewing of their messages. And, with respect to agility in respond receiving, this is mainly due to the nature of Twitter as a real-time social network.

However, we believe that these results could be improved applying the routing questions: identifying an expert on the subject of the question, the answer could come faster and with higher quality. Horowitz and Kamvar [6] established a correlation between social query and the village paradigm: when an individual in a village looking for information, before consult the libraries, he turns first to the most intelligent people he knows.

2.2 About Expertise Finding

In fact, the problem that this paper aims to address can be understood as an Expertise Finding Problem. However, usually the Expertise Finding involves a context much large of candidates. The work proposed in this paper deals with the detection of specialists in much smaller subset of the entire social network (the set of friends/followers of the questioner) and thus the conditions under which a friend is marked as a specialist in each case differ because of the context that is analyzed ,i.e., the Expertise Finding addressed here is personalized. Moreover, usually Expertise Finding involves only the discovery of who owns knowledge about a given topic, while (as will be presented in Section 4) the algorithm proposed here involves multiple criteria.

Sarda et al. [18] deals with the identification of experts in Orkut⁶ using two criteria: knowledge and confidence. However, this work has focused on the mining of expert opinion about products and not necessarily on the resolution of questions. Smirnova and Balog [20] propose a Recommendation Technique to identify experts using two criteria: knowledge gain (calculate using Bayes Theorem) and contact time (the calculation will vary with the type of social network, but is distance, though some metric, between the questioner and the expert). However, we believe that the model for contact time (which would be the time needed to receive answers) proposed in this work is not consistent with reality, because it not measures the availability of the other user to provide the answer.

⁶ <http://orkut.com>

2.3 About Routing Questions

The query router consists of an algorithm of recommendation (or a technique) that objectives find an expert present in a group and direct a question to him/her [21]. In [2], Banerjee and Basu presented a probabilistic and decentralized model for the routing questions problem. This means that there is not an entity that makes all decisions and the Routing algorithm works based on the probability of actions taken in past be repeated. Other work is the [5] that presents a centralized model: the iLink is a global entity that decides who will receive the questions and, in some cases, is also able to offer answers.

Some examples of systems that implement a model of query router are the Aardvark [6], a social network that belongs to Google, Q-Sabe [1], an academic tool for exchange of information focused on education. Both systems consist of CQA Sites where users publish questions (questioners) that are directed to other users (respondents) and these can choose to answer or ignore the question, and AskWho [13], a plugin to Facebook which aims present friends of the questioner as possible answers to a question.

2.4 About the Differential of Our Research

The researches of Andrade et al. [1], Davitz et al. [5] and Horowitz and Kamvar [6] propose query router techniques and developed environments where they works. Our research follows the reverse path. We propose a Query Router algorithm that works in a pre-existent and popular social network: the Twitter, one of most popular social networks currently and which, apparently, will benefit of our technique. In [13] is presented a plugin to Facebook, but AskWho do not use any special technique to match friends, consisting only in a search engine which compare the content of the question with the friends. The differential our research in relation to the works cited is that we take as our starting point a specific and popular social network and we build our model to fit in the context of this network. Moreover, our model inherits characteristics of other related work, like trust model and friendship. We presented these concepts of a way that they can be easily adaptable in other contexts. We believe that the query router problem can be treated as a multi-criteria decision making problem, instead a probabilistic problem considered by previous work. For this reason, another differential of our research is the solution of the model using WPM, a strategy for making decisions with multiple criteria, considered adequate for the amount of variables involved [22] and that also allows a dynamic evolution of our technique trough the addition of new criteria.

In [21], we present our Formal Model do the Query Router Problem to the context of Twitter. The work that will be presented below shows our model in an algorithm form, easily compressive and which exposes new features and adaptations of the technique.

3 Twitter

Twitter is a kind of microblog, a variant of the blogs that have some type of limitation of the content, where users can tweet (post a message) on any topic using 140 characters [4]. In less than three years, Twitter gained such popularity that became the microblog with the largest number of users [4] and awakening the interest of the scientific community about it [8,11]. In January 2010 the microblog counted more than 73 million users and in March, month in which it completed four years, reached the mark of 10 billion posted messages [4].

Via Twitter users can follow other users and can be followed by other users. In the context of Twitter, to follow a user means exposing publicly interest in the content posted by him. Among the reasons that lead a user to follow another one are admiration, friendship and reciprocity, for example. In addition, the user may want to follow other user by considering that content posted by him is relevant. The tweets (posts) may or may not be publics and any user is allowed to refer to others within a tweet. Because of these features many users use the microblog as a public chat [7].

When a user publish a public question in the Twitter, if it will not be answered quickly the chances of be visualized and answered in future are lowest because the question will down in timeline of followers of him. After publish a question, probably, not all users who follow the questioner will see the question. Among the users that visualize, only a few will provide an answer and there is no guarantee that any of these answers will satisfy the information needs of the questioner. Some users will not provide an answer because, as the question was posted for all followers, they do not feel an obligation to help. And as time passes, the chances of that question be viewed and consequently answered in the future are lower, because it will fall positions on the timeline of the followers of the questioner.

We believe that when a tweet (question) is directed previously to someone the probability of it be visualized are much larger, because the user mentioned can filtrate the messages which mention him/her. When a user mention other user, the user mentioned, immediately, receive an email inform about the message. There is the possibility of the mentioned user disable these notifications, but any user can filter their mentions, as already commented. Given these facts, we believe that direct the question to someone, practically, guarantees that the message will be visualized, but there is no guarantee that the message will be answered, either on the quality of the response.

Appears evident that to direct a question increases the probability of it be visualized, while the probability of it be good answered depends to who it will be directed. A Formal Model to Query Router in Social Networks consists in a recommendation algorithm (technique) that analyses the information available on social network to infer who the user most able to respond the question. In Figure 1 is illustrated the query router process working.

The question is formulated without a mention. The Query Router algorithm (or Routing algorithm) analyzes the information about the followers of the questioners and ranks them according to aptitude to answer the question. The algorithm adds a mention in the question. In Figure 1 the question is being directed

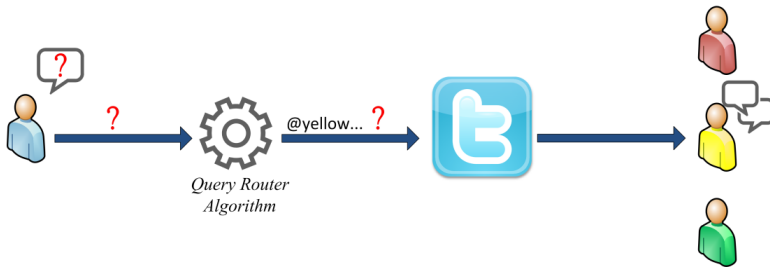


Fig. 1. Using the Query Router algorithm to Tweet a Question

to only one follower (the yellow user), but, as the algorithm ranks all user, will be possible send it to the n followers in bests positions.

In the next section, will be presented the Formal Model to Query Router, which is the proposal of this work.

4 Formal Model

Informally, the problem that the model proposes to solve is given a question posted by a user (questioner), find among his followers that user with the characteristics:

1. Knows the answer: if we direct the question to a follower who has no knowledge on the subject of some question, the quality of the response will be low and probably will not satisfy the information needs of the questioner;
2. Has the trust of the questioner: if we direct the question to a follower who has no confidence of the questioner, regardless of the answer, probably, the questioner could not believe in the responder and prefer continue to receive answers from other followers;
3. Provides the answer quickly: if we direct the question to a follower who does not access the social network often, a long time can spend until the questioner get your answer;

Thus, we need find a way to decide who user has the best combination of these three characteristics. Next, we will detail our Formal Model to Routing Questions in the Context of the Social Network Twitter.

4.1 The Social Network

The Twitter is a social network defined by the tuple $T = (U, R)$, where $U = \{u_1, u_2, u_3, \dots, u_{|U|}\}$ is a set of social network users and R corresponds to the set of relationships as $r_{i,j}$ between two users i and j , where $i \in U$ and $j \in U$. In the specific context of Twitter the relationships are not necessarily reciprocal, so $r_{i,j} \neq r_{j,i}$. The existence of relationship $r_{i,j}$ means that the user i is a follower of the user j .

An user u has these sets as attributes: $Followers_u$, $Following_u$ and M_u . The set $Followers_u \subset U - u$ contains all user x with whom u maintains a relationship of form $r_{x,u}$. The set $Following_u \subset U - u$ contains all user x with whom u maintains a relationship of form $r_{u,x}$. The set $M_u = \{m_{u,1}, m_{u,2}, m_{u,3}, \dots, m_{u,|M_u|}\}$ consists of all messages (tweets) posted by the user u . Each message m_u posted by u has the following attributes: d_{m_u} (corresponds the date that the message m_u was posted by the user u) and s_{m_u} (corresponds the string that represents the message content).

4.2 Problem Definition

The problem is to find a user $f_u \in Followers_u$ that has a higher probability of answering a question q_u that requires the knowledge $k_{f_u,q}$ and was published by user u in form of a message m_u . The calculation of this probability, called p_{q_u,f_u} , takes into account three abstract concepts:

1. $k_{f_u,q}$: knowledge of f_u in relation to the question q_u ;
2. t_{u,f_u} : trust of u on the user f_u ;
3. a_{f_u} : activity of f_u in social network.

So the problem can be summarized in find a user $f_u \in Followers_u$ whose tuple $(k_{f_u,q}, t_{u,f_u}, a_{f_u})$ maximizes the probability p_{q_u,f_u} , i.e., the user that has the best tuple $(k_{f_u,q}, t_{u,f_u}, a_{f_u})$.

The details of how we believe that these concepts are quantified are described with in [21]. Bellow, we present our Model in an algorithm form, to make easier understand it.

5 The Routing Algorithm

Figure 2 illustrates the pseudo code of the Routing algorithm used to qualify candidates.

First, it is verified the need of make or not make the synonymy expansion (line 1). The variable *DEFAULT_AMOUNT* (line 2) limits the max amount of synonyms that should be added to each word.

After synonymy expansion part, we calculate the attributes of each follower of the questioner: knowledge (line 6), trust (line 7) and activity (line 8). We define knowledge as an expertise that some user has about a topic. In social networks the knowledge of a user u is direct related with the content posted by him: M_u [5,6]. Trust is a measure that quantifies the faith that a user has over the information posted by another user (credibility) [19] and in our model is calculated based on friendship and similarity. The activity level corresponds to the frequency with the user post new tweets [21]. Calculate the tuple $(k_{f_u,q}, t_{u,f_u}, a_{f_u})$ for any user $f_u \in Followers_u$ is a simple task, but compare the tuples of two different users and decide which user is the best qualified to answer the question is not always a trivial task.

```

1  if (expansion == TRUE)
2    q = synonyms(q, DEFAULT_AMOUNT)
3
4  for each  $f_u$  in Followers $_u$  do
5    Begin
6       $k_{f_u,q}$  = knowledge( $f_u, q$ )
7       $t_{u,f_u}$  = similarity( $f_u, u$ ) * friendship( $f_u, q$ )
8       $a_{f_u}$  = activity( $f_u$ )
9       $f_u.victories$  = 0
10   End
11
12  /*Weight Product Model*/
13  for each  $f_{1_u}$  in Followers $_u$  do
14    for each  $f_{2_u}$  in Followers $_u$  do
15      Begin
16        if ( $f_{1_u} == f_{2_u}$ )
17          continue;
18
19         $k = k_{f_{1_u},q} \div k_{f_{2_u},q}$ 
20         $t = t_{u,f_{1_u}} \div t_{u,f_{2_u}}$ 
21         $a = a_{f_1} \div a_{f_2}$ 
22
23        comparison = ( $k^{RELEVANCE\_KNOWLEDGE\_LEVEL}$ ) *
24                    ( $t^{RELEVANCE\_TRUST\_LEVEL}$ ) *
25                    ( $a^{RELEVANCE\_ACTIVITY\_LEVEL}$ )
26
27        if (comparison ~ = 1)
28          Begin
29             $f_{1_u}.victories$  ++
30             $f_{2_u}.victories$  ++
31          End
32        else if (comparison ~ > 1)
33           $f_{1_u}.victories$  ++
34        else if (comparison ~ < 1)
35           $f_{2_u}.victories$  ++
36        End
37      sort_according_victories(Followers $_u$ )
38
39  return Followers $_u$ 

```

Fig. 2. Routing Algorithm

This way, as we already have commented, we consider that as a problem of decision making with multiple criteria (or multi-criteria decision making) and trying to find the best tuple among users $f_u \in Followers_u$, after calculate the attributes of each follower, we used the Weight Product Model (WPM) as a method for making decisions because it is the most appropriate for the conditions and context presented (dependence up to three variables) [22] (start in line 13).

The values *RELEVANCE_KNOWLEDGE_LEVEL* (line 23), *RELEVANCE_TRUST_LEVEL* (line 24) and *RELEVANCE_ACTIVITY_LEVEL* (line 25) are called factors of importance, must be set according to user need and their sum must results in 1, i.e., $(A+B+C+\dots = 1)$ [14]. When the user wants to prioritize the speed of response (to get answers quickly) he must establish a high value for the factor *RELEVANCE_ACTIVITY_LEVEL*, in case the user wishes to prioritize the answers from friends (because it requires a very personal response) he must establish a high value for the factor *RELEVANCE_TRUST_LEVEL*

and, finally, when the user wants to prioritize the knowledge that his friends have about the domain of the question (to find good answers) it must establish a high value for the factor *RELEVANCE_KNOWLEDGE_LEVEL*.

Still about WPM we use a variable called *comparison* to compare users in pairs. If *comparison* > 1 , then f_1 is superior to f_2 and we put 1 in position (f_1, f_2) of the matrix and 0 in position (f_2, f_1) (line 32). If *comparison* < 1 , then f_2 is superior to f_1 and we put 1 in position (f_2, f_1) of the matrix and 0 in position (f_1, f_2) (line 34). If *comparison* = 1, then f_1 is equivalent to f_2 and we put 1 in position (f_1, f_2) of the matrix and 1 in position (f_2, f_1) (lines 28 and 29). But in our approach we use a little confidence interval to decide. For this reason, we use in pseudo code the symbols to express: $\sim =$ (near), $\sim >$ (a little larger) and $\sim <$ (a little smaller).

We summarize the amount of victories of each user (start in line 26) and after compare all user we order them according with the number of victories (line 37). At the end, this ordered list also represents the relevance order of each user and it is returned as a solution of the algorithm (line 39).

We decided to address the problem as a multi-criteria decision making because it makes the algorithm easily expandable. The addition of new criteria, for example reciprocity [10] or the latency time (time between the last message and the current instant), requires only: (1) the addition of its calculation to each follower, (2) its usage in the ratios that calculates the value of the *comparison* (start in line 23) and (3) the calibration of the factors of relevance of each criteria.

6 Evaluation and Results

This section describes the details of evaluation process and then discusses how validation of the Routing algorithm proposed in this paper was performed. Additionally, the obtained results were reported.

6.1 Methodological Aspects

To validate the Routing algorithm proposed was draw an experiment whose objective was to ascertain its ability to reflect, trough recommendations, what happened in real world. For the research, nine people posted on Twitter 29 questions which were answered by 44 users. These questions are designed by the participants themselves, were visible to all followers and anyone that wanted to could answer it. Each question was considered a trial, being the input variables: the question (q), the user information (u) and the list of users who responded on Twitter (*real responders*). For each question was requested to Routing algorithm a list of recommendation (*recommended responders*). The list of recommended users was then compared with the list of users who answered the question in the real world and we analyze the recall rate obtained. As already commented, in order to analyze the benefits that the synonymy expansion would bring in the Routing algorithm, the same experiment was performed only passing in different an extended version of the question.

The study was conducted with nine persons who publish 29 questions and involved the processing of a graph composed for 1201 users, 131.962 messages and 2.047.305 links between users. All information used in research may be available by contacting any of the authors. To conduct the study, the proposed Routing algorithm was implemented in Java, for extracting data from the social network we used Twitter Streaming API⁷, for application of Natural Language Processing (NLP) techniques was used BrazilianAnalyzer⁸, that belongs to the Lucene API, the thesaurus used for the synonym expansion was from the website “Thesaurus da Língua de Portuguesa do Brasil”⁹.

6.2 Results and Discussion

In Figure 3 is showed the amount of true positive obtained by the Routing algorithm proposed in that the size of list of recommendations grows for the two versions compared (with and without synonym expansion). The horizontal axis represents the size of the list of recommendations and the vertical axis the total number of true positive within the 44 possible.

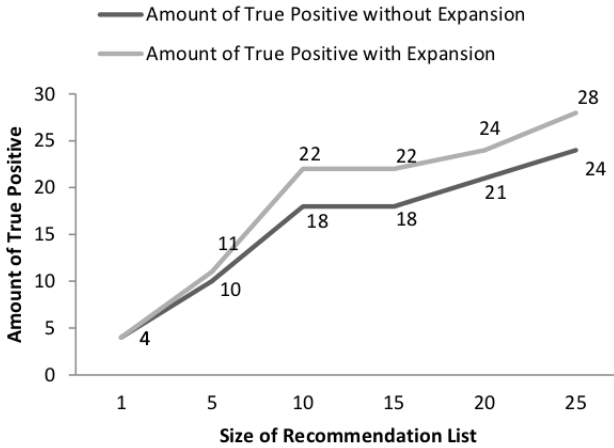


Fig. 3. Amount of True Positive of the Routing Algorithm.

Looking at Figure 3, apparently the combination with the expansion technique achieved better performance than a simple Routing algorithm (without synonymy expansion). Moreover, in both situations, even with a list of twenty five recommendations for each question, we do not match all 44 users. Even so,

⁷ <https://dev.twitter.com/docs/streaming-api>

⁸ http://lucene.apache.org/core/old_versioned_docs/versions/3_0_0/api/contrib-analyzers/org/apache/lucene/analysis/br/BrazilianAnalyzer.html

⁹ <http://alcor.concordia.ca/~vjorge/Thesaurus/>

the true positive score using 10 recommendations (P@10) was considered positive, being respectively 40% (18/44) and 50% (22/44) to a simple Routing algorithm and for the combination with the synonymy expansion. For this reason we choose this size to continue the investigation.

Using ten recommendations (P@10), the recall rate of the Routing algorithm without synonymy expansion for each one of the twenty nine trial was, in order: 0%; 30%; 100%; 100%; 0%; 0%; 0%; 100%; 100%; 50%; 0%; 100%; 0%; 0%; 100%; 0%; 0%; 50%; 0%; 100%; 0%; 100%; 50%; 0%; 0%; 100%; 50%; 100%; 100%. We can see many of the recall values are most extreme; this happen because most of the questions posted (59%) was answered by only one person. In this situation, is only possible to obtain a recall 0 (0%) or 1 (100%). To evaluate if the Routing algorithm hit at least 50% of cases, each trial where happen at least one true positive ($recall > 0$) was considered as one and the cases where no match ($recall = 0$) will be considered zero. Thus, the percentage of success obtained was 55% (16/29). To analyze whether this value is statistically significant was performed a one-tailed binomial test where he obtained a p -value of 0.3555 for $\alpha = 0.05$. This means that there is no significance to accept $H_{a,1}$. Thus, $H_{0,1}$ was the first hypothesis accepted for this work, i.e., the Routing algorithm proposed, statistically, did not get successes by more than half of the trials.

Using ten recommendations (P@10), the recall rate of the Routing algorithm with synonymy expansion for each one of the twenty nine was, in order: 50%; 30%; 100%; 100%; 100%; 100%; 100%; 100%; 100%; 50%; 0%; 100%; 0%; 0%; 100%; 100%; 0%; 50%; 0%; 100%; 0%; 100%; 50%; 0%; 0%; 100%; 0%; 100%; 100%. Thus, using the same adaptation before, the percentage of success obtained by the Routing algorithm combined with the synonymy expansion was 69% (20/29). To analyze whether this value is statistically significant was performed a one-tailed binomial test where was obtained a p -value of 0.03071 for $\alpha = 0.05$. This means that the hypothesis $H_{a,2}$ was accepted, i.e., the Routing algorithm combined with the synonymy expansion obtained successes in more than half of the trials.

Finally, the successes rates of each trial were compared to analyze if the technique with synonymy expansion is superior. Initially, a study was conducted aiming to verify the normality of the recall rates of both distributions using the Shapiro-Wilk Test which resulted in a p -value 6.582e-06 for combination technique and 4.73e-06 for the technique without the expansion. This means that both distributions are not normal. Then, using the Wilcoxon Signed Rank Test, we check if the distribution of recall rates by the Routing algorithm combined with synonymy expansion is superior to technique without expansion. The result was a p -value of 0.03299 to a $\alpha=0.05$, what means that the hypothesis $H_{a,3}$ also was accepted, i.e., the Routing algorithm proposed combined with a synonymy expansion obtained recall rates better than the same Routing algorithm without expansion.

The results showed that the combination of the Routing algorithm with the synonymy expansion got successes rates in more than half of the trials; however it is considered that such results are no so expressive. To examine whether the

recall rate for each trial (P@10) was superior to 50% was carried out again the Wilcoxon Signed Rank Test. This time, the *p-value* was 0.5981 for the recall distribution without synonymy expansion and 0.4007 for the Routing algorithm with synonymy expansion. These *p-values* indicate that the average value of successes per trial is not 50%. This negative result, in particular, is due to the high variance in success rates (many extreme values) and the statistical test did not guarantee a reliable range for the average of successes rate per trial. This raises the following question: *why have so many successes rates equal to 0%*? During the study, it was noted that the proposed task was naturally difficult. When posting a question that is visible to all, the questioner is held hostage of various random factors such as, for example, anyone who see the question might want (or not) answer it, if the question is not viewed by anyone quickly chances of someone to answer it in the future are smaller, people we never expect an answer can answer, or, e.g., if the perfect individual according to the Routing algorithm never see the question because of other random reasons (e.g., tired, blackout, did not pay the Internet bill and the service was suspended) he/she never answer it. However, if the study results was considered very positive, because being the main propose of this work to infer who are the best candidates to answer a particular question, the fact that the recommendation match with what happens in the real world consists of a predictive validity of the conceptual model, but little refers to the quality of the recommendation. Thus, we credit the negative trials (where the successes rate was 0%) mainly to the random factors mentioned above.

7 Conclusion

This paper presented a proposal for a Query Routing algorithm to Twitter. The purpose of this Routing algorithm is to find, among the followers of a given user (questioner), the individual most able to provide the answer. The differential of this research compared to the work here described is due to the fact that we take as our starting point a social network that does not fit into the category of Q&A Site, but where the users usually publish questions. Furthermore, the Routing algorithm was developed with facilities to be easily adaptable to the context of other social networks and to be also easily expansible. Finally we presented the model solution by using the WPM. To validate the proposed Routing algorithm a case study was performed where it was evaluated its ability to predict who would answer a question in the real world.

This study indicated that only when combined with a technique of expanding the synonymic, the Routing algorithm gets hit by more than half of the trials. This combination, when compared with the same technique without expansion, was more efficient even with respect to the rates of correct answers for each test. These results demonstrate that the proposed routing policy can still be improved.

An immediate future work includes a qualitative evaluation of the recommendations by the own questioner, besides we want increase our predict rate to 75%

of trials. Another interesting further research is a study on which factor is most important on the recommendation of experts: knowledge ($k_{f_u,q}$), trust (t_{u,f_u}) or activity (a_{f_u}); and if its importance depends on the type/topic [15] of question or another particular factor. We also proceed an investigation to establish whether the direction of questions to a user (or a small number of users) is more effective than post the question to all followers. Furthermore, aiming to improve the results obtained by Routing algorithm we can apply semantic web techniques or analyze the insertion of Bayes Theorem to calculate the probability based on the same criteria or (when the algorithm is available to Twitter community) to implement the idea presented in [20] and [13], which shows to questioner a list of recommendations and him/her takes autonomy to decide to who will direct the question.

References

1. Andrade, J.C., Nardi, J.C., Pessoa, J.M., de Menezes, C.S.: Qsabe-um ambiente inteligente para endereçamento de perguntas em uma comunidade virtual de esclarecimento. In: LA-WEB 2003 (2003)
2. Banerjee, A., Basu, S.: A social query model for decentralized search. In: Proceedings of the 2nd Workshop on Social Network Mining and Analysis, vol. 124. ACM, New York (2000)
3. Bender, M., Crecelius, T., Kacimi, M., Michel, S., Parreira, J.X., Weikum, G.: Peer-to-peer information search: Semantic, social, or spiritual. *IEEE Data Eng. Bull.* 30(2), 51–60 (2007)
4. da Silva, M.L.H.: O Twitter dentro do Universo da Cibercultura Uma Abordagem Teórica da Ferramenta. *intercom.org.br*, 1–15 (2010)
5. Davitz, J., Yu, J., Basu, S., Gutelius, D., Harris, A.: iLink: search and routing in social networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 931–940. ACM (2007)
6. Horowitz, D., Kamvar, S.D.: The anatomy of a large-scale social search engine. In: Proceedings of the 19th International Conference on World Wide Web, pp. 431–440. ACM (2010)
7. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. *First Monday* 14(1), 8 (2009)
8. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM (2007)
9. Konstan, J.A.: Introduction to recommender systems: Algorithms and evaluation. *ACM Trans. Inf. Syst.* 22, 1–4 (2004)
10. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proceedings of the First Workshop on Online Social Networks, WOSP 2008, p. 19 (2008)
11. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors. In: Most, pp. 591–600 (2010)
12. Lima, W.T., Branco, C.F.C., Barbosa, P.: Sistemas de recomendação de notícias nas mídias sociais buscam substituir o gatekeeping dos meios de comunicação de massa. *Comunicação & Inovação*, 36–45 (2009)
13. Liu, C.: AskWho: Finding Potential Answerers for Status Message Questions in Social Networks. *agora.cs.illinois.edu*, 1–5 (2010)

14. Miller, D.W., Starr, M.K.: Executive Decisions and Operations Research. Prentice-Hall, NJ (1969)
15. Morris, M.R., Teevan, J., Panovich, K.: What do people ask their social networks, and why? A Survey Study of Status Message Q&A Behavior. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 1739–1748. ACM Press, New York (2010)
16. Paul, S.A., Hong, L., Chi, E.H.: Is twitter a good place for asking questions? a characterization study. In: ICWSM (2011)
17. Ramalho, F., Robin, J.: Avaliação empírica da expansão de consultas baseada em um thesaurus: aplicação em um engenho de busca na web. RITA 10(2), 9–28 (2004)
18. Sarda, K., Gupta, P., Mukherjee, D., Padhy, S., Saran, H.: A Distributed Trust-based Recommendation System on Social Networks. In: 2nd IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb 2008). IEEE (December 2008)
19. Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Weikum, G.: Efficient top-k querying over social-tagging networks. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, p. 523 (2008)
20. Smirnova, E., Balog, K.: A User-Oriented Model for Expert Finding. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 580–592. Springer, Heidelberg (2011)
21. De Souza, C.C., de Magalhães, J.J., De Barros Costa, E.: A Formal Model To The Routing Questions Problem In The Ccontext Of Twitter. In: IADIS International Conference WWW/Internet, ICWI 2011 (2011)
22. Triantaphyllou, E., Mann, S.H.: An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox. Decision Support Systems 5(3), 303–312 (1989)
23. Ylan, Q.: Mui and Peter Whoriskey. Facebook passes Google as most popular site on the Internet, two measures show. The Washington Post (2010)

Towards a Goal Recognition Model for the Organizational Memory

Marcelo G. Armentano and Analía A. Amandi

ISISSTAN Research Institute (CONICET-UNICEN)
Campus Universitario, Paraje Arroyo Seco, Tandil, 7000, Argentina
{marcelo.armentano, analia.amandi}@isistan.unicen.edu.ar

Abstract. Automatically building a model of the different goals underlying a workflow is very important for an organization's memory since we will be able to capture the implicit knowledge that is hosted in the employees. The automatic recognition of the goal that motivates an employee to execute a particular sequence of tasks is crucial to determine what tasks are expected to be performed next in order to achieve that goal within the dynamics of the organization. Furthermore, an early recognition of the employee's goal can also prevent deviations in his/her behavior from the expected behavior by providing personalized assistance. In this article we propose a model to capture regularities in the activities carried out by employees of an organization when they are pursuing different goals. An experimental evaluation was conducted in order to determine the validity of our approach and promising results are reported.

Keywords: goal recognition, process mining, organizational memory.

1 Introduction

Knowledge is a very important value of any organization. Properly manage knowledge is crucial because a large part of it is volatile since it is hosted on individuals. This fact produces what is known as Corporate Amnesia [16], a problem characterized by the regular loss of knowledge and experience of the organization. All these skills and experience that characterizes the “know how” of any organization is called Organizational Memory.

Both corporate amnesia and organizational memory are part of a new vocabulary associated with a larger discipline known as Knowledge Management. Knowledge Management comprises a range of practices used by organizations to identify, create, represent and distribute knowledge.

The organizational memory comprises the documentation of the organization, objects or artifacts that are stored in the library or online database of the corporation and that can be used by employees familiar with the specific events and experiences of the organization. The physical evidence is known as explicit knowledge while the evidence which is hold by the employees is known as implicit knowledge. Both types of knowledge are important for efficient decision making,

to learn from past experiences and to try to replicate successful situations and avoid repeating past mistakes.

In this article we analyze the use of Variable Order Markov models to capture regularities in the activities carried out by employees of an organization when they are pursuing different goals. This model is built from the observation of the tasks that an employee performs and the goal that motivates the execution of those tasks. A learning algorithm is then executed to build a model of the tasks necessary to achieve different goals enabling this knowledge to persist in the organization, contributing to the organizational memory. Finally, this can be used to detect the goal that an employee has at any time and it will help to identify deviations from expected behavior.

This article is organized as follows. Section 2 presents some background of workflow mining and goal recognition. Section 3 describes some related work in both areas. Section 4 presents the proposed approach to perform goal recognition in process models and Section 5 describes the experiments performed to validate our approach. Finally, in Section 6 we present our conclusions.

2 Workflow Mining and Goal Recognition

The increasing use of technology to align the business processes of an organization towards the same goal has made a strong trend towards process-oriented information systems, which have a whole infrastructure to support such business processes. Despite all the advantages obtained from the use of these systems, the continuing growth and dynamics of organizations make business processes grow in number and complexity, resulting in an increasing difficulty to support its design and its rapid adaptation to changes. To help overcome these problems, a research area known as *workflow mining* [1] has emerged. Taking data from the results of the execution of processes, which are derived from the execution logs, workflow mining is based on applying different data mining techniques to obtain additional knowledge such as building a new model of a given process in order to compare it to the original process, detecting deviations from the original process or improving the process definition itself.

In this context, the automatic recognition of the goal that motivates an employee to execute a particular sequence of tasks is crucial to determine what tasks are expected to be performed next in order to achieve this goal within the dynamics of the organization. An early recognition of the employee's goal can prevent deviations of his/her behavior from the expected behavior by providing personalized assistance.

The term that has been introduced to describe the process of inferring the intentions of a subject based on the tasks that she performs in a given environment is *plan recognition*. There are currently two main approaches to the problem of *plan recognition*: consistency approaches and probabilistic approaches. Consistency approaches [15, 17] aim at reducing the set of candidates intentions by eliminating those that can not be explained by the tasks performed by the subject. On the other hand, probabilistic approaches [8, 12, 6, 11, 13, 3, 4] explicitly

accounts for the uncertainty associated with the subject's intentions and enables to create a probabilistic ranking of the possible intentions the subject might have in a given moment. Both approaches can lead to accurate predictions under the assumption that the plan library is complete and correct. However, probabilistic approaches can find the most probable intention if the observations at a specific time can lead to more than one possible intention while consistency approaches can not decide between them and must wait for a unique intention consistent with the subject's actions in order to take a decision.

Plans libraries are domain dependent and usually a domain expert encodes them manually. This task is tedious and prone to errors, and does not consider the fact that different employees can achieve a specific goal through different paths. For this reason the automatic acquisition of plan libraries is desired.

A similar situation occurs with the vast majority of *workflow management systems* (WMS) that have been created in the past and that are designed for static and structured processes to be executed in a stable environment. However, the environments in which processes are inserted nowadays create a situation in which workflows can be applied in highly dynamic scenarios. On the other side, processes have grown to a large scale and its life cycle has shortened due to the increasing complexity of the work itself and to the current dynamics of organizations. For this reason it can be inferred that designing processes is a hard task which demands too much time and, furthermore, they are always discrepancies between the process designed and the actual process.

Techniques used in workflow mining aim at redesigning, re-engineering and restructuring processes are carried out after analyzing the patterns detected at run-time. The mining process is based on extracting information from the execution of processes and automatically deriving a model which explains the events recorded in the workflow's logs. All workflow systems keep the result of the execution of its processes in log files. Therefore all information associated with these processes are recorded in those logs, such as the tasks performed, the people who performed such tasks, and the start and end time of each task. The knowledge discovered by applying process mining techniques is used in different ways to improve the original processes and to adapt them to the current scenarios.

The main difference between the process mining approach and the approach proposed in this article is that we will not focus on obtaining an explicit design of the underlying workflow. Instead, we seek to obtain a model of the tasks involved in the underlying workflow in order to detect the goal that motivates the execution of those tasks. With this information available, the system will be able to provide personalized assistance in the execution of those tasks and also to detect behaviors that do not fit the expected flow of activities.

3 Related Work

Cook and Wolf investigated the mining of processes in the context of software engineering processes. In [10] three methods for process discovery are described: one using neural networks, one using an algorithmic approach and the third

using a Markovian approach, concluding that the latter two methods are more effective. The algorithmic approach builds a finite state machine where states are fused if its possible behavior in the following k steps is identical. The Markovian approach uses a mixture of algorithmic and statistical methods and has the fundamental advantage of being robust in the presence of noise. The difference between the approach presented in this article and the Markovian approach proposed by Cook and Wolf is that the latter used Markov chains of first and second order, while we use variable order Markov models that are able to model both short and long time dependencies, depending on the statistical information contained in the logs.

Wen et. al [20] presented an algorithm based on two types of events that indicate the beginning and completion of tasks. This information is used to explicitly detect parallelism. Together with the causality information obtained from the activities log, they derive relationships between tasks which are then used to create a Petri net modeling the underlying process. The proposed algorithm, called β -algorithm, is an extension of the α -algorithm proposed in [2], solving some of the limitations of the latter such as the detection of short cycles. A disadvantage of this approach is that it is not probabilistic. That is, the model obtained is able to model different paths to follow to comply with a given process, but it is not able to model which of these paths is more likely to be executed by the employee. Furthermore, this approach is not robust to the presence of noise in the training data.

The two studies described above focus on building a model of the activities corresponding to a workflow and not on the on-line detection of the goal that motivates an employee to perform a set of tasks. This line of work corresponds to the area of goal (or plan) recognition. In this direction we consider the prediction of the goal of a subject as an inherently uncertain task. That is why we seek for a knowledge representation capable of capturing and modeling this kind of uncertainty. Non-probabilistic approaches [15,17] have the disadvantage of not being able to decide to what extent the observed evidence supports any particular hypothesis about the goal of the subject. This is an important consideration so as to generate a probabilistic ranking with the different possible explanations for a set of actions performed by the subject. Bayesian networks and Markov models are two of the most used representations that deal with this type of information.

Bayesian networks have been used successfully in previous studies of goal recognition [8,3]. While it is possible to achieve good results using this knowledge representation, learning Bayesian networks from logs is expensive and requires training examples with a constant number of tasks. However, this is not the case when we look at the tasks performed by an employee in an organization. Furthermore, there may be different ways to meet a goal and each of these ways may require the execution of a different number of tasks. These problems are often treated using a sliding window with a fixed size so as to segment the activities log, but this introduces fictional examples that can result in an incorrect detection of the causalities among the activities.

There are other two fundamental problems in using Bayes networks as a representation of knowledge in this domain. The first is that the result of the propagation of evidence in the network is not sensitive to the order in which the activities are observed. This means that if the sequence of activities performed by an employee is entered into the network in a totally different order, the result of the evidence propagation would be exactly the same. The second problem is that Bayesian network, by definition, are not capable of representing cycles.

On the other hand, Blaylock and Allen [5,6] studied the use of Markov models of first and second order and hidden Markov models for goal recognition in the domain of a Linux system. They defined the problem of goal recognition as a classification problem in which given a sequence of observations, the algorithm searches for the goal that more probably explain those observations.

Markov chains of first order can be derived without any prior knowledge of the model's structure, but the Markov assumption is not realistic in the domain of process mining. Markov chains of higher order have the disadvantage of having an exponential growth in the state space as the order of the model increases to capture longer time dependencies. For this reason we propose to use variable order Markov models, which extend Markov chains of fixed order by allowing the number of random variables that determine each state of the model depend on different specific contexts. These contexts represent sequences of previous observation assuming that future states are independent of past states earlier this context.

4 Proposed Approach

4.1 Learning a Workflow's Goals Model from Examples

Markov models are a natural way of modeling sequences of events observed along time. In its simplest form, a Markov chain is a stochastic process with the Markov property. Having the Markov property means that, given the present state, future states are independent of the past states. In other words, the description of the present state fully captures all the information that could influence the future evolution of the process. Future states will be reached through a probabilistic process instead of a deterministic one. At each step the system may change its state from the current state to another state, or remain in the same state, according to a certain probability distribution. The changes of state are called transitions, and the probabilities associated with various state-changes are called transition probabilities.

Markov chains of fixed order are a natural extension in which the future state is dependent on the previous m states. Although this extension is beneficial for many domains, there are some main drawbacks in the use of these models. First, only models with very small order are of practical value since there is an exponential grow in the number of states of Markov chains as their order is increased. Second, for sequences of tasks performed by an employee to achieve a given goal, the probability of the next performed task is not always determined by the same

fixed number of previous tasks. There is usually a variable length previous “context” that determines the probability distribution of what the employee may perform next.

Hidden Markov Models are an alternative way of modeling natural sequences. Although these models are a powerful and popular representation, there are theoretical results concerning the difficulty of their learning [18].

Variable Order Markov (VOM) models arose as a solution to capture longer regularities while avoiding the size explosion caused by increasing the order of the model. In contrast to the Markov chain models, where each random variable in a sequence with a Markov property depends on a fixed number of random variables, in VOM models this number of conditioning random variables may vary based on the specific observed realization, known as *context*. These models consider that in realistic settings, there are certain realizations of states (represented by contexts) in which some past states are independent from the future states conducting to a great reduction in the number of model parameters.

Algorithms for learning VOM models over a finite alphabet Σ attempt to learn a subclass of Probabilistic Finite-state Automata (PFA) called Probabilistic Suffix Automata (PSA) which can model sequential data of considerable complexity. Formally, a PSA can be described as a 5-tuple $(Q, \Sigma, \tau, \gamma, \pi)$, where Q is a finite set of states, Σ is the task universe, $\tau : Q \times \Sigma \rightarrow Q$ is the transition function, $\gamma : Q \times \Sigma \rightarrow [0, 1]$ is the next task probability function, where for each $q \in Q$, $\sum_{\sigma \in \Sigma} \gamma(q, \sigma) = 1$, $\pi : Q \rightarrow [0, 1]$ is the initial probability distribution over the starting states, with $\sum_{q \in Q} \pi(q) = 1$.

A PFA is a PSA if the following property holds. Each state in a PSA M is labeled by a sequence of tasks with finite length in Σ^* and the set of sequences S labeling the states is suffix free. Σ is the domain task universe, that is the finite set of tasks that the employee can perform in the domain. A set of sequences S is said to be suffix free if $\forall s \in S, Suffix^*(s) \cap S = \{s\}$, where $Suffix^*(s) = \{s_i, \dots, s_l | 1 \leq i \leq l\}$ is the set of all possible suffixes of s , including the empty sequence ϵ . For every two states q_1 and $q_2 \in Q$ and for every task $\sigma \in \Sigma$, if $\tau(q_1, \sigma) = q_2$ and q_1 is labeled by a sequence s_1 , then q_2 is labeled by a sequence s_2 that is a suffix of $s_1 \cdot \sigma$.

In contrast to m -order Markov models, which attempt to estimate conditional distributions of the form $Pr(\sigma|s)$, with $s \in \Sigma^N$ and $\sigma \in \Sigma$, VOM algorithms learn such conditional distributions where context lengths $|s|$ vary in response to the available statistics in the training data. Thus, PSA models provide the means for capturing both large and small order Markov dependencies based on the observed data. In [4] an algorithm for learning such models in an incremental way is proposed.

Learning a workflow model from activities logs has the main advantage that we do not need any additional information about the domain being modeled more than the tasks that can be performed in the domain. We will be able to learn regularities in the employees' behavior just by analyzing the trace examples

observed in the logs. In the following section we describe how we use PSA's models to perform goal recognition.

4.2 Recognizing an Employee's Goals withing a Workflow

To perform goal recognition, we build a PSA model for each different goal in the domain. By having a separate model for each goal, we will be able to track several goals that are being pursued simultaneously by the employee.

Conventionally, to compute the probability assigned by a PSA k to a given sequence of observations, we should compute $P_{PSA_k}(r) = \prod_{i=1}^N \gamma(s_{i-1}, r_i)$, where $\gamma(s_{i-1}, r_i)$ is the probability value assigned in state s_{i-1} to the observed task r_i , and will select the PSA that assigns the maximum probability as the PSA corresponding to the employee's intention. However, as the employee continues performing tasks, the total cumulative probability value assigned by each PSA will become smaller and smaller as we are multiplying values in the range $(0, 1]$. Furthermore, we must consider the uncertainty related to the moment in which the employee starts a new plan to achieve a new goal. We will face with a continuous stream of tasks and we should be able to recognize changes in the employee's current goal. Moreover, the goal recognition process should not be affected by the execution of noisy tasks. The problem we are facing is not a classical problem of classification as we do not predict a "class" (goal) after observing a complete sequence of tasks. In this domain, we should be able to predict the most probable employee goal after each performed task, and the limit between sequences of tasks corresponding to different goals might be fuzzy.

To tackle this problem we use an *exponential moving average* on the prediction probability $\gamma(s, \sigma)$ at each step in each PSA as the predicted value for each corresponding employee goal. Moving averages are one of the most popular and easy to use tools to smooth a data series and make it easier to spot trends. An exponential moving average (EMA) [14] is a statistic for monitoring a process that averages the data in a way that gives less and less weight to data as time passes. The weighting for each step decreases exponentially, giving much more importance to recent observations while still not discarding older observations entirely. By the choice of a weighting factor $0 \leq \lambda \leq 1$, the EMA control procedure can be made sensitive to a small or gradual drift in the process. Alternatively, λ may be expressed in terms of N time periods, where $\lambda = \frac{2}{N+1}$.

EMA_t expresses the value of the EMA at any time period t . EMA_1 is set to the a priori probability of the first observed task σ . Then, the computation of the EMA at time periods $t \geq 2$ is done according to equation \square

$$EMA_t = \lambda \gamma_{PSA_i}(s, \sigma) + (1 - \lambda) EMA_{t-1} \quad (1)$$

The parameter λ determines the rate at which *older* probabilities enter into the calculation of the EMA statistic. A value of $\lambda = 1$ implies that only the most recent measurement influences the EMA. Thus, a large value of λ gives more weight to recent probabilities and less weight to older probabilities; a small

value of λ gives more weight to older probabilities. The value of λ is usually set between 0.2 and 0.3 [14] although this choice is somewhat arbitrary and should be determined experimentally.

To sum up, the goal recognition process works as follows: as the user performs activities in the application at issue we keep making the corresponding state transitions in each PSA and computing the exponential moving average of the transition probability of the performed tasks given each PSA. At each step, we will own a probabilistically ranked set of PSAs which correspond to the most probable goals the user may have at each moment.

5 Experimental Evaluation

5.1 Experiment Set-Up

Process logs corresponding to real-world business process are hardly available. Harder to get are process logs in which different goals have been identified and associated with the corresponding log traces. For these reasons, we have tested our technique with a simulated scenario, as in [19] and [9]. The main benefit of using simulation is that properties such as noise can be controlled. We defined five different workflows using PLG [7], a framework that enables the generation of random business processes according to some specific user-defined parameters. The generated workflows represent five different goals that the employees can pursue by performing the same set of 18 activities. Figure 1 shows the Petri net representation of one of the workflows used in the experiment and Table 1 shows some statistics for the five process’s goals used in the experiment.

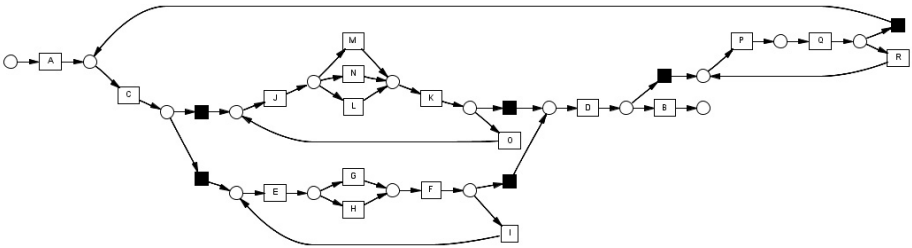


Fig. 1. Petri net representation of the workflow for one of the goals used in the experiment

Next, for each workflow, we use PLG to generate 100 traces and filtered out duplicate traces. Finally, we trained five different VOM models to be used in the experiments. In the following section we present the goal recognition performance using this dataset.

Table 1. Statistics for the processes used in the experiment

	G1	G2	G3	G4	G5
Number of AND patterns	0	2	2	1	1
Number of XOR patterns	3	1	1	2	1
Number of loops	4	2	0	1	1
Maximum number of AND branches	0	3	2	3	2
Maximum number of XOR branches	3	3	2	3	2
Total number of activities	18	18	12	18	11

5.2 Goal Recognition Performance

We evaluated our approach in terms of online accuracy and convergence, defined as follows:

- *Online accuracy.* For a certain goal, online accuracy is defined as the number of time-steps in which the employee’s current goal was the most probable goal divided by the number of performed tasks. In other words, if we assume that the recognizer makes predictions after each observed task, the online accuracy of a sequence $Seq = \sigma_1, \sigma_2, \dots, \sigma_N$ for a given goal q is computed as $accuracy_{online}(q, Seq) = \frac{\sum_{i=1}^N best_q(\sigma_i)}{N}$, where $best_q(\sigma_i) = \begin{cases} 1 & \text{if } q(\sigma_i) = q_{best}(\sigma_i) \\ 0 & \text{otherwise} \end{cases}$, $q(\sigma)$ is the resulting EMA value assigned by PSA q after observing the task σ given the previous context, and $q_{best}(\sigma)$ is the highest resulting EMA value assigned by all PSAs after observing task σ given the previous context
- *Convergence.* When applied to a specific goal, this criterion indicates the average percentage of observations after which a recognized goal converges to the correct answer. It measures how fast the recognition process of a goal converges to the correct answer. If the algorithm predicted correctly the actual employee’s goal from the time-step t to the time-step corresponding to the last performed task, then the convergence is computed as $convergence(q, seq) = \frac{N-t+1}{N}$, with $not_best_q(\sigma_{t-1})$ and $best_q(\sigma_j), \forall j \ t \leq j \leq N$. The time-step t is called convergence point.

We ran our experiments by doing a leave-one-out cross validation over the collected logs for each workflow. We tested different values for the smoothing constant, ranging from 0.1 to 1.0 with intervals of 0.1. Figure 2 shows the resulting average values for the five workflow’s goals.

We obtained better convergence and online accuracy with a value of $\lambda = 0.2$. This is equivalent to a (smoothed) sliding window on the last nine observed actions (remember that λ can be expressed as $\lambda = \frac{2}{N+1}$, where N is the number of last actions that concentrate the highest weights in the EMA computation). As expected, accuracy is higher than convergence in all cases since the recognizer is often “confused” at the beginning of the observed sequence of tasks, making some good and bad predictions until it finally converges. With $\lambda = 0.2$ convergence is

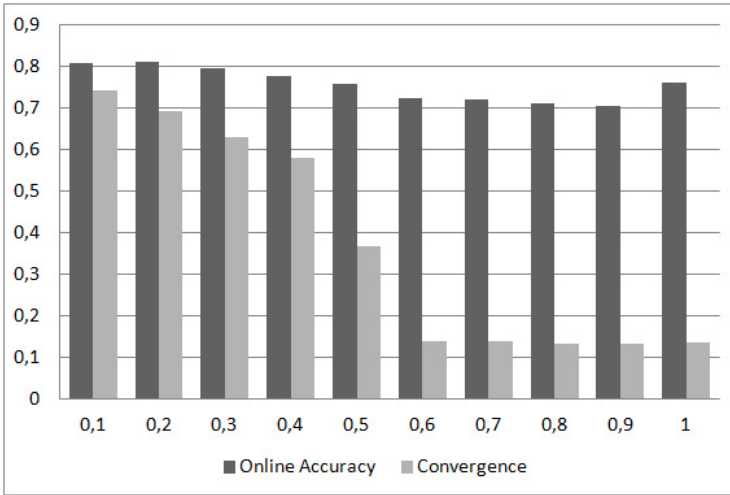


Fig. 2. Accuracy and convergence for different values of the smoothing constant λ

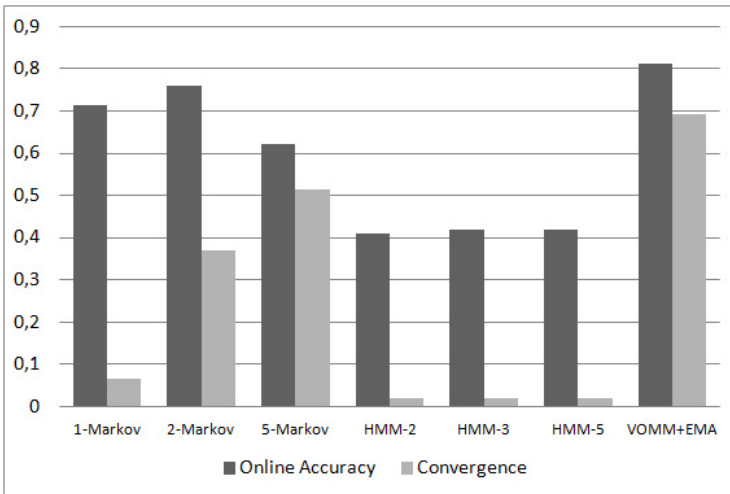


Fig. 3. Comparison with fixed-order and hidden Markov models

reached on average when only 30% the sequence has been observed, confirming our hypothesis that the smoothing constant enables an earlier detection of the user's current goal.

In order to compare these results with competing approaches, we performed the same experiments using fixed-order Markov models and Hidden Markov models. Figure 3 shows a comparison of online accuracy and convergence obtained for Markov models of order 1, 2 and 5 (indicated as 1-Markov, 2-Markov and

5-Markov in Figure 3, respectively), HMM with 2, 3 and 5 hidden states (indicated as HMM-2, HMM-3 and HMM-5 in Figure 3, respectively), and those obtained with VOM models with EMA smoothing (VOMM+EMA).

We can see that VOM models outperformed fixed order Markov models and HMM for the two metrics considered. There is an increase in convergence as we use a longer context for fixed-length Markov models. However, using fixed contexts longer than two does not improve online accuracy. On the other hand, HMM gave the worst results and trend to keep similar average performance when we vary the number of hidden states. By using VOM models with EMA smoothing we obtained 81.16% online accuracy and 69.33% convergence, an improvement of 5.23% and 17.96% respectively as compared to 2-Markov and 5-Markov, which resulted to be the best competing models for each metric. This confirms our hypothesis that VOM models enable a better modeling of the sequences of tasks needed to achieve different goals within a workflow, keeping short contexts when possible but using longer contexts when they give a better prediction than a shorter context.

6 Conclusions

In this article we addressed the problem of building a model of the different goals that an employee can pursue within the workflow of an organization. This model is automatically built from example sequences of tasks performed to achieve each goal. This fact is very important to the organizational memory of a company since we will be able to capture the implicit knowledge that is hosted in the employees in the form of a model representing how different goals are commonly achieved by them.

We found that Variable Order Markov models are a good alternative for capturing regularities in the employees' behavior in their task of accomplishing different goals. VOM models extend the well-known Markov chains by making variable the order of the model. In this way, longer order contexts are used only if they enable better predictions than shorter contexts. We also proposed the use of an exponential moving average for smoothing the predicted probabilities in order to better capture the underlying tendency in the goal followed by the employee.

Finally, we evaluated the proposed approach in a simulated scenario with five different goals. We conclude that our approach facilitates an early and accurate prediction of the employee's goal, when only 30% of the length of the sequence to achieve a given goal has been observed. We obtained an improvement of 5.23% in online accuracy and 17.96% in convergence compared to Markov models of second and fifth order, respectively.

One limitation of our approach is that we need example sequences of tasks representing how different goals in the workflow are achieved. Workflow logs are rarely tagged with this kind of information. Expressiveness is also reduced since we are not reconstructing the underlying workflow of a log. Nevertheless the graphical representation of the VOM model in the form of a PSA can give us an idea of contexts of tasks that determine the different transitions in the model.

Finally, VOM models are sensitive to the number of actions in the training sequences. On the one hand, if the provided training sequences are too short, we cannot take advantage of the variable order part of our models, since there is not enough statistical information to learn about. In this case, VOM models will be equivalent in performance to fixed order Markov models. On the other hand, if the model for a given goal is trained using short sequences and the model for another goal is trained with longer sequences containing the same symbols, the first model will be predicted after observing the same sequence of actions. This happens because the probability mass is distributed among fewer contexts and the corresponding goal will be predicted with higher values. A similar problem occurs with the EMA computation. If sequences are too short, the memory of the model will not be activated and using a smoothing constant with higher values can lead to better prediction results.

References

1. van der Aalst, W.: *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer (2011)
2. van der Aalst, W., Weijters, T., Maruster, L.: Workflow mining: Discovering process models from event logs. *IEEE Trans. on Knowl. and Data Eng.* 16, 1128–1142 (2004)
3. Armentano, M., Amandi, A.: Personalized detection of user intentions. *Knowledge-Based Systems* 24(8), 1169 (2011)
4. Armentano, M., Amandi, A.: Modeling sequences of user actions for statistical goal recognition. In: *User Modeling and User-Adapted Interaction*, vol. 22(3), pp. 281–311. Springer (2012), ISSN:0924-1868 (Print) 1573-1391 (Online)
5. Blaylock, N., Allen, J.: Corpus-based, statistical goal recognition. In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico*, pp. 1303–1308 (2003)
6. Blaylock, N., Allen, J.: Recognizing instantiated goals using statistical methods. In: Kaminka, G. (ed.) *IJCAI Workshop on Modeling Others from Observations (MOO 2005)*, Edinburgh, pp. 79–86 (2005)
7. Burattin, A., Sperduti, A.: Plg: a framework for the generation of business process models and their execution logs. In: *Proceedings of the 6th International Workshop on Business Process Intelligence (BPI 2010)*, Hoboken, New Jersey, USA (2010)
8. Charniak, E., Goldman, R.P.: A bayesian model of plan recognition. *Artificial Intelligence* 64(1), 53–79 (1993)
9. Claes, J., Poels, G.: Merging Computer Log Files for Process Mining: An Artificial Immune System Technique. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) *BPM Workshops 2011, Part I. LNBIP*, vol. 99, pp. 99–110. Springer, Heidelberg (2012)
10. Cook, J.E., Wolf, A.L.: Discovering models of software processes from event-based data. *ACM Trans. Softw. Eng. Methodol.* 7, 215–249 (1998)
11. Duong, T.V., Phung, D.Q., Bui, H.H., Venkatesh, S.: Human behavior recognition with generic exponential family duration modeling in the hidden semi-markov model. In: *International Conference on Pattern Recognition*, vol. 3, pp. 202–207 (2006)
12. Geib, C., Goldman, R.: Partial observability and probabilistic plan/goal recognition. In: *IJCAI 2005 Workshop on Modeling Others from Observations, Edinburgh, Scotland* (2005)

13. Hu, D.H., Yang, Q., Li, Y.: An algorithm for analyzing personalized online commercial intention. In: Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising, ADKDD 2008, pp. 27–36. ACM, New York (2008)
14. Hunter, J.S.: The exponentially weighted moving average. *Journal of Quality Technology* 18(4), 203–209 (1986)
15. Kautz, H.: A formal theory of plan recognition and its implementation. In: Allen, J.F., Kautz, H.A., Pelavin, R., Tenenber, J. (eds.) *Reasoning About Plans*, pp. 69–125. Morgan Kaufmann Publishers, San Mateo (1991)
16. Kransdorff, A.: *Corporate Amnesia: Keeping the Know-How in the Company*. Butterworth Heinemann (1998)
17. Rich, C., Sidner, C.L., Lesh, N.: COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine* 22(4), 15–26 (2001)
18. Ron, D., Singer, Y., Tishby, N.: The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning* 25(2-3), 117–149 (1996)
19. Weijters, A.J.M.M., van der Aalst, W.M.P.: Rediscovering workflow models from event-based data using little thumb. *Integr. Comput.-Aided Eng.* 10, 151–162 (2003)
20. Wen, L., Wang, J., Aalst, W.M., Huang, B., Sun, J.: A novel approach for process mining based on event types. *J. Intell. Inf. Syst.* 32, 163–190 (2009)

SART: A New Association Rule Method for Mining Sequential Patterns in Time Series of Climate Data

Marcos Daniel Cano¹, Marilde Terezinha Prado Santos¹, Ana Maria H. de Avila²,
Luciana A.S. Romani³, Agma J.M. Traina⁴, and Marcela Xavier Ribeiro¹

¹ Computer Science Department, Federal University of Sao Carlos, São Carlos - SP, Brazil
{marcos_cano, marilde, marcela}@dc.ufscar.br

² Cepagri, Unicamp, Campinas - SP, Brazil
avila@cpa.unicamp.br

³ Embrapa Agriculture Informatics, Campinas - SP, Brazil
luciana@cnptia.embrapa.br

⁴ Computer Science Department, University of Sao Paulo, São Carlos, Brazil
agma@icmc.usp.br

Abstract. Technological advancement has enabled improvements in the technology of sensors and satellites used to gather climate data. The time series mining is an important tool to analyze the huge quantity of climate data. Here, we propose the *Sequential Association Rules from Time series* - SART method to mine association rules in time series that keeps the information of time between related events through an overlapped sliding-window approach. Also the proposed method mines association rules, while the previous ones produce frequent sequences, adding the semantic information of confidence, which was not previously defined by sequential patterns. Experiments were conducted with real data collected from climate sensors. The results showed that the proposed method increases the number of mined patterns when compared with the traditional sequential mining, revealing related events that occur over time. Also, the method adds the semantic information related to the confidence and time to the mined patterns.

Keywords: sequential association rule, time series data mining, sliding window.

1 Introduction

Human knowledge comes from historical and current observations that are analyzed for decision making. Thus, the study of time series is important for the prediction and identification of possible patterns that occur in the flow of time. Time Series are generated and studied in several areas such as economy, health, telecommunication, weather, geosciences and more recently in remote sensing. One of the fields that employ time series analysis is the Agrometeorology. Climatology and agrometeorology researchers are interested in finding sets of related events. For example, they wish to relate the impact of a temperature change in the production of sugarcane.

With the development and the improvement of the data acquisition technology, climate data from surface stations, remote sensors, weather radars or sensor network have been increased, generating terabytes of data to be analyzed. Data mining is the computation area responsible for the developing of automatic or semi-automatic techniques to analyze the gathered data. A very important area in the data mining is the mining of association rules that concerns to the search for relations of occurrences among items in a database. The traditional association analysis is intra-transactional, since it reveals only relations among items that occurred in the same transaction. On the other hand, when working with time series, the association rules should be inter-transactional. The inter-transactional association rules comprises in patterns relating items occurrence over one or more transactions. In this sense, intra-transaction rules are a subset of the inter-transaction ones [16].

When working in the climate domain, if the classical association rule mining is employed, it is possible to obtain rules relating events at the same time. For instance, “in a given moment, if the temperature is low, then the precipitation is low”. That rule associates items found in the same transaction, so it is intra-transactional. One of the problems of it is that it cannot capture the time evolution between the climate phenomena, such as, “if in this month the temperature is high, then in the next month the precipitation will be high”. The last rule is an example of an inter-transactional rule and the major advantage of it is the possibility of finding patterns along a period of time.

Often, in the analysis of patterns, meteorologists and agrometeorologists use statistical models based on principal components analysis, cluster analysis, frequency distribution, geostatistics, non-parametric statistics, among others. However, there are few specific data mining techniques for climate and remote sensing databases [7].

This paper proposes a new method, named SART - *Sequential Association Rules from Time series* - to mine association rules in time series sequences. The rules are mined employing a sliding window that allows the mining of patterns relating event occurrences over time. The method broadens the exploratory power of the sequential methods of association rule mining, including the time dimension in the search process. The experiments show that our proposed method produces semantically richer patterns and in larger number, when compared to the traditional sequential association mining.

The paper is organized as follows. Section 2 presents a review of the literature regarding sequential analysis and inter-transactional association rules. In Section 3, describes the proposed method. In Section 4 presents the experiments. Finally, Section 5 presents the conclusions and future work.

2 Background and Related Work

Association Rules were proposed by Agrawal et Al. [1] to solve the problem of discovering sets of items that frequently occur together in transactions of a database.

Let $I = \{I_1, \dots, I_n\}$ be a set of literals, called items or candidates. A set $X \in I$ is called itemset. An itemset X with k elements is called itemset- k . Let R be a table

with tuples t involving elements that are subsets of I . The tuple t supports an itemset X , if $X \subseteq t$. An association rule is an expression of the form $X \rightarrow Y$, where X and Y are itemsets. X is called body or antecedent and Y is called head or consequent of the rule $X \rightarrow Y$. The *support* of an itemset X is the ratio between the number of tuples in R that support X and the total number of tuples of R . The confidence of a rule $X \rightarrow Y$ is the proportion of tuples containing X that also contains Y . The problem of association rule mining, as it was firstly stated, involves finding rules that satisfy the constraints of minimum support (*minsup*) and minimum confidence (*minconf*) specified by the user.

Traditional association rule mining techniques relate data items disregarding the sequence of their occurrences. However, when analyzing climate data it is necessary to consider the sequence of items occurrence. In this sense, sequential pattern mining was introduced in [2] as a problem of finding sequences of item occurrences derived from the association rule mining problem. In the last decade, techniques have been proposed to discover sequential patterns in temporal data [2,4,5]. Most of the earlier algorithms for sequential pattern mining are based on the Apriori property [17]. The Apriori property states that any sub-pattern of a frequent pattern (a pattern that satisfies the restriction of *minimum support*) must also be frequent. Apriori-like algorithms adopt a multiple-pass candidate generation-and-test strategy. In order to reduce the time taken to generate candidates, algorithms based on pattern growth were developed. In this sense, most algorithms of sequential association rules mining are based on Apriori or on pattern growth approaches [8].

The GSP algorithm [4], which is based on Apriori, iteratively counts the occurrences of candidate sequences (potential frequent sequences), requiring as many database scans as the longest frequent sequence of the database. The SPADE algorithm [9] is also based on Apriori. It adopts a vertical data format where the original problem is decomposed into smaller sub-problems that can be solved in main-memory using lattice search techniques. Another example of Apriori-based algorithm is FITI [11], which identifies inter-transactional patterns formed by extended-items. Hu [13] introduces a multi-time-interval sequential pattern, which reveals the time-intervals between all pairs of items in a pattern. He proposes an Apriori-based algorithm called MI-Apriori.

In [6] E-Apriori and EH-Apriori were proposed. As previously reported in [10], the most costly operation are L_1 (identification of the set of frequent itemsets of size 1) and L_2 (identification of the set of frequent itemsets of size 2) iterations, i.e., L_1 and L_2 dominate the total cost of mining. A large number of L_1 elements results in a large number of itemsets to process C_2 (set of candidate itemsets – potential frequent itemsets of size 2). In order to build a smaller set C_2 , EH-Apriori adopts a hashing technique similar to the one proposed in [10] to filter out unnecessary candidate two-extensive itemsets in advance, so all possible two-extensive itemsets are inserted into the Hash table.

The PrefixSpan [5] algorithm explores a pattern-growth approach and adopts a divide-and-conquer strategy. The database is recursively projected into a set of smaller projected databases based on the current sequential pattern. The sequential patterns are grown in each projected database by exploring only locally frequent fragments.

The main idea of PrefixSpan is: any frequent subsequences can always be found by growing frequent prefixes. Given a sequence database S the PrefixSpan algorithm can be defined in a recursive sequence of steps [5]:

PrefixSpan(α , i , $S|\alpha$)

1. Scan $S|\alpha$ once, find the set of frequent items b such that
 - b can be assembled to the last element of α to form a sequential pattern; or
 - $\langle b \rangle$ can be appended to α to form a sequential pattern.
2. For each frequent item b , append it to α to form a sequential pattern α' , and output α' ;
3. For each α' , construct α' -projected database $S|\alpha'$, and call PrefixSpan(α' , $i+1, S|\alpha'$).

The major advantage of PrefixSpan is that it does not require the candidate generation step as the Apriori-like approaches, reducing the computation cost of successive scans in the database. PrefixSpan mines projected databases. In this study, our database is a long continuous sequence. Thus, we adapted the PrefixSpan algorithm to mine a continuous sequence database.

Few PrefixSpan-based algorithms have been proposed. Examples of them are FreeSpan [14], I-PrefixSpan [15] and MI-PrefixSpan [13]. FreeSpan [13] creates projected databases based on the current set of frequent patterns without a particular ordering (i.e., growth direction), whereas PrefixSpan projects databases by growing frequent prefixes. FreeSpan is based on the following property: if an itemset X is infrequent, any sequence whose projected itemset is a superset of X cannot be frequent. FreeSpan mines sequential patterns by partitioning the search space and projecting the sequence subdatabases recursively based on the itemsets. Yet, the general idea of I-PrefixSpan [15] is to use a data structure to reduce the execution time and memory usage. Lee and Wang [18] proposed a method for mining frequent inter-transaction patterns consisting of two phases: the first one is prepared with two data structures, called a dat-list, which stores the information of items used to find the frequent inter-transaction patterns and an ITP-tree. The ITP-tree stores the frequent patterns discovered. In the second phase, the algorithm calls ITP-Miner (Inter-Transactions Patterns Miner) to mine all frequent inter-transaction patterns.

In general, algorithms of sequential pattern mining reveal patterns from event sequences, considering just the relation of order among the events but not the time interval taken among them. Moreover, algorithms of sequential pattern mining reveal frequent itemset sequences and not association rules. When working in climate domain, the time between events should be taken into account, since time is an essential feature to characterize climate phenomena. Also, it is important to preserve the relation of cause and consequence that exists in association rules to enhance the analysis of climate changes. In this paper, we propose to adapt a sequential pattern mining technique to work in times series analysis, allowing the mining of sequential association rules taking into account the time resolution between the events and the confidence of the mined rules.

The proposed method pre-processes time series using a sliding window approach producing sequences. The sequences are submitted to the mining algorithm, which is an adaptation of PrefixSpan. This adaptation employs a new support measure that computes the items occurrences regarding time series sequences. The method defines a new measure of confidence that quantifies the strength of a sequential association rule. In addition, the method incorporates in the mining algorithm the rule generation step and the calculus of the confidence measure, allowing the mining of association rules, which was not previously implemented in sequential pattern mining approaches.

3 Proposed Method: Sequential Association Rules from Time Series (SART)

We propose a new method, called *Sequential Association Rules from Time series* - SART, to mine sequential patterns from time series. The proposed method and the definitions employed to understand it are given, as follows.

A **unidimensional time series** sequence S is defined as a sequence of pairs (a_i, t_i) , with $i = \{1, \dots, n\}$:

$S = [(a_1, t_1), \dots, (a_i, t_i), \dots, (a_n, t_n)]$, where $(t_1 < \dots < t_i < \dots < t_n)$ and each a_i is an event, i.e. the value of attribute a at time t_i .

A **multidimensional time series** sequence S is a sequence of sets of events in the format $(a_i b_i \dots m_i, t_i)$, with $i = \{1, \dots, n\}$, where $a_i, b_i \dots m_i$ are events that occurs at the time value t_i . A **multidimensional time series** sequence S is defined as:

$S = [(a_1 b_1 \dots m_1, t_1), \dots, (a_i b_i \dots m_i, t_i), \dots, (a_n b_n \dots m_n, t_n)]$, where $(t_1 < \dots < t_i < \dots < t_n)$,

Since the SART method works using a sequential approach, its first step consists of transforming an input time series S in a database D of data sequences. This is performed using a sliding window approach, as described below.

A **window** w of an n -dimensional time series S is a block of events that occurs in a continuous interval, starting at time t_s and ending at time t_e such that events t_s and t_e belongs to S . The **size** $d = |w|$ is the number of consecutive sets of events $(a_i b_i \dots m_i, t_i)$ occurred at time t_i kept by the window, $0 \leq d < n$. A **sliding window** $W[j]$ at position j of the time series S is a walking window that scans sub-sequences $S_j \subset S$, starting at time t_s . Two consecutive sliding windows S_j and S_{j+1} , $0 \leq j < n$, may overlap, i.e. $S_j \cap S_{j+1} = S' \mid S' \subset S_j$ and $S' \subset S_{j+1}$. The size $v = |S'|$, $0 \leq v < d$ is called **overlap size**. Recall that the overlap size should be smaller than the window size in order to keep the sliding of the window. For the proposed method, we considered constant the window size d and the overlap size v . The number of time measures walked at each step of the sliding window is $s = d - v$. Therefore, we can define a sliding window $W[j]$ at the position j of a time series sequence S as a subsequence S_j :

$W[j] = S_j = [(a_j, b_j, \dots, t_j), (a_{j+1}, b_{j+1}, \dots, t_{j+1}), \dots, (a_{j+d}, b_{j+d}, \dots, t_{j+d})]$, $j = sk, k = 1 \dots n/s$.

The **first step** of SART is to process the time series S producing sequences S_j , employing two input parameters: the window size d and the overlap size v .

Let an itemset X be a set of database items that occurs simultaneously. An *itemset sequence* is defined as $P = \langle (X_1), \dots, (X_i), \dots, (X_m) \rangle$, where $m \geq 1$ and X_i is the i^{th} itemset occurred in time order. An itemset is delimited by parentheses. If an itemset has only one element (*itemset-1*), it can be represented without the parentheses notation. In a sequence P , the time order between itemsets occurrences is: the itemset X_i occurs before X_{i+1} .

The **second step** of SART consists of finding *frequent itemset sequences* P . An itemset sequence is **frequent** if it satisfies minimum support *minsup*. The support of an **itemset sequence** P is defined as:

$$sup(P) = \frac{|P|}{(v + 1)|D|} \tag{1}$$

where $|D|$ is the number of sequences from the database D , $|P|$ is the number of occurrences of the itemset sequence P in the database D , v is the overlap size of the sliding window approach that map the time series S in sequences S_j .

We define a sequential association rule generated from the itemset sequence $P = \langle (X_1), (X_2), \dots, (X_i), \dots, (X_m) \rangle$ as an expression of the form:

$$X_1 \rightarrow Z, \text{ where } Z = \langle (X_2), \dots, (X_m) \rangle, \text{ and } X_i \text{ occurs before } X_{i+1}$$

The rule $X_1 \rightarrow Z$ indicates that, if the starting sequence itemset X_1 occurs, than the remaining of the itemset sequence Z tends also occurs. A sequential association rule relates the starting itemset of a sequence to the subsequent ones.

To the best of our knowledge, there are no approaches in the literature that calculates the confidence measure for sequential association rules. However, the confidence measure is an important indication of the statistical strength of the rule. We define the **confidence** of a sequential association rule $X_1 \rightarrow Z$ as:

$$conf (X_1 \rightarrow Z) = \frac{sup(X_1)}{sup(Z)} \tag{2}$$

The confidence measures the probability of the subsequence of itemsets Z occurs after X_j . Since the itemset X_j is the set of events that triggers Z , the confidence, as stated by Equation 2, is also called the *trigger-confidence* of the rule.

The **third step of** SART consists of finding strong rules from the frequent sequences produced in the second step. The **strong rules** are the ones that satisfy the minimum confidence input threshold *minconf*. To perform this, each frequent itemset sequence $P = \langle (X_1), (X_2), \dots, (X_i), \dots, (X_m) \rangle$ is employed to generate the rules $X_1 \rightarrow Z$, where $Z = \langle (X_2), \dots, (X_m) \rangle$, and the confidence of it is calculated according to Equation 2. The method outputs the strong rules and eliminates the non-strong ones. Algorithm 1 describes the steps of the proposed method SART.

Algorithm 1. Method SART

Input: multidimensional time series $S=[(a_1, b_1, \dots, m_1, t_1), \dots, (a_n, b_n, \dots, m_n, t_n)]$, window size d , overlap size v , minimum support $minsup$, minimum confidence $minconf$

Output: set R of strong sequential association rules $X_1 \rightarrow Z$

Step 1: Scan the time series S using the sliding window $W[j]$ producing the set W of time series sequences

1 **for** $k=1$ to n/s

2 $s=d-v$; $j=s \times k$

3 $W[j] = S_j = [(a_j, b_j, \dots, t_j), (a_{j+1}, b_{j+1}, \dots, t_{j+1}), \dots, (a_{j+d}, b_{j+d}, \dots, t_{j+d})]$

4 $W = \cup W_j$

Step 2: Find the set F of frequent itemset sequences from W

5 **for** each itemset sequence $P \subset W$

6 calculate $sup(P)$ according to Eq.(1) using *PrefixSpan*

7 $F = \cup P \mid P \subset W \wedge sup(P) \geq minsup$

Step 3: Find the set R of strong association rules from F

8 **for** each itemset sequence $P \subset F$

9 generate a rule of the form $X_1 \rightarrow Z$, where $Z = \langle (X_2), \dots, (X_m) \rangle$

10 calculate $conf(X_1 \rightarrow Z)$ according to Eq. (2)

11 $R = \cup X_1 \rightarrow Z \mid conf(X_1 \rightarrow Z) \geq minconf$

12 **return** R

The first step of the method (see Algorithm 1, lines 1 to 4) produces sequences from time series using a sliding window approach that divides the time series in sequences. Each sequence has the sequence of itemsets occurred in a window of duration d and overlap v . This step has low computational cost because the time series is scanned once.

The second step (see Algorithm 1, lines 6 to 7) consists in determining the frequent sequences. This step is the most computational expensive one, since it generates the candidate sequences and determines the support of them. In this step, we employed the pattern-growth *PrefixSpan* [5] approach with an adaptation in the support counting. Our approach takes into account the overlap produced in the process of mapping time series in sequences (Step 1 of SART, see Algorithm 1, lines 1 to 4) and employs a different way of support counting expressed in Equation 1.

The third step of SART (see Algorithm 1, lines 8 to 12) consists in generating association rules from the frequent itemset sequences found in step 2 and determining the confidence of them (see Equation 2). The method output the set of all strong rules found, i.e. the ones that satisfy the minimum confidence threshold.

The method SART broadens the previous sequence pattern mine methods since it:

- produces rules, while the previous ones produce only frequent sequences;
- add the semantic information of confidence, which was not previously defined by sequential patterns;
- adapts the support counting to the mining of time series using a sequential approach.

Several experiments were conducted in order to evaluate the applicability of the proposed method. Three of the most representative ones are described in the next section.

4 Experiments

We implemented the method SART in Java. The experiments were taken on a Pentium Core i7, 2.8 GHZ computer with 8GB of RAM and a SATA hard disk. We performed the experiments employing agrometeorological time series. Prior to applying the SART mining method, the time series values, which are continuous, were discretized. The discretization process was performed using the Omega algorithm [3]. Omega is a supervised discretization algorithm, which creates data intervals avoiding data inconsistencies and having linear computational cost.

We tested three configurations of sequential mining of the climate time series. In all the configurations, the input data were previously discretized by Omega. The configurations are detailed as follows:

- *Configuration 1* - PrefixSpan: the sequential association rule mining algorithm PrefixSpan was applied over the discretized dataset;
- *Configuration 2* - PrefixSpan ($v = V_1, d = V_2$): the sequential association rule mining algorithm PrefixSpan was applied over the discretized dataset previously processed to generate sequences. The sequence generation was performed using the Step 1 of SART (see Algorithm 1, lines 1- 4). The values $v = V_1$ and $d = V_2$ are the input of overlap size and window size of the sequence generation process;
- *Configuration 3* - SART ($v = V_1, d = V_2$): the method SART (see Algorithm) was entirely applied to mine the discretized dataset. The values $v = V_1$ and $d = V_2$ are the input of overlap size and window size of the method.

4.1 Experiment 1

In this experiment, we employed the Araraquara dataset. The dataset Araraquara was collected from the Brazilian System of Agrometeorological Monitoring (Agritempo - <http://www.agritempo.gov.br/>). It contains agrometeorological data monthly collected from Araraquara, a Brazilian city in the São Paulo State, comprising values of precipitation, maximum temperature (Tmax), minimum temperature (Tmin), NDVI (*Normalized Difference Vegetation Index*) and WRSI (*Water Requirement Satisfaction Index*).

Table 1. Example of data from Araraquara dataset

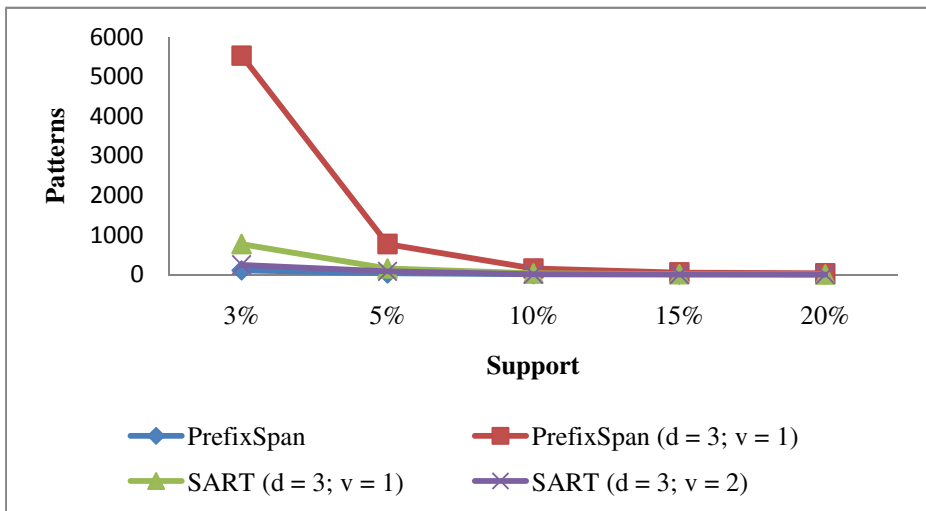
Precipitation	Tmax	Tmin	NDVI	WRIS
30.50	29.02	16.95	0.57	0.63
08.90	24.30	12.70	0.53	0.57
47.50	25.70	13.96	0.43	0.39

The Araraquara dataset was submitted to Omega for discretization. Table 2 shows the data from Table 1 discretized by Omega.

Table 2. Meteorological data discretized by Omega algorithm

Precipitation	Tmax	Tmin	NDVI	WRIS
1[21.89-33.00[2[29.02-29.14[3[16.93-17.28[4[0.50-0.63[5[0.63-0.66[
1[08.89-10.19[2[23.64-24.76[3[12.69-13.02[4[0.50-0.63[5[0.56-0.62[
1[46.59-49.50[2[25.60-26.89[3[13.85-13.89[4[0.30-0.50[5[0.25-0.56[

We submitted the Araraquara dataset to the mining of sequential patterns according to the three configurations described in the beginning of the section. We varied the minimum support (*minsup*) input to the values 3%, 5%, 10%, 15% and 20% and compared the patterns and the number of patterns mined using the traditional sequential pattern mining (Configurations 1 and 2) and the SART method. In order to have a fair comparison, we employed the minimum confidence (*minconf*) input as zero, since the configuration 1 and 2 does not employ the minimum confidence values to mine rules. Figure 1 shows the graph of number of mined patterns.

**Fig. 1.** Number of patterns mined for the Araraquara dataset varying the support

In Configuration 1, the original time series was submitted to PrefixSpan. Employing $minsup=3\%$, 106 patterns were generated. Increasing $minsup$ to 5%, 10%, 15% and 20%, the respective values of 32, 11, 7 and 3 patterns were generated.

In Configuration 2, PrefixSpan works with the sequence generation process. The new sequences generated leads to increase the number of sequences and the total number of tuples, thus, increasing the number of patterns generated. Employing $minsup=3\%$, 5,535 patterns were generated. Increasing $minsup$ to 5%, 10%, 15% and 20%, the respective values of 774, 155, 54 and 31 patterns were generated.

In Configuration 3, SART was applied using different settings for the parameter ν (overlap size). Given parameters ($d = 3$; $\nu = 1$), employing $minsup=3\%$, 774 patterns were generated. Increasing $minsup$ to 5%, 10%, 15% and 20%, the respective values of 774, 155, 54 and 31 patterns were generated. It was obtained a larger number of patterns in relation to the Configuration 1, but less than the Configuration 2, where the parameters were applied with the traditional measure of support and many patterns became accepted because of the reduction in the number of tuples resulted by the sequence generation process.

To examine the influence of overlap on the number of rules, this parameter has been set to $\nu=2$ on the SART method. Employing $minsup=3\%$, 247 patterns were generated. Increasing $minsup$ to 5%, 10%, 15% and 20%, the respective values of 83, 12, 5 and 2 patterns were generated. The results show a reduction in the number of patterns mined with the increase in the overlap size. This occurs because of the support calculation (see Equation 1) that compensates the tuple replication generated by the increase in the overlap size.

An example of a mined pattern using Configuration 1 (PrefixSpan over the discretized dataset) is:

$$Tmax[25.60-26.89[\text{NDVI}[0.30-0.50[, \text{support}=15\%,$$

This pattern means that when the maximum temperature is between 25.60°C and 26.89°C, NDVI is between 0.30 and 0.50, and these items occur in 15% of the dataset. Employing the Configuration 3 (SART method over the discretized dataset previously processed to generate sequences), using $d=3$ and $\nu=1$, the same pattern is generated with the support = 18%, obtaining a small difference in the value. This indicates that the sequence generation and support counting processes performed by SART do not significantly change the values of itemset counting regarding tuple occurrences.

An example of pattern extracted in Configuration 2 is follows:

$$(WRSI[0.99-1.00[), (NDVI[0.50-0.63[), (NDVI[0.50-0.63[), \text{support}=22\%$$

This second pattern means that the frequent sequence of values $WRSI = [0.99 -1.00[$, $NDVI = [0.50- 0.63[$ and $NDVI = [0.50-0.63[$ occurs in 22% of the dataset. Generally, Configuration 2 presents higher support values for itemsets because of the reduction in the number of tuples in the database (106 to 41) after applying the first step of SART.

After entirely performing SART (Configuration 3, using $d=3$ and $v=1$), the following rule was mined:

$(Prec[0.00-1.10[Tmax[25.60-26.89[) \rightarrow (NDVI[0.30-0.50[WRSI[0.25-0.56[), (NDVI[0.30-0.50[), support = 4\%, confidence = 57\%$

This rule means that:

"**If** the Precipitation is between 0.00 and 1.10mm and the maximum temperature is between 25.6 and 26.8 °C, **then**, a month later, NDVI is between 0.30 and 0.50 and WRSI is between 0.25 and 0.56, and, a month later, the NDVI value remains the same".

Observe that SART (Configuration 3) output rules, as an expression of trigger itemset and consequence itemset sequence, and not just frequent itemsets, as the output of the Configurations 1 and 2.

4.2 Experiment 2

In this experiment, we employed the Piracicaba dataset. This dataset was collected from Embrapa, which is a public research institution under the Ministry of Agriculture, Livestock and Supply. The dataset Piracicaba has three attributes, being each one the average value of: monthly minimum temperature (*tmin*), monthly maximum temperature (*tmax*) and monthly rain (*precipitation*). The dataset contains the values of the attributes measured for a period of 48 years for Piracicaba, a Brazilian region in the State of São Paulo.

This experiment was subjected to Configurations 1, 2 and 3 described in the beginning of this section, applying different set of parameters of windows size and overlap ($d = 12$; $v = 11$), whereas finding patterns occurring in a period of 1 year.

Figure 2 shows the number of patterns generated by PrefixSpan (Configuration 1) and the SART method (Configuration 3) for different input values. PrefixSpan (Configuration 1) mined 11 patterns setting minimum support $minsup = 3\%$. Two patterns with two items stand out:

- $(Tmin[14.94-15.46[Prec[0.52-6.63[), support = 4\%$;
- $(Tmin[17.06-17.43 [Prec[0.52-6.63 [), support = 4\%$.

These patterns indicate that the occurrence of minimum temperature values between 14.94 and 17.43 are associated with rainfall between 0.52 and 6.63mm.

Applying Configuration 2 to the experiment, we obtained 190 patterns with support = 3%. As previous experiment, the number of patterns is increased by the reduction of the total number of tuples caused by running the first step of SART.

Running Configuration 3 (SART), we set the values of window size and overlap to cover 12 months, but in this case, due to the small variation in precipitation values along the year, many of the rules that were generated had associated rainfall over the months. For example, the rule " $(Prec[0.52-6.63[) \rightarrow (Prec[0.52-6.63[), \dots, (Prec[0.52-6.63[), over nine months of the year, with support = 4\%$ and confidence = 50%" was mined by SART ($d = 12, v = 11$). The rule means "**if** the Precipitation is between 0.52 and 6.63 mm **then**, for the next 8 months, the Precipitation

value remains in the same interval". This is an example of rule obtained because of the coarse granularity of data interval generated by the discretization process.

The number of patterns mined by SART, as show in Figure 2, increases with the increase in the window size d . This occurs because of increasing d , we enlarge the maximum size of the itemset sequences that are mined by the method. Employing $d=12$ means that the method analyze 12 months of consecutive events to mine patterns.

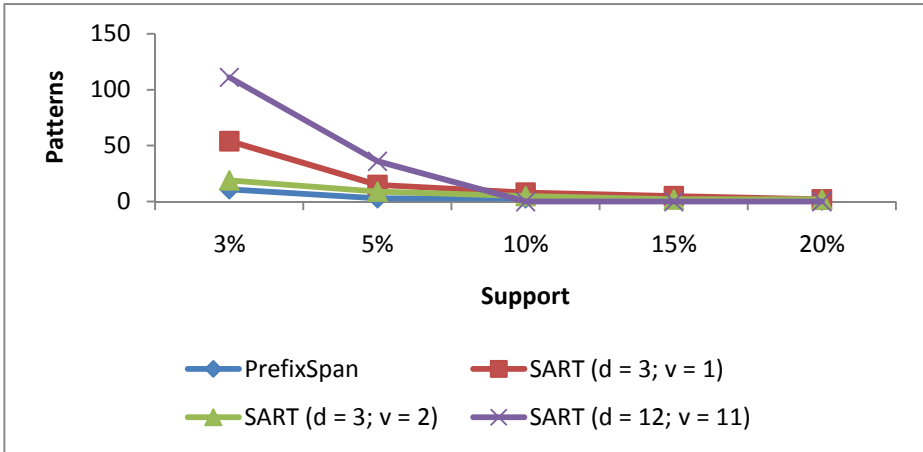


Fig. 2. Number of patterns mined for the Piracicaba dataset varying the support

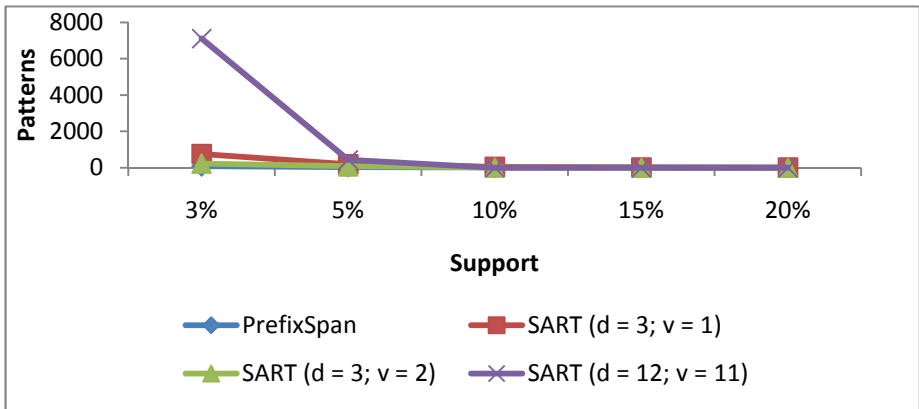
4.3 Experiment 3

Piracicaba is a city of São Paulo State, Brazil, which is an important sugar-canner producer. In this experiment, we employed the dataset Piracicaba-Productivity, provided by Cepagri (Center for Weather and Climate Research Applied to Agriculture) and CTC (Sugarcane Technology Center), which has four attributes, being each one the average value of: monthly minimum temperature ($tmin$), monthly maximum temperature ($tmax$), monthly rain ($precipitation$) and monthly productivity of sugarcane ($productivity$ - tonne per hectare). The dataset contains data from Piracicaba region in the period from 2003 to 2009.

In this experiment, we investigated the number of generated patterns with PrefixSpan, comparing it with the SART Method with the three Configurations described in the beginning of the section. Table 3 shows each setting with the number of mined patterns. Observe (see Table 3) that SART (Configuration 3) reduces the number of patterns in relation to PrefixSpan (Configuration 2) with $d = 3$ and $v = 1$,

Table 3. Number of patterns generated for Configurations 1, 2 and 3 by applying different values of support over dataset Piracicaba-Productivity

Method	3%	5%	10%	15%	20%
PrefixSpan	79	36	12	7	4
PrefixSpan ($d = 3; v = 1$)	2718	744	192	79	42
SART ($d = 3; v = 1$)	748	192	42	13	6
SART ($d = 3; v = 2$)	228	77	17	4	2
SART ($d = 12; v = 11$)	7109	431	0	0	0

**Fig. 3.** Number of patterns generated for Configurations 1, 2 and 3 by applying different values of support over dataset Piracicaba-Productivity

In this dataset, using the SART method it is possible to identify association rules with higher confidence. The following rules were generated by setting the parameters window size $d = 3$ and overlap size $v = 1$.

- a. $(Tmin[18.23-18.51[\text{Prec}[0.58-6.76[) \rightarrow (Tmax[29.61-29.87[Prod[85.58-87.26[)$,
support= 2%, confidence=100%, meaning:

"If in a month the minimum temperature is between 18.23 and 18.51°C and the precipitation is between 0.58 and 6.76 mm, **then**, after at most 2 months, the maximum temperature is between 0.58 and 6.76°C and the sugarcane production is between 85.58-87.26 tonne per hectare".

- b. $(Tmin[19.54-19.67[\text{Prec}[0.58-6.76[) \rightarrow (Prec[0.58-6.76[Prod[0.00-80.59[)$,
support= 2%, conf=100%

"If in a month the minimum temperature is between 19.54 and 19.67°C and precipitation is between 0.58 and 6.76 mm, **then**, after at most 2 months, the precipitation is between 0.58 and 6.76 mm and the cane production is between 0.00 and 80.59 ton".

Comparing rules (a) and (b) generated from this data set, it is possible to notice that an increase of about 1 degree in the minimum temperature produces a decrease of productivity in the next months.

The experiments shows that method SART produces rules, while the previous ones produce only frequent sequences, adding the semantic information of confidence, which was not previously defined by sequential patterns. This important because a rule has an important relation of cause and consequence that a frequent pattern does not bring. Also, the method adapts the support counting to the mining of time series and adds the time information about the related events that occur in a rule.

5 Conclusions and Further Research

In this paper we proposed a new method called SART for mining sequential patterns in time series using a sliding window approach. Our proposed method SART broaden the previous sequence pattern mine methods, producing rules, while the previous ones produce only frequent sequences. The method adapts the measure of support to work with time series using a sequential approach. Also, the method defines the confidence measure (the *trigger-confidence*) for a sequential pattern. The experiments showed that the method mines a larger number of patterns when compared with the tradition association rule mining, revealing associations regarding itemset occurrences over the time. Also, the mined association rules and their confidence values bring more semantic information, than the frequent itemset sequences mined by the previous methods. In addition, the time information between the rule itemsets broadens the applicability and semantic information of the mined patterns.

For future research, we consider conducting larger series of real world data and developing a visual data mining tool to present the rules in a visual grid to obtain an overview of their distribution through the measures of support and confidence, introducing useful filter tools to increase the cognition gain during the mined patterns analysis.

Acknowledgments. This work has been supported by FAPESP (Sao Paulo State Research Foundation), CNPq (National Council for Scientific and Technological Development), CAPES (Brazilian Federal Funding Agency for Graduate Education Improvement) and Microsoft-Research.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: 4th Int. CFDOA, Chicago, IL, pp. 69–84 (1993)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Yu, P.S., Chen, A.S.P. (eds.) Proceedings of the 11th International Conference on Data Engineering (ICDE 1995), pp. 3–14. IEEE Press, Taipei (1995)

3. Ribeiro, M.X., Traina, A.J.M., Traina, J.C.: A new algorithm for data discretization and feature selection. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 953–954. ACM, Fortaleza (2008)
4. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: ICEDT, Avignon, France, pp. 3–17. Springer (1996)
5. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of the 17th International Conference on Data Engineering, pp. 215–224. IEEE Computer Society, Washington, DC (2001)
6. Lu, H., Feng, L., Han, J.: Beyond intratransaction association analysis: mining multidimensional intertransaction association rules. *ACM Trans. Inf. Syst.* 18, 423–454 (2000)
7. Romani, L.A.S., et al.: Clearminer: a new algorithm for mining association patterns on heterogeneous time series from climate data. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 900–905. ACM, New York (2010)
8. Subramanyam, R.B.V., Goswami, A.: A fuzzy data mining algorithm for incremental mining of quantitative sequential patterns. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 13, 633–652 (2005)
9. Zaki, M.J.: Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.* 42, 31–60 (2001)
10. Park, J.S., Chen, M.-S., Yu, P.S.: An effective hash-based algorithm for mining association rules. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pp. 175–186. ACM, New York (1995)
11. Tung, A.K., Angelis, L., Vlahavas, I.: Breaking the barrier of transactions: mining intertransaction association rules. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 297–301. ACM, New York (1999)
12. Feng, L., Yu, X.J., Lu, H., Han, J.: A template model for multidimensional intertransactional association rules. *The VLDB Journal* 11, 153–175 (2002)
13. Hu, Y., Huang, T.C., Yang, H., Chen, Y.: On mining multi-time-interval sequential patterns. *Knowledge Engineering* 68(10), 1112–1127 (2009)
14. Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., Hsu, M.-C.: Freespan: Frequent pattern-projected sequential pattern mining. In: Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD 2000), Boston, MA, pp. 355–359 (2000)
15. Saputra, D., Dayang, R.A.R., Foong, O.M.: Mining sequential patterns using I-prefixSpan. *International Journal of Computer Science and Engineering* 2, 14–16 (2008)
16. Berberidis C., Angelis L., Vlahavas I.: Inter-transaction Association Rules Mining for Rare Events Prediction. In: Proc. (companion volume) 3rd Hellenic Conference on Artificial Intelligence (SETN 2004), Samos, Greece (2004)
17. Zhao, Q., Bhowmick, S.S.: Sequential Pattern Matching: A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore (2003)
18. Lee, A.J.T., Wang, C.-S.: An efficient algorithm for mining frequent inter-transaction patterns. *Inf. Sci.* 177(17), 3453–3476 (2007)

Author Index

- Abdullah, Noryusliza IV-364
Abdullah, Nurul Azma IV-353
Abid, Hassan III-368
Adewumi, Adewole IV-248
Afonso, Vitor Monte IV-274, IV-302
Agarwal, Suneeta IV-147
Aguiar, Rui L. III-682
Aguilar, José Alfonso IV-116
Ahmadian, Kushan I-188
Alarcon, Vladimir J. II-578, II-589
Albuquerque, Caroline Oliveira III-576
Alfaro, Pablo IV-530
Ali, Salman III-352
Alizadeh, Hosein III-647
Almeida, Regina III-17
Alonso, Pedro I-29
Alves, Daniel S.F. I-101
Amandi, Analía A. III-698, III-730
Amaral, Paula III-159
Amjad, Jaweria III-368
Ammar, Reda A. I-161
Amorim, Elisa P. dos Santos I-635
An, Deukhyeon III-272
Anderson, Roger W. I-723
Angeloni, Marcus A. I-240
Antonino, Pablo Oliveira III-576
Aquilanti, Vincenzo I-723
Arefin, Ahmed Shamsul I-71
Armentano, Marcelo G. III-730
Arnaout, Arghad IV-392
Aromando, Angelo III-481
Asche, Hartmut II-347, II-386, II-414, II-439
Aydin, Ali Orhan IV-186
Ayres, Rodrigo Moura Juvenil III-667
Azad, Md. Abul Kalam III-72
Azzato, Antonello II-686
- Bae, Sunwook III-238
Balena, Pasquale I-583, II-116
Balucani, Nadia I-331
Barbosa, Ciro I-707
Barbosa, Fernando Pires IV-404
Barbosa, Helio J.C. I-125
- Baresi, Umberto II-331
Barreto, Marcos I-29
Baruque, Alexandre Or Cansian IV-302
Bastianini, Riccardo I-358
Batista, Augusto Herrmann III-631
Batista, Vitor A. IV-51
Battino, Silvia II-624
Beaver, Justin IV-646
Bencardino, Massimiliano II-548
Berdún, Luis III-698
Berenguel, José L. III-119
Bernardino, Heder S. I-125
Berretta, Regina I-71
Biehl, Matthias IV-40
Bimonte, Sandro II-373
Bisceglie, Roberto II-331
Bitencourt, Ana Carla P. I-723
Blecic, Ivan II-481, II-492
Boavida, Fernando II-234
Bollini, Letizia II-508
Boratto, Murilo I-29
Borg, Erik II-347, III-457
Borruso, Giuseppe II-624, II-670
Boulil, Kamal II-373
Braga, Ana Cristina I-665
Brumana, Raffaella II-397
Bugs, Geisa I-477
Burgarelli, Denise I-649
Bustos, Víctor III-607
- Caiaffa, Emanuela II-532
Calazan, Rogério M. I-148
Caldas, Daniel Mendes I-675
Callejo, Miguel-Ángel Manso I-462
Camarda, Domenico II-425
Campobasso, Francesco II-71
Campos, Ricardo Silva I-635
Candori, Pietro I-316, I-432
Cannatella, Daniele II-54
Cano, Marcos Daniel III-743
Cansian, Adriano Mauro IV-286
Carbonara, Sebastiano II-128
Cardoso, João M.P. IV-217
Carmo, Rafael IV-444

- Carneiro, Joubert C. Lima e Tiago G.S. II-302
- Carpené, Michele I-345
- Carvalho, Luis Paulo da Silva II-181
- Carvalho, Maria Sameiro III-30, III-187
- Casado, Leocadio G. III-119, III-159
- Casado, Leocadio Gonzalez I-57
- Casas, Giuseppe B. Las II-466, II-640, II-686
- Casavecchia, Piergiorgio I-331
- Castro, Patrícia F. IV-379
- Cavalcante, Gabriel D. IV-314
- Cecchi, Marco I-345
- Cecchini, Arnaldo II-481, II-492
- Ceppi, Claudia II-517
- Cermignani, Matteo I-267
- Cerreta, Maria II-54, II-168, II-653
- Chanet, Jean-Pierre II-373
- Charão, Andrea Schwertner IV-404
- Cho, Yongyun IV-613, IV-622
- Cho, Young-Hwa IV-543
- Choe, Junseong III-324
- Choi, Hong Jun IV-602
- Choi, Jae-Young IV-543
- Choi, Jongsun IV-613
- Choi, Joonsoo I-214
- Choo, Hyunseung III-259, III-283, III-324
- Chung, Tai-Myoung III-376
- Ci, Song III-297
- Cicerone, Serafino I-267
- Ciloglugil, Birol III-550
- Cioquetta, Daniel Souza IV-16
- Clarke, W.A. IV-157
- Coelho, Leandro I-29
- Coletti, Cecilia I-738
- Conrado, Merley da Silva III-618
- Corea, Federico-Vladimir Gutiérrez I-462
- Coscia, José Luis Ordiales IV-29
- Costa, M. Fernanda P. III-57, III-103
- Costantini, Alessandro I-345, I-401, I-417
- Crasso, Marco IV-29, IV-234, IV-484
- Crawford, Broderick III-607
- Crocchianti, Stefano I-417
- Cuca, Branka II-397
- Cui, Xiaohui IV-646
- Cunha, Jácome IV-202
- da Luz Rodrigues, Francisco Carlos III-657
- Danese, Maria III-512
- Dantas, Sócrates de Oliveira I-228
- da Silva, Paulo Caetano II-181
- Daskalakis, Vangelis I-304
- de Almeida, Ricardo Aparecido Perez IV-470, IV-560
- de Avila, Ana Maria H. III-743
- de By, Rolf A. II-286
- de Carvalho, Andre Carlos P.L.F. III-562
- de Carvalho Jr., Osmar Abílio III-657
- Decker, Hendrik IV-170
- de Costa, Evandro Barros III-714
- de Deus, Raquel Faria II-565
- de Felice, Annunziata II-1
- de Geus, Paulo Lício IV-274, IV-302, IV-314
- Delgado del Hoyo, Francisco Javier I-529
- Dell'Orco, Mauro II-44
- de Macedo Mourelle, Luiza I-101, I-113, I-136, I-148
- de Magalhães, Jonathas José III-714
- De Mare, Gianluigi II-27
- Dembogurski, Renan I-228
- de Mendonça, Rafael Mathias I-136
- de Miranda, Péricles B.C. III-562
- de Oliveira, Isabela Liane IV-286
- de Oliveira, Wellington Moreira I-561
- de Paiva Oliveira, Alcione I-561
- Deris, Mustafa Mat I-87, IV-340
- De Santis, Fortunato III-481
- Désidéri, Jean-Antoine IV-418
- de Souza, Cleyton Caetano III-714
- de Souza, Éder Martins III-657
- de Souza, Renato Cesar Ferreira I-502
- de Souza Filho, José Luiz Ribeiro I-228, II-712
- De Toro, Pasquale II-168
- Dias, Joana M. III-1
- Dias, Luis III-133
- do Nascimento, Gleison S. IV-67
- Donato, Carlo II-624
- do Prado, Hércules Antonio III-631, III-657
- dos Anjos, Eudisley Gomes IV-132
- dos Santos, Jefersson Alex I-620

- dos Santos, Rafael Duarte Coelho IV-274, IV-302
- dos Santos, Rodrigo Weber I-635, I-649, I-691, I-707
- dos Santos Soares, Michel IV-1, IV-16
- e Alvelos, Filipe Pereira III-30
- El-Attar, Mohamed IV-258
- Elish, Mahmoud O. IV-258
- El-Zawawy, Mohamed A. III-592, IV-83
- Engemaier, Rita II-414
- Eom, Young Ik III-227, III-238, III-272
- Epicoco, Italo I-44
- Esmael, Bilal IV-392
- Ezzatti, Pablo IV-530
- Falcinelli, Stefano I-316, I-331, I-387, I-432
- Falcone, Roberto II-508
- Fanizzi, Annarita II-71
- Farage, Michèle Cristina Resende I-675
- Farantos, C. Stavros I-304
- Farias, Matheus IV-444
- Fechine, Joseana Macêdo III-714
- Fedel, Gabriel de S. I-620
- Felino, Antônio I-665
- Fernandes, Edite M.G.P. III-57, III-72, III-103
- Fernandes, Florbela P. III-103
- Fernandes, João P. IV-202, IV-217
- Ferneda, Edilson III-631, III-657
- Ferreira, Ana C.M. III-147
- Ferreira, Brigida C. III-1
- Ferreira, Manuel III-174
- Ferroni, Michele I-358
- Fichtelmann, Bernd II-347, III-457
- Fidêncio, Érika II-302
- Figueiredo, José III-133
- Filho, Dario Simões Fernandes IV-274, IV-302
- Filho, Jugurta Lisboa I-561
- Fiorese, Adriano II-234
- Fonseca, Leonardo G. I-125
- Formosa, Saviour II-609
- Formosa Pace, Janice II-609
- Fort, Marta I-253
- França, Felipe M.G. I-101
- Freitas, Douglas O. IV-470
- Fruhwrith, Rudolf K. IV-392
- García, I. III-119
- García, Immaculada I-57
- Garrigós, Irene IV-116
- Gasior, Wade IV-646
- Gavrilova, Marina I-188
- Gentili, Eleonora III-539
- Geraldes, Carla A.S. III-187
- Gervasi, Osvaldo IV-457
- Ghandehari, Mehran II-194
- Ghazali, Rozaida I-87
- Ghiselli, Antonia I-345
- Ghizoni, Maria Luísa Amarante IV-588
- Girard, Luigi Fusco II-157
- Gomes, Ruan Delgado IV-132
- Gomes, Tiago Costa III-30
- Gonschorek, Julia II-208, II-220
- Gonzaga de Oliveira, Sanderson Lincohn I-172, I-198, I-610
- Görlich, Markus I-15
- Greco, Ilaria II-548
- Grégio, André Ricardo Abed IV-274, IV-286, IV-302
- Guardia, Hélio C. IV-560
- Gupta, Pankaj III-87
- Hahn, Kwang-Soo I-214
- Haijema, Rene III-45
- Han, Jikwang III-217
- Han, JungHyun III-272
- Han, Yanni III-297
- Handaga, Bana IV-340
- Hasan, Osman III-419
- Hashim, Rathiah II-728
- Hendrix, Eligius M.T. I-57, III-45, III-119, III-159
- Heo, Jaewee I-214
- Hong, Junguye III-324
- Huang, Lucheng I-447
- Ibrahim, Rosziati IV-353, IV-364
- Igounet, Pablo IV-530
- Ikhu-Omoregbe, Nicholas IV-248
- Im, Illkyun IV-543
- Imtiaz, Sahar III-339
- Inceoglu, Mustafa Murat III-550
- Iochepe, Cirano IV-67
- Ipbuker, Cengizhan III-471
- Ivánová, Ivana II-286
- Izkara, Jose Luis I-529

- Jeon, Jae Wook III-311
 Jeon, Woongryul III-391
 Jeong, Jongpil IV-543
 Jeong, Soonmook III-311
 Jino, Mario IV-274, IV-302
 Jorge, Eduardo IV-444
 Jung, Sung-Min III-376
- Kalsing, André C. IV-67
 Kang, Min-Jae III-217
 Karimipour, Farid II-194
 Kasprzak, Andrzej IV-514, IV-576
 Kaya, Sinasi III-471
 Khalid, Noor Elaiza Abdul II-728
 Khan, Salman H. III-339
 Khan, Yasser A. IV-258
 Khanh Ha, Nguyen Phan III-324
 Kim, Cheol Hong IV-602
 Kim, Hakhyun III-391
 Kim, Iksu IV-622
 Kim, Jeehong III-227, III-238, III-272
 Kim, Junho I-214
 Kim, Young-Hyuk III-248
 Kischinhevsky, Mauricio I-610
 Kluge, Mario II-386
 Knop, Igor I-707
 Komati, Karin S. II-739
 Kopeliovich, Sergey I-280
 Kosowski, Michał IV-514
 Koszalka, Leszek IV-576
 Koyuncu, Murat IV-234
 Kwak, Ho-Young III-217
 Kwon, Keyho III-311
 Kwon, Ki-Ryong IV-434
 Kwon, Seong-Geun IV-434
- Ladeira, Pitter Reis II-548
 Laganà, Antonio I-292, I-345, I-358,
 I-371, I-387, I-401, I-417
 Lago, Noelia Faginas I-387
 Laguna Gutiérrez, Víctor Antonio
 III-618
 Lanorte, Antonio III-481, III-512
 Lanza, Viviana II-686
 Lasaponara, Rosa III-481, III-497,
 III-512
 Le, Duc Tai III-259
 Lederer, Daniel II-263
 Le Duc, Thang III-259
 Lee, Dong-Young III-368, III-376
- Lee, Eung-Joo IV-434
 Lee, Hsien-Hsin IV-602
 Lee, Jae-Gwang III-248
 Lee, Jae-Kwang III-248
 Lee, Jae-Pil III-248
 Lee, Jongchan IV-613
 Lee, Junghoon III-217
 Lee, Kwangwoo III-391
 Lee, Sang Joon III-217
 Lee, Suk-Hwan IV-434
 Lee, Yunho III-391
 Leonel, Gildo de Almeida II-712
 Leonori, Francesca I-331
 Li, Yang III-297
 Lim, Il-Kwon III-248
 Lima, Priscila M.V. I-101
 Lin, Tao III-297
 Liu, Yi IV-100
 Lobarinhas, Pedro III-202
 Lobosco, Marcelo I-675, I-691, I-707
 Loconte, Pierangela II-517
 Lomba, Ricardo III-202
 Lombardi, Andrea I-387
 Lopes, Maria do Carmo III-1
 Lopes, Paulo IV-217
 Lou, Yan I-447
 Lubisco, Giorgia II-517
 Luiz, Alfredo José Barreto III-657
- Ma, Zhiyi IV-100
 Macedo, Gilson C. I-691, I-707
 Maffioletti, Sergio I-401
 Maleki, Behzad III-647
 Mancini, Francesco II-517
 Mangialardi, Giovanna II-116
 Manuali, Carlo I-345
 Marcondes, Cesar A.C. IV-470
 Marghany, Maged III-435, III-447
 Marimbaldo, Francisco-Javier Moreno
 I-462
 Marinho, Euler Horta IV-632
 Martins, Luís B. III-147
 Martins, Pedro IV-217
 Martucci, Isabella II-1
 Marucci, Alessandro II-532
 Marwala, T. IV-157
 Marwedel, Peter I-15
 Marzuoli, Annalisa I-723
 Mashkoor, Atif III-419
 Masini, Nicola III-497

- Mateos, Cristian IV-29, IV-234, IV-484
 Mazhar, Aliya III-368
 Mazón, Jose-Norberto IV-116
 McAnally, William H. II-578, II-589
 Medeiros, Claudia Bauzer I-620
 Mehlawat, Mukesh Kumar III-87
 Meira Jr., Wagner I-649
 Mele, Roberta II-653
 Melo, Tarick II-302
 Messine, F. III-119
 Miao, Hong I-447
 Milani, Alfredo III-528, III-539
 Min, Changwoo III-227, III-238
 Min, Jae-Won III-376
 Misra, A.K. IV-157
 Misra, Sanjay IV-29, IV-147, IV-234,
 IV-248
 Miziolek, Marek IV-514
 Mocavero, Silvia I-44
 Mohamad, Kamaruddin Malik IV-353
 Mohamed Elsayed, Samir A. I-161
 Monfroy, Eric III-607
 Monteserin, Ariel III-698
 Montrone, Silvestro II-102
 Moreira, Adriano II-450
 Moreira, Álvaro IV-67
 Moscato, Pablo I-71
 Moschetto, Danilo A. IV-470
 Müller, Heinrich I-15
 Mundim, Kleber Carlos I-432
 Mundim, Maria Suelly Pedrosa I-316
 Mundim, Maria Suely Pedrosa I-432
 Munir, Ali III-352, III-368
 Murgante, Beniamino II-640, II-670,
 III-512
 Murri, Riccardo I-401
 Musaoglu, Nebiye III-471

 Nabwey, Hossam A. II-316, II-358
 Nakagawa, Elisa Yumi III-576
 Nalli, Danilo I-292
 Nawi, Nazri Mohd I-87
 Nedjah, Nadia I-101, I-113, I-136, I-148
 Nema, Jigyasu III-528
 Neri, Igor IV-457
 Nesticò, Antonio II-27
 Neves, Brayan II-302
 Nguyên, Toàn IV-418
 Niu, Wenjia III-297
 Niyogi, Rajdeep III-528

 Nolè, Gabriele III-512
 Nunes, Manuel L. III-147

 O'Kelly, Morton E. II-249
 Oliveira, José A. III-133
 Oliveira, Rafael S. I-649
 Oreni, Daniela II-397
 Ottomanelli, Michele II-44

 Pacifici, Leonardo I-292, I-371
 Pádua, Clarindo Isaías P.S. IV-51
 Pádua, Wilson IV-51
 Pallottelli, Simonetta I-358
 Panaro, Simona II-54
 Pandey, Kusum Lata IV-147
 Paolillo, Pier Luigi II-331
 Park, Changyong III-283
 Park, Gyung-Leen III-217
 Park, Junbeom III-283
 Park, Sangjoon IV-613
 Park, Young Jin IV-602
 Parvin, Hamid III-647
 Parvin, Sajad III-647
 Pathak, Surendra II-589
 Pauls-Worm, Karin G.J. III-45
 Peixoto, Daniela C.C. IV-51
 Peixoto, João II-450
 Pepe, Monica II-397
 Perchinunno, Paola II-88, II-102
 Pereira, Gilberto Corso I-491
 Pereira, Guilherme A.B. III-133, III-187
 Pereira, Óscar Mortágua III-682
 Pereira, Tiago F. I-240
 Pessanha, Fábio Gonçalves I-113
 Pigozzo, Alexandre B. I-691, I-707
 Pimentel, Dulce II-565
 Pingali, Keshav I-1
 Pinheiro, Marcello Sandi III-631
 Pirani, Fernando I-316, I-387, I-432
 Piscitelli, Claudia II-517
 Poggioni, Valentina III-539
 Pol, Maciej IV-576
 Pollino, Maurizio II-532
 Poma, Lourdes P.P. IV-470
 Pontrandolfi, Piergiuseppe II-686
 Poplin, Alenka I-491
 Pozniak-Koszalka, Iwona IV-576
 Pradel, Marilys II-373
 Prasad, Rajesh IV-147
 Prieto, Iñaki I-529

- Proma, Wojciech IV-576
 Prudêncio, Ricardo B.C. III-562
- Qadir, Junaid III-352
 Qaisar, Saad Bin III-339, III-352,
 III-407
 Quintela, Bárbara de Melo I-675, I-691,
 I-707
- Ragni, Mirco I-723
 Raja, Haroon III-368, III-407
 Rajasekaran, Sanguthevar I-161
 Rak, Jacek IV-498
 Ramiro, Carla I-29
 Rampini, Anna II-397
 Rasekh, Abolfazl II-275
 Re, Nazzareno I-738
 Renhe, Marcelo Caniato II-712
 Resende, Rodolfo Ferreira IV-632
 Rezende, José Francisco V. II-302
 Rezende, Solange Oliveira III-618
 Ribeiro, Hugo IV-202
 Ribeiro, Marcela Xavier III-667, III-743
 Riveros, Carlos I-71
 Rocha, Ana Maria A.C. III-57, III-72,
 III-147
 Rocha, Bernardo M. I-649
 Rocha, Humberto III-1
 Rocha, Jorge Gustavo I-571
 Rocha, Maria Célia Furtado I-491
 Rocha, Pedro Augusto F. I-691
 Rodrigues, António M. II-565
 Romani, Luciana A.S. III-743
 Rosi, Marzio I-316, I-331
 Rossi, Elda I-345
 Rossi, Roberto III-45
 Rotondo, Francesco I-545
 Ruiz, Linnyer Beatrys IV-588
- Sad, Dhiego Oliveira I-228
 Salles, Evandro O.T. II-739
 Salles, Ronaldo M. IV-326
 Salvatierra, Gonzalo IV-484
 Sampaio-Fernandes, João C. I-665
 Samsudin, Nurnabilah II-728
 Sanches, Silvio Ricardo Rodrigues
 II-699
 Sanjuan-Estrada, Juan Francisco I-57
 Santos, Aduino IV-588
 Santos, Maribel Yasmina III-682
 Santos, Marilde Terezinha Prado
 III-667, III-743
 Santos, Teresa II-565
 Sarafian, Haiduke I-599
 Saraiva, João IV-202, IV-217
 Sarcinelli-Filho, Mario II-739
 Schirone, Dario Antonio II-1, II-17,
 II-88
 Sciberras, Elaine II-609
 Scorza, Francesco II-640
 Selicato, Francesco I-545, II-517
 Selicato, Marco II-144
 Sellarès, J. Antoni I-253
 Sertel, Elif III-471
 Shaaban, Shaaban M. II-316, II-358
 Shah, Habib I-87
 Shukla, Mukul IV-157
 Shukla, Ruchi IV-157
 Silva, António I-571
 Silva, João Tácio C. II-302
 Silva, José Eduardo C. I-240
 Silva, Rodrigo I-228
 Silva, Rodrigo M.P. IV-326
 Silva, Roger Correia I-228
 Silva, Valdinei Freire da II-699
 Silva Jr., Luneque I-113
 Silvestre, Eduardo Augusto IV-1
 Simões, Flávio O. I-240
 Simões, Paulo II-234
 Singh, Gaurav II-286
 Skouteris, Dimitris I-331
 Soares, Carlos III-562
 Song, Hokwon III-227, III-238
 Song, Taehoun III-311
 Soravia, Marco II-599
 Soto, Ricardo III-607
 Souza, Cleber P. IV-314
 Stanganelli, Marialuce II-599
 Stankute, Silvija II-439
- Tajani, Francesco II-27
 Tasso, Sergio I-358, IV-457
 Tavares, Maria Purificação I-665
 Teixeira, Ana Paula III-17
 Teixeira, José Carlos III-174, III-202
 Teixeira, Senhorinha F.C.F. III-147,
 III-202
 Tentella, Giorgio I-417
 Thom, Lucinéia IV-67

- Thonhauser, Gerhard IV-392
 Tilio, Lucia II-466, II-686
 Timm, Constantin I-15
 Tori, Romero II-699
 Torkan, Germano II-17
 Torre, Carmelo Maria I-583, II-116,
 II-144, II-157
 Traina, Agma J.M. III-743
 Treadwell, Jim IV-646
 Tricaud, Sebastien IV-314
 Trofa, Giovanni La II-144
 Trunfio, Giuseppe A. II-481, II-492
 Tsoukiàs, Alexis II-466
 Tyrallová, Lucia II-208, II-220
- Usera, Gabriel IV-530
- Vafaiezhad, Ali Reza II-275
 Vargas-Hernández, José G. I-518
 Varotsis, Constantinos I-304
 Vaz, Paula I-665
 Vecchiocattivi, Franco I-316, I-432
 Vella, Flavio IV-457
 Verdicchio, Marco I-371
 Verma, Shilpi III-87
 Versaci, Francesco I-1
 Vieira, Marcelo Bernardes I-228, II-712
- Vieira, Wesley IV-444
 Vyatkina, Kira I-280
- Walkowiak, Krzysztof IV-498, IV-514
 Wang, Hao III-283
 Wang, Kangkang I-447
 Weichert, Frank I-15
 Won, Dongho III-391
 Wu, Feifei I-447
- Xavier, Carolina R. I-635
 Xavier, Micael P. I-691
 Xexéo, Geraldo B. IV-379
 Xu, Yanmei I-447
 Xu, Yuemei III-297
- Yanalak, Mustafa III-471
- Zaldívar, Anibal IV-116
 Zenha-Rela, Mário IV-132
 Zhang, Tian IV-100
 Zhang, Xiaokun IV-100
 Zhang, Yan IV-100
 Zhao, Xuying IV-100
 Zito, Romina I-583
 Zunino, Alejandro IV-29, IV-234,
 IV-484