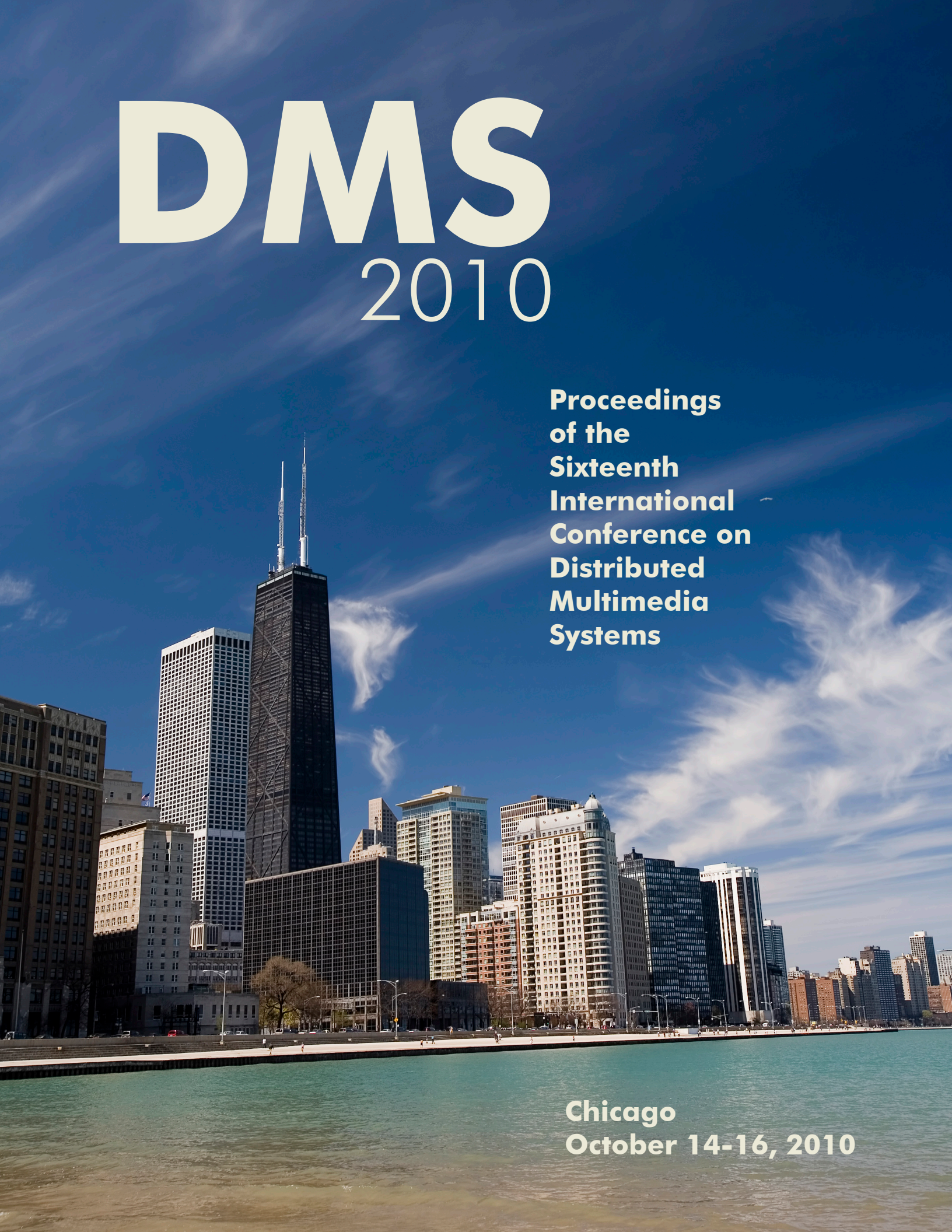
DMS 2010

Proceedings of the Sixteenth International Conference on Distributed Multimedia Systems



Chicago
October 14-16, 2010

PROCEEDINGS

DMS 2010

The 16th International Conference on Distributed Multimedia Systems

Sponsored by

Knowledge Systems Institute Graduate School, USA

Technical Program

October 14 - 16, 2010

Hyatt Lodge at McDonald's Campus, Oak Brook, Illinois, USA

Organized by

Knowledge Systems Institute Graduate School

Copyright © 2010 by Knowledge Systems Institute Graduate School

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written consent of the publisher.

ISBN 1-891706-28-4 (paper)

Additional Copies can be ordered from:
Knowledge Systems Institute Graduate School
3420 Main Street
Skokie, IL 60076, USA
Tel:+1-847-679-3135
Fax:+1-847-679-3166
Email:office@ksi.edu
http://www.ksi.edu

Proceedings preparation, editing and printing are sponsored by
Knowledge Systems Institute Graduate School, USA

Printed by Knowledge Systems Institute Graduate School

DMS2010 Foreword

Welcome to DMS 2010! This conference proceedings document the presentations for the 16th International Conference on Distributed Multimedia Systems (DMS2010).

Multimedia has been growing from strength to strength and continues to expend in a huge number and varied subject areas in the digital world, with ever more exciting and important research and development for both academic and industrial domains. The DMS conference is an international conference series, which covers a wide spectrum of paper presentations, technical discussions and demonstrations in the fields of distributed multimedia computing. The main conference themes have been organized, according to a formula consolidated during past editions, into a number of thematic tracks offering to the conference attendants and to readers a convenient way to explore this vast amount of knowledge in an organized way. Two additional workshops which extended the main conference offerings, and complemented the conference program are: the International Workshop on Distance Education Technologies (DET 2010), and the International Workshop on Visual Languages and Computing (VLC 2010).

This year, the theme of the Conference is "Globalization and Personalization", looking at both ends of the continuum exploring new possibilities and cutting edge state of the arts. Researchers from twenty-six countries submitted papers this year: Austria, Belgium, Brazil, China, Colombia, Germany, India, Iran, Italy, Japan, Jordan, Malaysia, Oman, Pakistan, Portugal, Russian Federation, Saudi Arabia, South Africa, Spain, Sweden, Switzerland, Taiwan, Thailand, United Kingdom, United States and Yugoslavia, offering a truly "distributed" atmosphere and "globalization" nature to the conference.

The sixty papers included in this proceedings are the results of a highly selective review process to pick the most creative work to fit into a very limited available time for the conference.

First and foremost, we would like to thank all the authors who have submitted their works to DMS2010. Thanks for the quality of the submissions and all the innovative results and studies and thank you for choosing DMS as the place to present it. All of the authors who submitted papers, both accepted and regretfully rejected, are responsible for upholding the vitality and cutting edge level of the DMS program. We would like to take this opportunity to acknowledge all the authors!

Thanks to the Program Committee members who have offered enormous blocks of time from their busy schedules to carefully read and evaluate all the submissions. This program could not exist without their contributions.

As program co-chairs, our work was made much easier by following the procedures developed, refined and passed along by the Steering Committee Chair Dr S.K. Chang and fantastic supports from the staff of the Knowledge Systems Institute Graduate School of Computer and Information Sciences, including Dennis Chi, David Huang, Rachel Lu and Gabriel Smith.

We are pleased to include two invited keynotes in the program. Thanks to our invited keynotes: Prof David Hung-Chang Du from the University of Minnesota, Minneapolis and Prof Yi Deng from the University of North Carolina at Charlotte.

We are grateful to many people without whom this Conference would not be possible. Thanks to old and new friends, collaborators, institutions and organisations who have supported the conference.

Last but not least, thanks to all participants of DMS2010. On behalf of the DMS2010 programme committees, we are delighted to welcome you to the 16th International Conference on Distributed Multimedia Systems (DMS2010).

DMS 2010 Program co-Chairs:
Paolo Nesi – DSI – University of Florence, Italy. nesi@dsi.unifi.it, www.dsi.unifi.it/~nesi
Kia Ng, ICSRiM – University of Leeds, UK. kia@kcng.org, www.kcng.org

The 16th International Conference on Distributed Multimedia Systems (DMS 2010)

October 14 - 16, 2010
Hyatt Lodge at McDonald's Campus, Oak Brook, Illinois, USA

Organizers & Committees

Steering Committee Chair

Shi-Kuo Chang, *University of Pittsburgh, USA*

Conference Co-Chairs

Augusto Celentano, *Universita Ca Foscari di Venezia, Italy*
Atsuo Yoshitaka, *JAIST, Japan*

Program Co-Chairs

Paolo Nesi, *University of Florence, Italy*
Kia Ng, *University of Leeds, UK*

Program Committee

Arvind K. Bansal, *Kent State University, USA*
Tolga Capin, Bilkent *University, Turkey*
Augusto Celentano, *Universita Ca Foscari di Venezia, Italy*
Ing-Ray Chen, *Virginia Tech (VPI&SU), USA*
Shu-Ching Chen, *Florida International University, USA*
William Cheng-Chung Chu, *Tunghai University, Taiwan*
Gennaro Costagliola, *Univ of Salerno, Italy*
Alfredo Cuzzocrea, *University of Calabria, Italy*
Andrea De Lucia, *Univ. of Salerno, Italy*
Sergio Di Martino, *Universita degli Studi di Napoli Federico II, Italy*
David H. C. Du, *Univ. of Minnesota, USA*
Larbi Esmahi, *National Research Council of Canada, Canada*

Daniela Fogli, *Universita degli Studi di Brescia, Italy*
Farshad Fotouhi, *Wayne State University, USA*
Kaori Fujinami, *Tokyo University of Agriculture and Technology, Japan*
Moncef Gabbouj, *Tampere University of Technology, Finland*
Ombretta Gaggi, *Univ. of Padova, Italy*
Stefan Goebel, *ZGDV Darmstadt, Germany*
Jivesh Govil, *Cisco Systems Inc., USA*
Angela Guercio, *Kent State University, USA*
Niklas Hallberg, *FOI, Sweden*
Hewijin Christine Jiau, *National Cheng Kung University, Taiwan*
Joemon Jose, *University of Glasgow, UK*
Erland Jungert, *Linkoping University, Sweden*
Egidijus Kazanavi ius, *Kaunas University of Technology, Lithuania*
Wolfgang Klas, *University of Vienna, Austria*
Yiannis Kompatsiaris, *Informatics and Telematics Institute - CERTH, Greece*
Yau-Hwang Kuo, *National Cheng Kung University, Taiwan*
Fuhua Lin, *Athabasca University, Canada*
Alan Liu, National *Chung Cheng Univeristy, Taiwan*
Jonathan Liu, *University of Florida, USA*
Satoshi Nakamura, *Kyoto University, Japan*
Paolo Nesi, *University of Florence, Italy*
Kia Ng, *University of Leeds, UK*
Vincent Oria, *New Jersey Institute of Technology, USA*
Sethuraman Panchanathan, *Arizona State Univ., USA*
Antonio Piccinno, *Univ. of Bari, Italy*
Fabio Pittarello, *University of Venice, Italy*
Giuseppe Polese, *University of Salerno, Italy*
Monica Sebillo, *Universita di Salerno, Italy*
Buntarou Shizuki, *University of Tsukuba, Japan*
Peter Stanchev, *Kettering University, USA*
Mahbubur Syed, *Minnesota State University, USA*
Bjorn Thuresson, *KTH/CSC/HCI, Sweden*
Christian Timmerer, *Klagenfurt University, Austria*
Genny Tortora, *University of Salerno, Italy*
Joseph E. Urban, *Texas Tech University, USA*
Athena Vakali, *Aristotle University, Greece*
Ellen Walker, *Hiram College, USA*
Chi-Lu Yang, *Institute for Information Industry, Taiwan*
Atsuo Yoshitaka, *JAIST, Japan*
Ing Tomas Zeman, *Czech Technical University, Czech Republic*
Kang Zhang, *University of Texas at Dallas, USA*

Publicity Co-Chairs

Kao-Shing Hwang, *National Chung Cheng University, Taiwan*

Proceedings Cover Design

Gabriel Smith, *Knowledge Systems Institute Graduate School, USA*

Conference Secretariat

Judy Pan, *Chair, Knowledge Systems Institute Graduate School, USA*
Dennis Chi, *Knowledge Systems Institute Graduate School, USA*
David Huang, *Knowledge Systems Institute Graduate School, USA*
Rachel Lu, *Knowledge Systems Institute Graduate School, USA*

International Workshop on Distance Education Technologies (DET 2010)

October 14 - 16, 2010
Hyatt Lodge at McDonald's Campus, Oak Brook, Illinois, USA

Organizers & Committees

Workshop Co-Chairs

Qun Jin, *Waseda University, Japan*
Paolo Maresca, *University Federico II, Napoli, Italy*

Program Co-Chairs

Rita Francese, *University of Salerno, Italy*
Mahbub Syed, *Minnesota State University, USA*
Kazuo Yana, *Hosei University, Japan*

Program Committee

Giovanni Adorni, *University of Genova, Italy*
Tim Arndt, *Cleveland State University, USA*
Yuan-Sun Chu, *National Chung Cheng University, Taiwan*
Massimo De Santo, *Universita degli Studi di Salerno, Italy*
Pierpaolo Di Bitonto, *University of Bari, Italy*
Rita Francese, *University of Salerno, Italy*
Kao-Shing Hwang, *National Chung Cheng University, Taiwan*
Pedro Isaias, *Open University, Portugal*
Qun Jin, *Waseda University, Japan*
Yasir Khalifa, *University of Pittsburgh, USA*
Paolo Maresca, *University Federico II, Napoli, Italy*
Andrea Molinari, *University Federico II, Napoli, Italy*
Teresa Roselli, *University of Bari, Italy*
Veronica Rossano, *University of Bari, Italy*
Nicoletta Sala, *University of Italian Switzerland, Switzerland*
Giuseppe Scanniello, *University of Salerno, Italy*
Hui-Kai Su, *Nanhua University, Taiwan*

Mahbub Syed, *Minnesota State University, USA*
Kazuo Yana, *Hosei University, Japan*

International Workshop on Visual Languages and Computing (VLC 2010)

October 14 - 16, 2010
Hyatt Lodge at McDonald's Campus, Oak Brook, Illinois, USA

Organizers & Committees

Workshop Chairs

Gem Stapleton, *University of Brighton, UK*

Program Chair

Giuseppe Santucci, *University of Rome, Italy*

Program Committee

Dorothea Blostein, *Queen's University, Canada*
Paolo Buono, *University of Bari, Italy*
Alfonso F. Cardenas, *University of California, USA*
Kendra Cooper, *University of Texas at Dallas, USA*
Maria Francesca Costabile, *University of Bari, Italy*
Gennaro Costagliola, *University of Salerno, Italy*
Philip Cox, *Dalhousie University, Canada*
Sergiu Dascalu, *University of Nevada, USA*
Davide De Chiara, *University of Salerno, Italy*
Vincenzo Del Fatto, *University of Salerno, Italy*
Aidan Delaney, *University of Brighton, UK*
Vincenzo Deufemia, *University of Salerno, Italy*
Stephan Diehl, *University of Trier, Germany*
Jing Dong, *The University of Texas at Dallas, USA*
Filomena Ferrucci, *University of Salerno, Italy*
Andrew Fish, *University of Brighton, UK*
Paul Fishwick, *University of Florida, USA*
Manuel J. Fonseca, *INESC-ID, Portugal*
Vittorio Fuccella, *University of Salerno, Italy*
Dorian Gorgan, *Technical University of Cluj-Napoca, Romania*

Mao Lin Huang, *University of Technology, Sydney, Australia*
Lars Knipping, *Technische Universitat Berlin, Germany*
Jun Kong, *North Dokota State University, USA*
Robert Laurini, *University of Lyon, France*
Benjamin Lok, *University of Florida, USA*
Piero Mussio, *University of Milan, Italy*
Luca Paolino, *University of Salerno, Italy*
Joseph J. Pfeiffer, *New Mexico State University, USA*
Giuseppe Polese, *University of Salerno, Italy*
Steven P. Reiss, *Brown University, USA*
Giuseppe Santucci, *University Di Roma, Italy*
Monica Sebillo, *University of Salerno, Italy*
Gem Stapleton, *University of Brighton, UK*
Nik Swoboda, *Universidad Politecnica de Madrid, Spain*
Stefano Valtolina, *University of Milan, Italy*
Athanasios Vasilakos, *University of Western Macedonia, Greece*
Giuliana Vitiello, *University of Salerno, Italy*
Kang Zhang, *University of Texas at Dallas, USA*

Table of Contents

Foreword ……..... iii

Conference Organization ……... v

Keynote

Challenges in a New Data Explosion Environment
David Hung-Chang Du……..... xxi

From Computing to Informatics
Yi Deng ……..……..…….....……..... xxii

Content distribution

A Global Data Source for Temporal Web Mining (S)
Mireille Samia……..…….. 3

Metrics for best practice networks analysis (S)
Pierfrancesco Bellini, Daniele Cenni, Alice Fuzier, Paolo Nesi, Gianni Pantaleo, Michela Paolucci……..……..…….. 7

Software Architecture for Web-based Project Management System (S)
Antonio Marcos Neves Esteca, Rogéria Cristiane Gratão de Souza, Adriana Barbosa Santos…. 11

Software understanding and reuse

Software Clustering Using Drools Tool (S)
Nenad Rogulja, Branko Milosavljević, Zora Konjović 17

A model proposal for program comprehension
Francesca Arcelli, Marco Zanoni, Riccardo Porrini, Mattia Vivanti 23

Software Process Reuse Based on Knowledge and Process Components (S)
Andr´e Luiz Peron Martins Lanna, Carlos Alberto Marques Pietrobon 29

Slow Intelligence Systems & Panel

Ontological Filters for Slow Intelligence Systems
Shi-Kuo Chang, Emilio Zegarra, Chia-Chun Shih, Ting-Chun Peng, Francesco Colace and Massimo De Santo 33

Modeling Human Intelligence as A Slow Intelligence System
Tiansi Dong 41

SINMS: A Slow Intelligence Network Manager based on SNMP Protocol
Francesco Colace, Shi-Kuo Chang and Massimo De Santo 47

Building Topic/Trend Detection System based on Slow Intelligence (S)
Chia-Chun Shih, Ting-Chun Peng 53

Emergency Management and Security

A Generic Architecture for Surveillance Systems
Erland Jungert, Christina Grönwall, Niklas Hallberg, Birgitta Kylesten, Fredrik Lantz, Lars Eriksson 57

Privacy Enforcement by Identity Management in Smart Surveillance Systems (S)
H. Vagts, E. Krempel and J. Beyerer 64

Digital Home and HealthCare

RFID Applications in Hospitals – A Case Study for Emergency Department
Yen-Chieh Huang, Chih-Ping Chu, Yu-Shih Lin, Chia-Hao Kuo 70

A Technical Model Using Genetic Algorithm For the Cosmetic Tourism Business
Chi-Lu Yang, Yen-Chieh Huang, Yeim-Kuan Chang, Chih-Ping Chu 76

Networking and reconfiguration

Remote Query Processing in Wireless Sensor Networks using Coordinated Mobile Objects
Juzheng Li and Sol M. Shatz 82

Congestion Optimization in Data Networks
Doraid Dalalah 88

SICSDA: an Adaptive Configurable Distributed Software Architecture Applied to Satellite Control Missions (S)
Adriana Cursino Thome, Mauricio G. V. Ferreira, Joao Bosco S. Cunha 94

Media Distribution and e-commerce

A Media Sharing Platform Built With Open Source Software
Karel Braeckman, Robbie De Sutter, Mike Matton, Tine Blomme 98

Electronic Payment and Online Gaming (S)
Gerald Draxler, Johannes Sametinger, Andreas Wiesauer 104

Architectures for communities and management

An Engine for Local Communities Sharing Services and Content (S)
Pierfrancesco Bellini, Daniele Cenni, Vladimiro Corsi, Paolo Nesi 110

THE EXPLAINER: Using the Computer to Help Collaborate in Connecting the Dots to Resolve Complex Problems (S)
Donald V Steward, Josh Knight 114

Verification of Security Policy of Service Oriented Systems
Antonio Maña, Gimena Pujol 120

Multimedia Analysis and Processing

Software-Only M-JPEG Performance for Emerging Multimedia Applications (S)
Varun Kulkarni and Jonathan C.L. Liu 126

A 3D Curve Chain Code Representation for the Analysis of Human Motion Similarity
Francisco Torres, Edward Chow. 132

iRock Interactive Multimedia System
Chris Ellison, Bee Ong and Kia Ng 137

Media, metadata and semantics

Towards Hybrid Index Structures for Multi-Media Search Criteria
Richard Göbel, Carsten Kropf 143

Common Interoperability Framework for Defense Architectures, A Web Semantic approach (S)
Abdullah S. Alghamdi, Zeeshan Siddiqui 149

Assessing Open Archive OAI-PMH implementations
Emanuele Bellini, Marcel Aime Deussom, Paolo Nesi 153

Advanced Pattern Match techniques for identifying Similar Transactions in a Database System (S)
Thangaraj Karol Stuart , Zhiwei Xu 159

User Interface

A Display Adaptation Model for Nomadic Users (S)
Luis Carlos Niño-Tavera, Angela Carrillo-Ramos, Edgar Enrique Ruiz…….....…….....……..... 164

A Hybrid Approach for Adapting Web Graphical User Interfaces to Multiple Devices using Information Retrieved from Context
Carlos E. Cirilo, Antonio F. do Prado, Wanderley L. de Souza, Luciana A. M. Zaina……..…... 168

Agile User Interfaces in Disaster Management
Veselina Milanova, Kay-Uwe Schmidt, Jorg Dorflinger, Axel Spriestersbach, Susan Thomas... 174

Analysis of the UCD process of a web-based system
Carmelo Ardito, Paolo Buono, Maria F. Costabile…….....…….. 180

DET Workshop

Eclipse and Jazz Technologies for e-Learning

Enforcing Team Cooperation: an example of Computer Supported Collaborative Learning in Software Engineering (S)
Mauro Coccoli, Paolo Maresca, and Lidia Stanganelli …….…… 189

The assessment of knowledge requirements in a network of goal oriented organization
G. Franco, P. Maresca, G. Nota…….……..... 193

Learning Objects

A Model for Teaching Distributed Computing in a Distance-based Educational Environment
Petra le Roux, Alta van der Merwe, Aurona Gerber…… 200

How do archaeologists work? A multimedia digital object to learn their investigation process
C. Ardito, R. Lanzilotti…….....….....…….....…..... 206

An Adaptive Test for Learning Objects: item calibration
P. Di Bitonto, M. Laterza, T. Roselli, V. Rossano…….....…….....…..... 211

e-Learning

Barriers to Distance Learning in Web Accessibility for Persons with Blindness and Visual Impairment
Kathleen A. Stirbens, Angela Guercio, Charles Haiber, Joseph Williams… 216

A Novel Approach for Attention Management in E-learning Systems
Gennaro Costagliola, Mattia De Rosa, Vittorio Fuccella, Nicola Capuano, Pierluigi Ritrovato… 222

Full Formative Assessment based on Educational Video Games (S)
Salvatore Cuomo, Vittorio Fuccella, Aniello Murano… 228

Opencast, the Opencast Community and Opencast Matterhorn – A global community-driven project to build a management system for academic video (S)
Todd Jensen, Olaf A. Schulte… 232

e-Learning and e-Collaboration

Sakai 3: the Co-evolution of Product and Community
Clay Fenlason… 235

Towards Integration of Data Base Management System for Open-source Course Management Systems (S)
Yuki Terawaki, Yuichi Takahashi, Yuji Tokiwa, Yasushi Kodama, Kazuo Yana… 239

3D virtual space for a quality e-Learning environment in collaborative study (S)
Yoshiyuki Hino, Kazuo Yana, Hisato Kobayashi, Kimiyoshi Hayashi, Yasushi Kodama, Billy Pham… 243

VLC Workshop

Visual computing for expert communities

Visualizing the Commonalities Between Hierarchically Structured Data Queries
Chad Jones, Ryan Armstrong, Kwan-Liu Ma 251

Towards Security-aware Program Visualization for Analyzing In-lined Reference Monitors (S)
Aditi Patwardhan, Kevin W. Hamlen, and Kendra Cooper 257

Geovisualization

Immersive Gesture-based Interface for Effective 3D Geographic Data Manipulation (S)
Andrea F. Abate, Vincenzo Del Fatto, Michele Nappi, Stefano Ricciardi, Genoveffa Tortora 261

Augmented Map navigation through Customizable Mobile Interfaces
D. De Chiara, L. Paolino, M. Romano, M. Sebillo, G. Tortora, G. Vitiello 265

A design methodology for the development of GIS applications
Rosanna Cassino, Monica Sebillo, Maurizio Tucci 271

Formal theory of Visual Languages

Visual Spatio-Temporal Programming with a 3D Region Connection Calculus (S)
Julia Albath, Jennifer L. Leopold, Chaman L. Sabharwal 277

On the Relative Expressiveness of Second-Order Spider Diagrams and Regular Expressions (S)
Peter Chapman and Gem Stapleton 283

Visual Interfaces and Environments

Image Annotation System Using Visual and Textual Features
Swapnil Barai and Alfonso F. Cardenas 289

Reverse Engineering Visual Design Models from Code: What are the Issues? (S)
Kendra M.L. Cooper, Jeffrey Koch 297

CAVEMANDER: Creating 3-D Command-and-Control Scenarios for the CAVE Automatic Virtual Environment (S)
Muhanna Muhanna, Sermsak Buntha, Sohei Okamoto, Michael J. McMahon, Jr., Sergiu Dascalu, Frederick C. Harris, Jr. 300

Sketch Computing

Recognizing Sketches of Euler Diagrams Drawn with Ellipses
Aidan Delaney, Beryl Plimmer, Gem Stapleton, Peter Rodgers 305

Automated Clock Drawing Test through Machine Learning and Geometric Analysis (S)
Anupam Guha, Hyungsin Kim, Ellen Yi-Luen Do 311

Interpreting Gestures for Text Entry on Touch Screen Devices
Gennaro Costagliola, Vittorio Fuccella, Michele Di Capua 315

Evaluation of Paper-Pen based Sketching Interface (S)
Pankaj Rajan, Paul Taele, Tracy Hammond 321

Sketch-a-Doc: Using Sketches to Find Documents (S)
Manuel J. Fonseca, Daniel Goncalves 327

Reviewer's Index 331

Author's Index 333

Note: (S) means short paper.

Keynote I:

Challenges in A New Data Explosion Environment

David Hung-Chang Du, Qwest Chair Professor
Department of Computer Science and Engineering
University of Minnesota, Minneapolis

Abstract

The Internet today has grown to an enormously large scale. Devices large and small are connected globally from anywhere on the earth. With the rapid advancement of technology, we now also have cheap, small and smart devices with high computing power, powerful wireless communication capability and relatively large storage capacity. These devices are designed to improve our daily life by monitoring our environment, collecting critical data, and executing special instructions. These devices called sensors have gradually become an essential part of our Internet. They are embedded into physical systems (Cyber Physical Systems) and normal daily used objects (Internet of Things) to form a extremely large networks that we have ever witnessed. As a result, unprecedented amount of data are collected by these devices and are available via Internet. How to build such a system and manage and look for the desired information becomes a great challenge. In this talk, we will examine the challenges in this new computing and communication environment. These challenges include new data representations and models, data security and privacy issues, long-term data preservation methods, new data storage systems and even a new Internet architecture.

About David Hung-Chang Du

David Hung-Chang Du . received the B.S. degree in mathematics from National Tsing-Hua University, Taiwan, R.O.C. in 1974, and the M.S. and Ph.D. degrees in computer science from the University of Washington, Seattle, in 1980 and 1981, respectively. He is currently the Qwest Chair Professor at the Computer Science and Engineering Department, University of Minnesota, Minneapolis. He has served at US National Science Foundation as a Program Director from 2006-2008. His research interests include cyber security, sensor networks, multimedia computing, storage systems, high-speed networking, high-performance computing over clusters of workstations, database design and CAD for VLSI circuits. He has authored and co-authored more than 200 technical papers including more than 100 journal papers in these areas. Dr. Du is an IEEE Fellow and a Fellow of Minnesota Supercomputer Institute. He is currently served on a number of journal editorial boards. He has also served as Conference Chair and Program Committee Chair to several conferences in multimedia, database and networking areas. Most recently, he is the General Chair for IEEE Security and Privacy Symposium (Oakland, California) 2009 and Program Committee Co-Chair for International Conference on Parallel Processing 2009, and General Chair of ICDCS 2011.

Keynote II:

From Computing to Informatics

Yi Deng, Dean & Professor
College of Computing and Informatics
University of North Carolina at Charlotte

Abstract

In the 21st Century economy, every industry has become an .information industry, where computing, information and communication are not only essential utility, but more importantly, the core driver for competitiveness and innovation. The field of computing has evolved and expanded from its focus on computation to data, information and knowledge. The ubiquitous and deep penetration of information and technology, and exponential growth of scale and complexity of data give arises to a new discipline and industry of informatics - the integration, rather than application, of information and technology with specific disciplines and industries. This phenomenon will have transformational effect on research and education and on the future of our institutions. I will discuss the trends and their implications to our students, faculty, programs and organizations, and how our college plan to deal with these changes.

About Yi Deng

Yi Deng is the Dean and Professor of the College of Computing and Informatics at the University of North Carolina at Charlotte, a position he has held since July 2009. Prior to joining UNC-Charlotte, he had been the Dean and Professor of the School of Computing and Information Sciences at the Florida International University (FIU) from 2002 to 2009. He has authored or co-authored over 100 research papers in peer-reviewed journals and proceedings, and been awarded numerous research grants as the principal or co-principal investigator totaling over $21 million, most of which from the premier federal agencies. He has initiated and led many large scale multidisciplinary R&D and education projects and initiatives, founded and/or directed several research centers. He has been an active contributor to the professional and research community in various leadership capacities. He co-founded and co-led the Latin American Grid (LA Grid) Consortium and founded and directed the Consortium of Partnership for International Research and Education (PIRE), whose partners include leading institutions and research centers in the U.S., Mexico, Argentina, Spain, China, India and Japan. Yi Deng received his B.S. in Computer Science and Technology from the University of Science and Technology of China in 1983, and M.S. and Ph.D. in Computer Science from the University of Pittsburgh in 1990 and1992, respectively.

Proceedings
International Conference on
Distributed Multimedia Systems
(DMS 2010)

Co-Editors

Paolo Nesi, *University of Florence, Italy*
Kia Ng, *University of Leeds, UK*

A Global Data Source for Temporal Web Mining

Mireille Samia
St. Thaddeus Research, Düsseldorf, Germany
mireillesamia@hotmail.com

Abstract

Our concept, namely Temporal Web Mining (TWM) [5, 6, 8, 9], is an extension of temporal data mining and Web mining. TWM intends to interpret data in real time. It provides an architecture that aims at building a logical structured information source, namely a common reference data warehouse or a global data source, which for instance, records, combines, or shares data between multiple different sources. Moreover, to enhance data prediction, TWM intends to support or improve real-time data recording, data extraction, data sharing or data analysis by minimizing the delay in data transfer and without losing the reliability of our global data source. To achieve these goals, we propose the architecture of a multicast network for the TWM global data source. In this paper, we define TWM before presenting its architecture [5,6,8,9]. Then, we propose and study our multicast network model for the TWM global data source.

1. Introduction

Temporal Web Mining (TWM) concerns the Web mining of data with temporal information, and deals with real-time data [5, 6, 8, 9]. TWM aims at studying, analyzing, and predicting significan knowledge from historical data, current data, as well as real-time data extracted from any type of data store, data source, or information repository. By combining data from different sources, TWM aims at predicting consequences of actions and at associating a cause with a consequence, whenever possible, in order to give potentially useful information about future events. Hence, TWM can play a key role in the advance of different areas, such as volcanism and aviation, environmental sciences, medicine, engineering, finance and marketing. Temporal Web Mining provides an architecture, which describes how data with temporal information is gathered and combined (if needed), from different sources, in real time over the Web. By building a logical structured information source, namely a common reference data warehouse or a global data source, the TWM architecture can enhance the communication, coordination and integration of different societies, communities, and institutions. Our global data source for TWM establishes a data exchange system which can be used by different sources for data recording, data extraction, data sharing, or data analysis. More specificall , our global data source sends the requested information to one or more sites. Multiple different sites may make concurrent requests of the same data. The aim of our global data source is to multicast; i.e., to deliver the same requested data to the sites in question at more or less the same time, instead of having to send this data to just one of these sites at a given time. Moreover, to support the real-time issue, the minimization of delay in data transfer without losing the reliability of our global data area is needed. To achieve these goals, we propose the architecture of a multicast network for our global data source, which is responsible for handling data requests of multiple sites and where the delay in data transfer is minimized. In this paper, section 2 provides related work. In section 3, we defin TWM before presenting its architecture [5, 6, 8, 9]. In section 4, we propose and study our multicast network model for the TWM global source data. Section 5 concludes this paper, and points out future work.

2. Related work

Temporal Web Mining [5,6,8,9] aims at combining temporal data mining and Web mining in order to deal with real-time data. By having our common reference data warehouse (also known as global data source), a data exchange system between different communities can be created in order to record, exchange, use, process, and analyze data, assembled from different sources. Since multiple different sites may make concurrent requests of the same data, our global data source needs to multicast the requested data. Multicasting means that a piece of data (i.e. a packet) is sent to multiple sites at more or less the same time, instead of having to send that packet once for every site [4]. Multicasting has many advantages described next. It conserves bandwidth by forcing the network to do replication only when necessary [1]. Multicasting reduces the network load by sending a single packet from the source to an arbitrary number of receivers by replicating the packet within the network at points along a multicast delivery tree rooted at the packet's source [2]. Moreover, sources send packets to a multicast group, which contains the subscribed receivers [4].

TWM intends to improve data recording, data extraction, data sharing and data analysis, by optimizing the network performance. In TWM, the real-time issue comes into play. Consequently, a reliable global data source, as well as a reliable stream delivery, are needed. A real-time traffi requires minimal delays. However, multicast datagrams can be delayed. For this reason, we propose our multicast network model, where the presence of delays is considered. In this multicast network, routers are involved, in order to move packets from one network to another. By giving additional functionalities (such as backup and error correction) to these routers (or intermediate sites) and by findin the optimal number of intermediate sites, we enhance the performance of our multicast network, and hence, the performance of our TWM global data source.

3. Temporal Web Mining (TWM)

Definition of Temporal Web Mining Temporal data discovered by the application of temporal data mining techniques is used in the Web mining process in order to retrieve useful data with temporal information in real time over the Web. The derived useful data with temporal information discovered by the application of Web mining techniques is used again in the temporal data mining process [5, 6].

One of the application scenario of TWM is in volcanism and aviation [9]. In Figure 1, whenever a volcanic activity occurs, volcano data and aviation data are analyzed instantly, if required combined with other data (such as wind speed data, wind direction data) from *different sources* (such as a station where a wind sensor is located). The derived temporal data is forwarded, in real time over the Internet, from $Station_A$ to $Station_B$. In $Station_B$, in order to extract new relevant information, it is combined, if necessary, with new data (such as historical volcanic ash data, historical incidents of aircrafts entering ash clouds) from $Station_B$ or from *other different sources* (such as meteorological station, air traffi control center). Then, it is analyzed in order to discover new temporal data that can reveal key information about volcanic activities and volcanic ash dispersal. The new derived temporal data is sent back to the original source; i.e. to $Station_A$. Analyzing this data connect different stations and provide a possibility to reduce volcanic ash hazards and aircraft damages.

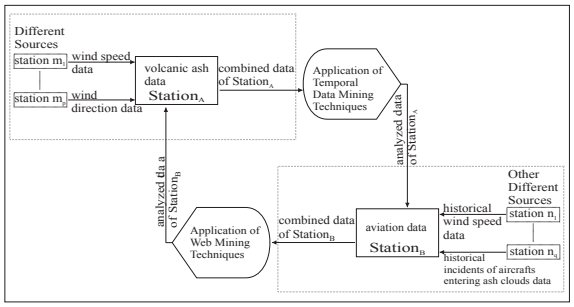


Figure 1. An application scenario for TWM in volcanism and aviation.

Description of the TWM architecture In Figure 2, the architecture of Temporal Web Mining [5,6] consists of three main components: original source component, destination source component and global data source component. *Temporal data* represents the data derived from the application of temporal data mining techniques. *New temporal data* is the data obtained from the application of Web mining techniques. In the *original source component*, *multiple different sources* interact, in real time, with each other in order to discover, retrieve, combine, and analyze data. One or more of these *multiple different sources* receive real-time data. The obtained temporal data is collected in order to be selected and explored. Applying temporal data mining techniques, this temporal data is analyzed. From the *temporal data* database, it is sent to the global data source component, where it is saved. Furthermore, in order to reveal more important information, this temporal data may be combined with other data retrieved from the global data source component. It can also be added to other data received from the destination source component. In this case, it is sent over the Internet to the destination source component. The *destination source component* provides an environment for decision support using mining techniques. The *new temporal data* database receives the analyzed temporal data from the original source component in real time and over the Web (cf. Figure 2). It is examined and, if needed, combined with auxiliary data acquired from *other multiple different sources* or from the global data source component. Then, it is analyzed in order to extract derived temporal data. The new temporal data is sent back, in real time over the Internet, from the *new temporal data* database to one or more sources of the original source component. It is also saved in the global data source component. The *global data source component* consists of the *global data source*

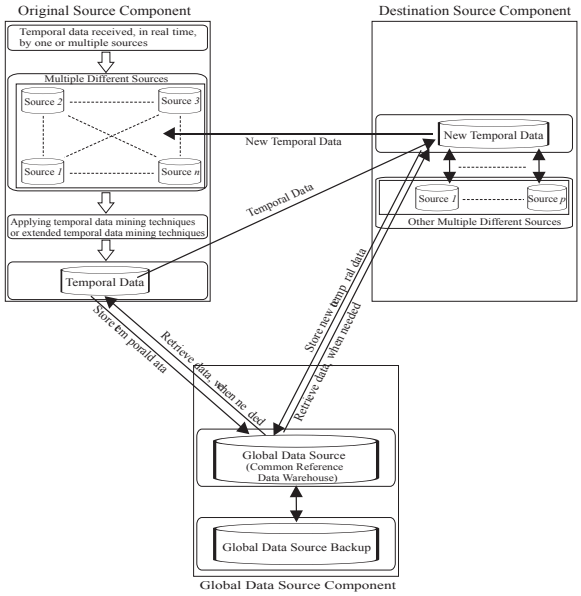


Figure 2. The TWM architecture.

database (also called common reference data warehouse) and the *global data source backup* database. The *global data source database* receives data from the original source component and the destination source component. It manages the history of the data or real-time data collected from these different sources. It also stores the analysis of the data derived from the application of temporal data mining and Web mining techniques in addition to the data received from different sources. It backups this data in the *global data source backup* database. Whenever needed, the *global data source component* forwards the requested data to the *original source component* or to the *destination source component*. Copies of the same data are sometimes requested similtaneously from different sites. Instead of sending that data to just one site at a given time, it can be sent to more than one site at more or less the same time by using our multicast network [7] (cf. Figure 3). Our multicast network is studied in section 4.

4. Our TWM global data source

Since our TWM global data source deals with data changing rapidly over time, the delay in data transfer should be minimized. This can be done by using our multicast network, which is a real-time multicast network, and its use is not only limited to TWM global data source. Without forgetting the importance of reliability in the multicast network and our global data source, our goal is to optimize the performance of our multicast network by minimizing the delay and by findin the optimal number of intermediate sites (intermediate routers). In this section, we study our multicast network and its architecture. The results of the related simulation are also discussed.

Assumptions considered in our multicast network model Figure 3 represents our multicast network model. Our simulation model involve 100 sites spread in the multicast network. The presence of packet loss is taken into account and hence, retransmission of lost packets are simulated and analyzed. Packets are generated at the source as a Poisson process with arrival rate λ. The arrival rate λ varies between 0.1 and 0.9. The service rate at the firs (respectively the second) stage of the network is μ (respectively σ). μ and σ are taken larger (or equal) to 1 so that the system response time does not diverge. In fact, λ must be smaller than μ and σ. The service rate at the intermediate sites (or routers) is taken equal to γ. γ is also larger or equal to 1. μ, γ, and σ are all considered to have an exponential distribution. In the multicast network model, the routers are active. They have backup functions and they retransmit any packet which is lost in the second stage of the network.

Architecture of our multicast network In Figure 3, the architecture of our multicast network is presented. It consists of a source node connected to a number of networks. Each of these networks is connected to an intermediate site

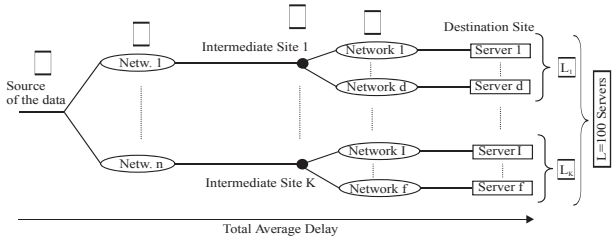


Figure 3. Multicast network.

(i.e. a router). Each router is connected to several networks, which lead to a number of destination sites. The total number of destinations is still equal to 100. Here, our major objective is to fin the optimal number of intermediate sites so that the overall performance of the network is optimized. In Figure 2, the *global data source component*, or more specifically, the *global data source* depicts the source node. The destination sites are those that requested concurrently the same data. In our case, the destination sites are the *original source component* and the *destination source component*.

Analysis of our multicast network We need to multicast packets to 100 destination sites. Packets are forwarded from the root (i.e. source of the data) along a distribution tree to each receiver in the multicast group. Each of the originated packets is sent from the source of the data to Netw where only a service rate μ is considered. Then, from each of the Netw, it is sent to a router (i.e. intermediate site) where a service rate γ is considered, and backup of a packet and error-correction are considered. Whenever a packet is corrupted, lost or duplicated, a NAK (Negative Acknowledgment) message is sent back to the source of data with the packet sequence number to be retransmitted. Each of the intermediate sites (or multicast routers) forwards the datagram to other sites where a service rate σ is considered, and error-correction is supported. Whenever a delivery error occurs, a NAK message is sent back to the related intermediate site asking it to retransmit the corresponding packet. The intermediate site will again retransmit the requested packet to the destination site.

Multicast network simulation results Let K and L be the number of routers (i.e. intermediate sites), and the number of destination sites, respectively. Note that $L_1 + L_2 + L_3 + + L_k = L = 100$.

We aim to prove that for a certain value of K and L, when backup and error recovery are used at routers (or intermediate sites), the network performance will enhance. In [3], it was proven that for general multi-link networks, the optimum value of K as a function of L (assuming all other parameters in the system are constant) is given by $K \approx L^{1/r}$, where $r > 1$. All we need to do is to show that the above result still hold for our multicast networks. In our simulations, we considered $L = 100$ nodes.

Intermediate site optimal number To fin the optimal number of intermediate sites of our multicast network, we fi the values of λ, μ, γ and σ. As state previously, the

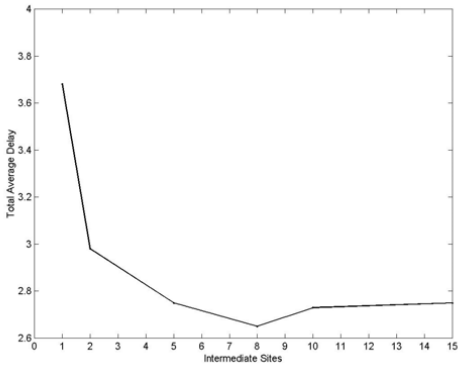


Figure 4. Total average delay variation as a function of intermediate sites.

number of intermediate sites K varies as we change r: $number_of_intermediate_sites = round(total_sites_number)^{1/r}$, where $r > 1$
$K = round(L)^{1/r}$, where $r > 1$
We assume that λ is equal to 0.9. The firs network service rate μ is equal to 1, while γ and σ are both equal to 2. The graph of Figure 4 shows the variation of the delay as a function of the number of intermediate sites K. From the graph of Figure 4, we deduce that the optimal number of intermediate sites is equal to 8, because the total average delay of the network at that point represents the minimal value of the curve. In fact, the network delay drops. So with 100 multicast users, we should have 8 intermediate sites. The firs 7 intermediate sites are connected each to 12 end-users (i.e. the cluster size is 12 users) and the 8th intermediate site is connected to 16 end-users. A decrease in the number of K decreases the load on the fir t stage, while it increases the load on the second stage. In fact, if the number K is very small, the K routers will have to bear a large load resulting from the handling of error correction and backup. This will lead to a larger overall delay since the routers will be like a bottleneck. While an increase in the value of K will decrease the load on the intermediate sites since the workload will be divided among all the routers. This will result in a decrease in the overall delay. Consequently, the use of the intermediate site optimal number in our multicast network enhances the performance of the network. By enhancing the performance of the multicast network, we improve the performance of our TWM global data source. More clearly, whenever data is needed to be multicasted to different destinations, data is forwarded from the source of the data (i.e. global data source) along a distribution tree to each receiver (i.e. destination site). The destination sites; i.e., the original source component or the destination source component, do not communicate directly with the global data source. They communicate with their corresponding intermediate site, which supports backup and error correction. Then, if an intermediate site (respectively a destination site) receives a corrupted piece of data or does not receive the requested piece of data, a message is sent back to the source of data (respectively the intermediate site), requesting it to retransmit the needed data. Consequently, the workload is distributed among the global data source and the intermediate sites. The workload on the global data source is minimized, which helps to reduce its response time. This resulted in minimizing the delay in data transfer, as well as data loss and data errors. Moreover, in our multicast network, the intermediate sites divides the total number of sites to clusters. Every intermediate site is responsible for one cluster of nodes. We showed that by taking the number of intermediate sites equal to its optimal value, we enhance the performance of the network and of the TWM global data source, by minimizing its total average delay, and keeping its reliability.

5. Conclusion and future work

TWM intends to support the real-time issue and to enhance, for instance data analysis of our global data source, by improving the network performance. We presented the TWM architecture and showed that by using our multicast network, the delay in data transfer and the network load are minimized. This resulted in better performance of our multicast network and our TWM global data source. Moreover, by findin the optimal number of intermediate sites, we found that the network performance and the performance of our global data source are optimized. A future work is to improve the time of data extraction by reducing the cost of a query execution, and by minimizing the cost of findin the appropriate path which retrieves the required data.

References

[1] E. Amir, S. McCanne, and R. H. Katz. Receiver-Driven Bandwidth Adaptation for Light-Weight Sessions. In *ACM Multimedia*, pages 415–426, 1997.

[2] D. Comer. *Internetworking with TCP/IP: Principles, Protocols, and Architecture*. Prentice Hall, 3rd edition, 1995.

[3] H. W. Maalouf. Optimisation of the Communication Network Performance of Distributed Systems with Resequencing Constraints. PhD Thesis. Digital Communications Section. Department of Electrical and Electronic Engineering, Imperial College, University of London. 1998.

[4] S. McCanne. Scalable Multimedia Communication with Internet Multicast, Light-weight Sessions, and the MBone, 1998.

[5] M. Samia et al. From Temporal Data Mining and Web Mining To Temporal Web Mining. In *Proceedings of the Sixth International Baltic Conference on Databases and Information Systems (BalticDB&IS'2004)*, volume 672, pages 232–245, Riga, Latvia, June 2004.

[6] M. Samia et al. From Temporal Data Mining and Web Mining To Temporal Web Mining. In *Databases and Information Systems: Selected Papers from the Sixth International Baltic Conference DB&IS'2004*, volume 118 of *Frontiers in Artificial Intelligence and Applications*, pages 91–102. IOS Press, 2005.

[7] M. Samia. Optimization of IP Multicast Networks Performance. Master's Thesis, Faculty of Natural and Applied Sciences, Department of Computer Science, Notre Dame University, Lebanon, June 2001.

[8] M. Samia. Temporal Web Mining. In *15. GI-Workshop über Grundlagen von Datenbanken (15th GI-Workshop on the Foundations of Databases)*, pages 27–31, Tangermünde, Germany, June 2003.

[9] M. Samia. Temporal Web Mining. PhD Thesis, Heinrich-Heine University of Düsseldorf. Shaker Verlag, August 2006.

Metrics for best practice networks analysis

Pierfrancesco Bellini, Daniele Cenni, Alice Fuzier,
Paolo Nesi, Gianni Pantaleo, Michela Paolucci
DISIT-DSI, Distributed Systems and Internet Technology Lab
Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze, Firenze, Italy
http://www.disit.dsi.unifi.it/, nesi@dsi.unifi.it, paolo.nesi@unifi.it

Abstract

Best practice networks are thematic social networks focused on sharing common area works and goals. The analysis of user behavior on social network is fundamental to tune services and stimulate the network growth. This is even more relevant for best practice collaborative networks where details about collaboration may lead to understand the effective activities and role of people in the several groups. In this paper, a set of metrics for analyzing user behavior related to download on a best practice network is presented. The metrics are currently active, as administrative tools for the Mobile Medicine best practice network.

1. Introduction

Many widespread Social Networks (SN) are mainly focused on contents (e.g., YouTube, Flickr, LastFm) [1], whereas others are more focused on: establishing relationships among users (Facebook, MySpace, Orkut, Friendster) [2]. SN analysis is typically focused on analyzing relationships among users, users and groups, in order to identify which are the most central users and groups, and, on the other hand, which are those that are frozen out, namely being those running the risk of losing interest in the network activities due to a serious lack of involvement.

On the other hand, Best Practice Networks, BPN, are thematic social networks where smaller groups of users share content, contacts and information for a common goal. Some of them are set up by thematic communities, associations and large companies to capitalize skill and knowledge, by facilitating communications and interactions among personnel and the growth of internal content repository and knowledge. As to BPN, to keep trace and analyze user activities is becoming more and more fundamental for tuning services and predicting needs, such as market trends, reactions of users, the product's acceptance, educational needs and problems, etc.

The main differences of Best practice Networks, BPN, with respect to classical massively widespread SNs may be summarized in what follows, where some of the main peculiar features of the BPNs have been highlighted. Therefore:

- BPNs present a larger percentage of active contributors, for both content and comments with respect to SN, the size of the internal communities are smaller. Users of BPN are typically more motivated to participate since BPN are strongly thematic and connected to the user's work. In SN, the effective/active contributors are typically much less than 10%, all the others are merely observers;
- BPNs are typically without advertising, since the business model is designed to provide a service, in order to understand user needs and it is not meant to make earnings with product advertising on a larger number of users.
- BPNs have to cope with a large variety of content types and not only with videos and images like it occurs on large SNs. Most SNs can manage only simple audiovisual content (e.g., images or videos, which are the simplest content type to be generated by users, for example via digital camera and/or phones). Many BPNs have to be able to cope with complex interactive content for education, edutainment, and/or entertainment experiences. This means having more complex semantic descriptors for those content items. At present, there are many multimedia and cross content formats.
- BPNs have to cope with a large variety of semantic computing activities/algorithms: for content adaptation and processing (also due to the complexity of content) of user contributions: user generated content, indexing of user contributions/ comments, etc. In BPN, the number of users and items is smaller, so that the complexity of semantic processing in terms of triples and ontology complexity may be larger, though still economically viable;
- BPNs can typically cover more platforms for providing services and content distribution; for example: PC, windows mobile PDAs, iPhone, etc.;
- BPNs have more complex content and services, therefore users/content profiles/descriptors have to

consider static as well as dynamic aspects [3]. In BPN, like in SN, dynamic aspects are much more relevant since they may be continuously updated.

This paper focuses on the experience in defining, realizing and analyzing a set of metrics for BPN where a large number of content types, devices and content models is presented (please note that this paper is only marginally on user behavior analysis and semantic computing; a complementary paper can be recovered from [3]). The experience refers to two BPNs based on cross media content semantic model: the XMF (cross media finder) (http://xmf.axmedis.org) (educational) and the Mobile Medicine (http://mobmed.axmedis.org) (medical and mobile). The proposed solutions have been developed by using AXMEDIS content and media content processing language [4].

2. Best Practice Network Architecture

The BPN model exploiting the cross media multichannel distribution has been created to provide support in distributing cross media content towards different kinds of device: PC, PDA and Mobiles (smarphones, iPhone, iPod, iPad).
The simplified architecture of the BPN considered solution is depicted in Figure 1, where main modules are reported: user management (registration and profiling), mobile support (mobile content distribution and monitoring), user generated content support (in connection with back office tools for content ingestion, processing and adaptation), collaborative support among users (chat, grouping, messaging, notifications, ..), indexing and querying (multilingual indexing, fuzzy support, semantic indexing, ..), recommendation support for content and users [3].

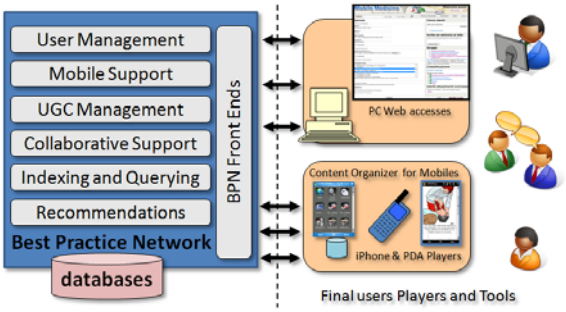


Figure 1 – Simplified BPN Architecture

All the capabilities are deployed and shown to the BPN users via different front-ends, specialized for device category: PC (Windows, Mac OS X, Linux, Unix, etc.), Smartphone, iPhone with a dedicated Application (see Mobile Medicine iPhone App, http://itunes.apple.com/en/app/mobile-medicine/id359865882), Windows Mobile PDA with a dedicated application called AxObjectFinder (see http://mobmed.axmedis.org/drupal/?q=en-US/help). The provided front-ends share the same database and user registration mechanism, and they may provide different services via different sets of languages. Moreover, different devices can support different modalities to access certain content: video may be accessed in download, progressive download (also called HTTP streaming), or even via P2P (only on XMF). These capabilities are only available for some devices and solutions. This is the typical complexity of a multichannel, multi-device, content management and distribution system for a BPN.

In BPN content types may range from single files: audio, video, images, documents and animations, simple slides, animations; to cross media such as: interactive guidelines/procedures, calculators and tools, questionnaires and data collection, leaning objects. The cross media content and file formats may be of several kinds ranging over: MPEG21, SCORM, ePub, XMF, HTML, SMIL, NewsML, SportML, METS, etc. In this context, cross media means objects produced in AXMEDIS MPEG-21 files formats [5].

3. User Behavior Metrics on BPN

BPNs organize and classify content, so that users can easily access, retrieve and share. Leveraging Mobile Medicine BPN features; an engine to monitor data has been realized in order to reason upon content descriptors, user profiles, ads profiles, user generated content descriptors, device capabilities descriptors, etc.
In order to provide the administrator with an effective tool able to identify the general trends of the BPN including downloads and user behavior, a set of metrics has been defined. Therefore, the identified set of metrics has been implemented. What follows is a selection of these metrics and the analysis of results about their application on Mobile Medicine BPN. The selected metrics are those that can highlight better the differences between the SNs and the BPNs.

Usage of Devices and Platform

As mentioned, a BPN like Mobile Medicine, offers a range of content formats that may be accessible by different types of devices and platforms. Figure 2 shows the distribution of device/platforms most preferred by users to access content (for both download and direct play). Most downloads are performed via PC-windows; however, also other operative systems (Linux, Mac Os X) and devices are used (PDA, iPhone, etc.). Therefore, the 12% of users did get access to content by using mobile devices (iPhone, AxObjectFinder), the 86% by using desktop.

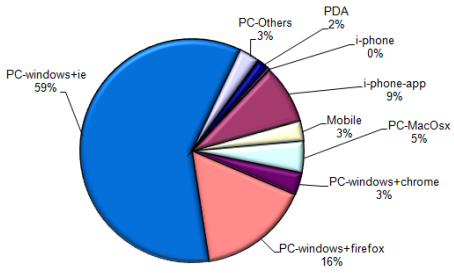


Figure 2: Distribution of content access per device/platform

Proposed and Preferred Content Formats

Given the great diversity of content the platform offers, it was necessary to identify which kinds of format are more appreciated by the BPN users with respect to those provided within BPN. In Figure 2, the distribution of accessed content format is presented. The distribution highlights that cross media content is the most accessed format, followed by video and HTML content pages (still in this content, interactive elements similarly cross media). On mobiles the most selected content formats are cross media with the 82% and video with 9%.

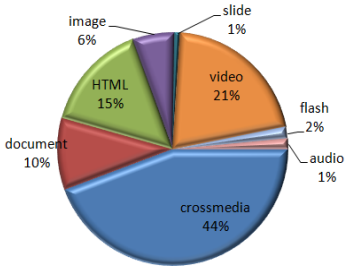


Figure 3: Distribution of accessed content formats

Figure 4 highlights the strong correlation between the content offered population in terms of formats and the actual distribution of content formats as depicted in Figure 2. This graph could estimate what object types are deprecated by users or what classes are more useful for them.

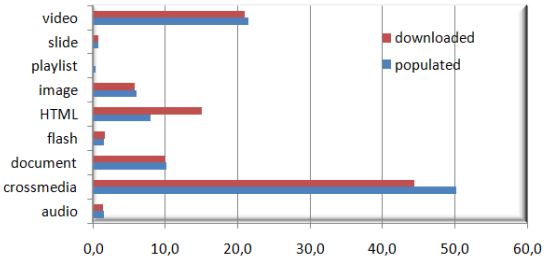


Figure 4: Proposed vs accessed content formats

Content Access Modality of Users

The content accessed via PC and mobiles may be downloaded or played via the HTTP connection in progressive download. This last modality is possible on PC and iPhone, while presently it is not viable on Windows Mobile PDA devices endowed with AxObjectFinder tool. Only 1% of content was accessed via AxObjectFinder, while 90% is accessed via online progressive download and only the 10% via download (including the accesses via AxObjectFinder). From this analysis, it is self-evident that users prefer to access the content online rather than to have it on their devices. On iPhone, it is possible to have online execution of content or download and successive offline play via the Mobile Medicine App. In this case, iPhone users of Mobile Medicine prefer accessing the videos (38%), HTML (26%) and PDF (21%) while only a small part of them prefer other cross media (please note that dosages and medical calculators are classified as HTML). PDA users prefer accessing the cross media for the 94%. For PDA, cross media include the typical medical tools for dosages, guidelines, checklists, procedures, etc. What was impressive is the adoption of iPhone for on-line video play. It seems to be the most preferred platform for video play.

4.2 Content Stability and User Fidelity

Another important factor is the assessment of content stability and user fidelity.

Object Stability: In SN, the distribution of content accesses is typically strongly exponential, with only a small percentage of content highly demanded. In BPN the distribution is still exponential but less explosive and more stable. On such grounds, an interesting metric would allow understanding which is the level of interest for a given content item within the BPN community, and thus which is the content permanence in the user's interests. To this end, S_5 has been defined to measure the so called Object Stability (see Figure 6a). It measures how many times a content object has been downloaded in a period of time. We have chosen a histogram as the most representative graph type:

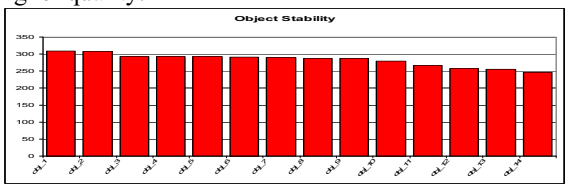
$$\begin{cases} x\text{-}assis = [o_1, o_2, \ldots, o_n] \\ y\text{-}assis = [sum_o_1, sum_o_2, \ldots, sum_o_n] \end{cases}$$

Where: o_i are the object (ids) downloaded and sum_o_i is the number of downloads for each object.

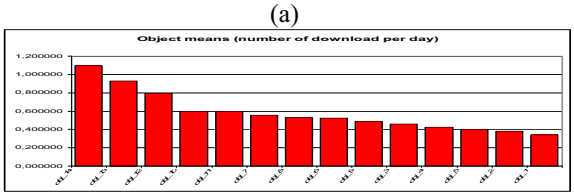
Moreover, it can be also interesting to normalize the distribution, with respect to the duration of the period of interest from first to last download in the period as reported in Figure 6b: date of first (d_f) and last (d_l) download, stability (sum_o_i), mean number of

downloads per day ($mean_o_i$, see Figure 6b) in the period. Where: $mean_o_i = \frac{sum_o_i}{d_l - d_f}$

The comparison of Figure 6a and 6b gives the evidence of the presence of more explosive content, content used more frequently in less number of days (see Obj-14), while others are more stable such as Obj-1. As to the BPN, it is very important to identify both kinds of object, since the most stable one has to provide a higher quality.



(a)



(b)

Figure 6: Object Stability (a) and mean number of download per day in the stability period (b).

User Fidelity (S_6)

The aim of ***User Fidelity*** measure is to assess in a simple shot which are the most active users in the BPN. Even in this case, a more accurate measure would lead to go into the distribution of their action over a time period such as that reported in Figure 7, where the distribution in the last 10 months is presented for a sample of users with different behaviors (some are returning users, other are sporadic, etc.). We have also analyzed the user behavior in terms of number of active days in the period, the so called *user fidelity*. This simple metric is strongly correlated to the number of content: preferred and uploaded (with a correlation value greater than 75%).

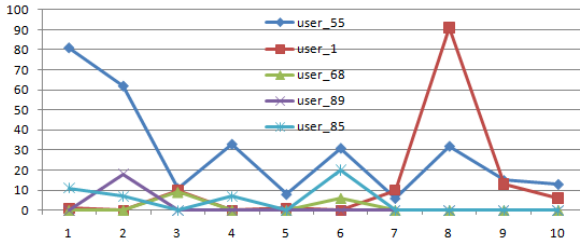


Figure 7: User download distribution in the period.

In order to assess the user behavior, it can be useful to see which is the user activity in the sections. In Mobile Medicine, the 63% of users access the portal to get/play a single content even if they return back later and in the next days as depicted in Figure 7. Only the 14.3% of them stay for two content items, etc.

5 Conclusions

In this paper, a set of direct metrics for the assessment of best practice networks, BPN, has been presented. The analysis of user behavior in BPN is fundamental to tune services and stimulate the network growth. In this paper, a set of metrics related to downloads, kinds of content, different behaviors on PC and mobile, and other aspects implied in BPN has been presented and analyzed with respect to their usage in classical SN. The presented metrics and work have been applied and are currently active to help the growth of Mobile Medicine BPN http://mobmed.axmedis.org. According to a first validation of the proposal, administrative users did find the direct estimation useful. The main discovered issues are related to the different behaviors of users on PC and mobiles, the user preferences with respect to the kind of content, and the identification of the most active users. The proposed metrics can be used as a complementary to the classical SN analysis tools which are more focused on understanding the user relationships.

Acknowledgments:

The authors would like to express their thanks to the many people involved in the validation and to the users of Mobile Medicine BPN; we trust in their understanding for being unable to quote them all in the paper.

References

[1] G. Kruitbosch, F. Nack. Broadcast Yourself on YouTube – Really? HCC'08, October 31, 2008, Vancouver, Canada.

[2] A. Nazir, S. Raza, C. Chuah. Unveiling Facebook: A Measurement Study of Social Network Based Applications. IMC'08, October 20–22, 2008, Vouliagmeni, Greece.

[3] P. Bellini, I. Bruno, A. Fuzier, P. Nesi, M. Paolucci, "Semantic Processing and Management for MultiChannel Cross Media Social Networking", Proceedings of 2009 NEM Submit "Towards Future Media Internet", 28-30 September 2009, Saint Malo France.

[4] P. Bellini, I. Bruno, P. Nesi, "A language and architecture for automating multimedia content production on grid'", Proc. of the IEEE Int. Conf.on Multimedia & Expo (ICME 2006), IEEE Press, Canada, 9-12 July, 2006.

[5] P. Bellini, I. Bruno, P. Nesi, D. Rogai, "'MPEG-21 for the multi-format and multi-channel content production, protection and distribution'", Proc. of the Int. Broadcasting Conf., IBC 2007, Amsterdam, Sept. 2007.

Software Architecture for Web-based Project Management System

Antonio Marcos Neves Esteca, Rogéria Cristiane Gratão de Souza, Adriana Barbosa Santos
UNESP – São Paulo State University, São José do Rio Preto, SP, Brazil
am.esteca@uol.com.br, {rogeria, adriana}@ibilce.unesp.br

Abstract

This article presents software architecture for a web-based system to aid project managing, conceptually founded on guidelines of the Project Management Body of Knowledge (PMBoK) and on ISO/IEC 9126, as well as on the result of an empiric study done in Brazil. Based on these guidelines, this study focused on two different points of view about project management: the view of those who develop software systems to aid management and the view of those who use these systems. The designed software architecture is capable of guiding an incremental development of a quality system that will satisfy today's marketing necessities, principally those of small and medium size enterprises.

1. Introduction

In the last decade, technological advances, market integrations, globalization and tough competition have driven companies to adopt managing strategies that make it possible to increase efficiency and effectiveness of productive processes. Therefore, it became crucial to increase internal communication capacity, to plan activities, to estimate and monitor time and costs, as well as to guarantee task quality [1]. Project management today is a discipline that defines in detail the software and techniques that meet managing requirements, and is becoming increasingly debated in academic and various productive mediums as it enables improving production, reduction of wasted time and money, better decisions, among others that benefit projects [2, 3]. Although there are various definitions for project, this concept can be understood, generally, as a unique process, consisting of a group of coordinated and controlled activities having dates for starting and finishing, to create a product, service or exclusive result, which will satisfy customers [4, 5].

To speed up the project managing process, different computational systems were developed to supply relevant information to managers in a fast, safe and consistent manner. Nevertheless, such systems as are available on the market have a restricted scope of the indispensable managing aspects contemplated by Project Management Body of Knowledge (PMBoK) [5], one of the most reputed guides to project managing. In this context, this work presents software architecture for a supporting web-based system of project managing activities capable of directing the development of a quality product which will contemplate the directives given in the PMBoK [5], the quality characteristics described in the ISO/IEC 9126 [6], as well as certain marketing needs.

2. Software architecture

As time goes by, more evident is the importance of using software systems, in the most varied areas, of an ever increasing size and complexity. The massive use of this technology, as well as the resulting need to obtain efficient, dependable and durable software systems, requires the previous preparation of a target software architecture which will make possible a description of the elements that make up a system, the interactions between those elements and the principles that lead to its evolution [7, 8, 9, 10].

Ref. [11] presents three advantages of designing a software architecture: (a) easier communications between those interested in the project, since a representation of the system makes it easier to be understood; (b) the necessity for a system analyses, since a definition of an architecture requires the study and reflection about the objectives to be reached, which require well planned solutions; and (c) large scale reutilizing, since a defined and documented architecture can be useful in projects which are different, though similar in their desired requirements, thus avoiding reworking.

Moreover, Ref. [12] points out that with a well planned architecture, maintenance is made easier and, consequently, the life cycle of the system may increase considerably.

Software architecture is a relatively new sub-area of software engineering, even though innumerable works have already been published about this subject. Generally, these works could be grouped into three principal categories:

1. Improvements of techniques and methods involved in the project, analyses and validation of the software architectures [13, 14, 15, 16], contributing to the evolution of this sub-area;
2. Disseminating existing concepts, techniques and methods about software architecture, involving didactic works, such as books [17, 18, 11] and scientific articles [19], contributing towards software architecture comprehension and development;
3. Usage of existing knowledge about software architecture

to propose architectures for systems that are being developed, contributing towards a better quality product [20, 21, 22].

This work is in line with the third category as it prioritizes the use of predefined concepts to construct quality architecture, as well as the fact that it directs the development of a system designed to improve the efficiency of managing activities.

3. Methods

The methodological process was divides into two stages:

Funtional requirements elicitation - In this stage was considered two different points of view about project managing activities: the point of view of the software system users and the developers.

Users' point of view: It was obtained based on a pilot survey, done electronically, which contained a convenience sample from the professionals involved in project managing activities from various segments in the São Paulo State, Brazil. The answerers filled in an online questionnaire developed specifically to assess 32 points elaborated considering the nine areas discussed in the PMBoK guide [5]. Importance and the applicability to implement each item in their daily activities were evaluated, using a 5 point Likert scale, with 1 being "Of very little importance" or "Not applicable" and 5 "Very important" or "Very applicable". As a complement, data was collected about the software used by the answerer and their opinion about its main deficiencies. A total of 39 answers were received during the data collecting period which ran from January to March, 2009. Most of them involved small and medium sized enterprises.

Developers' point of view: It involved a review of certain software systems available on the market, resumed in fifteen criteria using PMBoK [5] as reference and other six criteria software quality from ISO/IEC 9126 [6]. In all, nine computational systems were analyzed (four desktop and five web-based systems), which included software mentioned by the project managers that took part of the survey plus some by the authors.

From the analyses of the two points of view, a set of functional requirements was obtained.

Software architecture design – The purpose of this stage was to meet principal marketing necessities and supply a more effective support to project managing, substantiated on the answers to functional requirements.

The architectural project was done in a systematic manner and also considered aspects that contribute for the inclusion of characteristics that promote portability, maintainability and incremental development of the software system to be developed.

4. Requirements Analysis and Design

The functional requirements and the software architecture for web-based project management system are detailed below.

A. Functional requirements

Analyses of users' point of view became possible to prioritize the importance of managing aspects addressed in the PMBoK [5] guide, according to the opinion of the survey participants. Moreover, the survey helped to identify systems available at the market according to the level of popularity and whether or not they included special characteristics, that is, selected systems were those used by project managers or because they offered some type of differential in their available functions . The results are in Figure 1.

To include the developers' point of view as to alignment of the seletected systems to PMBoK guidelines and to the ISO/IEC 9126 software quality norms, a comparative analysis was performed with four desktop and five web-based systems. The results are in Table 1.

A qualification, which represented the frequency each topic was addressed in each system, was attributed to each analyzed topic. The qualifications are shown in Table 1 in discs: as greater the frequency, as larger the portion in black.

Regarding the scope of the systems as to the PMBoK guide, Table 1 shows that some topics, such as number 13, are well covered by most of the systems, while many others are not, such as with topics 1, 7 and 8. On the whole, it can be concluded that none of the analyzed systems provide a complete support to the principle processes described in the PMBoK guide. This may force the users to use a combination of project managing support systems, which would slow their work, to satisfy all their needs.

The assessment of the systems regarding ISO/IEC 9126 norms was done on the contents available to users by the firms that develop such systems. This made it impossible to analyze some attributes; therefore the assessment of the attribute Maintainability was qualified as null for all systems that are not Open Source (except MS Project, which although not being Open Source, allows users to perform evolutional maintenance with Visual Basic) since with the analysis of such systems it was not possible to verify how the maintenance operations were done. System analysis results of the quality norms made it possible to infer that each one of them meet the requirements of a certain set of quality attributes to the detriment of the rest. Due to this, it can be concluded that the quality of the analyzed systems is impaired as on the whole, they do not satisfy quality requirements. Therefore, there is a necessity to design a support for project management activities based on quality criterias, besides obtaining efficient, dependable and customed software systems.

Analyzing and combining information obtained from both views, it was possible to reach the results shown in Table 2, which lists the functional requirements according

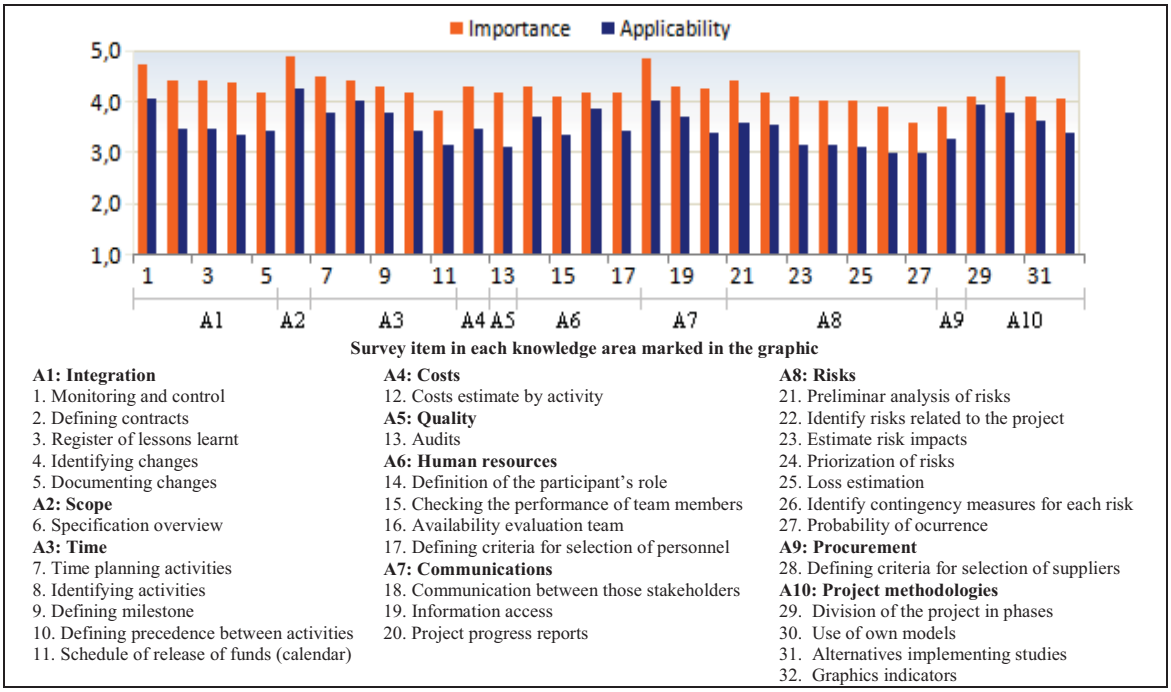


Figure 1. Mean of importance and applicability score for each survey item

Table 1. Comparative Analysis of Support Systems for Project Managing

	Criteria	Desktop systems				Web-based systems				
		MS Project	Project Planner	Mirax	Adapted Tracker	Dot Project	ACE Project	Cooper	Project Open	Net Project
Analysis based on PMBoK guide	1. Register of learnt lessons	○	○	○	○	◔	◔	○	○	○
	2. Storing documents related to the project	○	○	○	○	○	●	●	●	●
	3. Create WBS (Work Breakdown Sctruture)	○	○	●	○	○	○	○	○	●
	4. Plan / control costs	◑	◑	◑	◑	◕	○	◑	◑	◕
	5. Plan / control quality	◑	○	●	○	○	○	○	◔	○
	6. Plan / control manpower	◑	○	◑	○	◕	◑	◑	◕	◕
	7. Control changes	◔	○	◕	○	○	○	○	○	◕
	8. Control / managing risks	◔	◑	●	○	○	○	○	○	◕
	9. Control purchases	◑	◔	◑	◔	◑	○	◔	◑	◔
	10. Communication between those involved	◔	○	○	○	◕	●	◑	●	◑
	11. Control project's progress	●	◕	◕	◔	◑	◑	◑	◑	◑
	12. Control scope	◔	◔	◔	◔	◔	◔	◔	◔	◔
	13. Control time	◕	◕	◕	◕	◑	◕	◕	◑	●
	14. Emit reports	●	◕	◔	◔	●	◑	◑	●	◑
	15. Construct graphics	●	◔	◔	◑	◔	◔	◔	◔	◑

Analysis based on ISO/IEC 9126										
	1. Functionality	◕	◑	◔	◔	◕	◕	◑	◕	●
	2. Reliability	●	◔	◑	◔	◕	◑	◕	◑	◑
	3. Usability	◕	◕	◕	◕	◕	◕	◕	◕	●
	4. Efficiency	◑	●	●	●	●	◑	●	◑	●
	5. Maintainability	◔	○	○	○	◔	○	○	○	○
	6. Portability	◕	◕	◕	◑	◕	◕	◕	◕	◑

Legend: ○ - 0% ; ◔ - 12,5% ; ◔ - 25% ; ◑ - 37,5% ; ◑ - 50% ; ◕ - 62,5% ; ◕ - 75% ; ◕ - 87,5% ; ● - 100%

Table 2. System functional requirements

Knowledge areas	
1. Integration	**6. Human resources**
1.1 Managing contracts with schedules, costs and objectives defined at the beginning of the project	6.1 Managing project team
1.2 Register of events (lessons learnt)	6.1.1 Registering members of the team
1.3 Identifying and documenting changes along the execution of the project	6.1.2 Helping with the enlistment and selection of personnel
	6.1.3 Checking the performance of team members
	6.2 Managing contracts
2. Scope	**7. Communications**
2.1 Defining the scope of the project	7.1 Communication between those interested in the project
2.2 Creating a Work Breakdown Structure (WBS)	7.2 Production and distribution of custom made reports and Warnings
3. Time	7.3 Production of information for mobile equipments
3.1 Include activities	**8. Risks**
3.1.1 Identify activities	8.1 Managing risks
3.1.2 Estimate of starting and ending dates of the activities	8.1.1 Identify risks related to the project
	8.1.2 Estimate risk impacts
3.2 Register finished activities	8.1.3 Identify events related to the probability of an increase of risks
3.3 Warn project delays	8.1.4 Identify contingency measures for each risk
3.4 Development of a schedule of project activities	8.1.5 Monitoring the probability of an increase of risks
3.5 Schedule of release of funds (calendar)	8.2 Creating alerts about the occurrence of risks
4. Costs	**9. Procurement**
4.1 Costs estimate by activity	9.1 Register procurements
4.2 Budgeting	9.2 Register supplier contracts
5. Quality	**Remaining requirements**
5.1 Producing audit reports	Links with other software systems
	Existence of access levels

to the known areas as defined in the PMBoK [5], and even includes identified additional requirements.

B. Software architecture

Aware that the structuring of software architecture that contemplates functional requirements will make viable wide scope, quality and capable of meeting the real necessities of professionals. The first step taken during the execution of the architectural project was to place the software to be developed in its context, that is, external entities were defined with which the system interacts, as well as the nature of the interactions from the composition of the architectural context diagram [1], shown in Figure 2.

The entity **Graphic support** represents a set of libraries and functions capable of offering aid for the construction of all the graphic part of the modules of the system. **Utilities** represent independent System to Aid Project Managing-SAPM - software that can be coupled to provide various services, such as: calculators, spread sheets of calculations, etc. Both supply services to SAPM, while the entities **Managers** and **Project Team** use the services supplied by the system. To present a high level view of the structure of the system, Figure 2 shows also a global architectural structure for the SAPM, where two principal components of the system and the interactions between them are shown. The component **Access Control** offers the access control to the system, permitting listing and authentication of users, create projects and a division of the software into pre-established access levels. This way, certain information of the system can be seen only by authorized users. The component **Project Managing** is responsible for offering the functions inherent to project management.

After creating the global architectural structure, the architectural style for the SAPM was defined, where a style in layers was considered adequate, since it allows an incremental development of the system and offers good portability and maintenance. In Figure 3 the structure established for the SAPM is shown. The highest layer supplies the interface between the system and the user. Immediately below is the business layer project management. The lowest layer provides the graphic

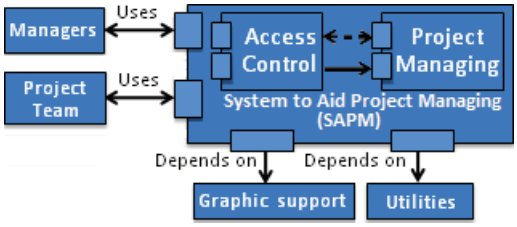


Figure 2. Architectural context diagram and global architectural structure

support and database, where data produced in the upper layer are stored. From identified functional requirements, SAPM architectural components were placed in modules. With that, the architectural diagram was structured, shown in Figure 4 which reveals the components of the system, its modules and the interactions between them. Following, a description of the SAPM modules is presented, which were established for each of the considered components.

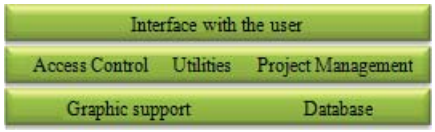


Figure 3. SAPM structure in layer

Access Control:

User identifier: Lists new users and authenticates previously listed users in the system. Should these operations not be successful, the user is notified. Moreover, the *User identifier* receives from the *Project manager* a list of members to be linked to the project and, then, must validate that list, verifying which members are already listed in the system. After validating, the *User identifier* sends a notification about the result of the operation to the *Project manager* and the members that were already listed in the system is sent to the *Content controller* where they will be registered with their respective access levels, defined by the users from the *Project manager*;

Project manager: Presents all the projects to which the users are linked (based on data registered by the *Content controller*), permitting a selection of said projects for later edition. Moreover, from this module users can list new projects. During the listing of a project or at a later stage, users must fill in general data about the project, declare the scope of the project, annex project starting contracts and draw up a list of members (showing members names and respective access levels) who can access/modify project data;

Content controller: Registers the user access level of each project. Also monitors the actions of each user in the whole system, blocking any action that is not permitted according to the users access level. Moreover, this module registers all actions taken by the users in the system.

Managing Project :

Integrated controller of changes: Receives and registers data about changes made in any of the modules of the component **Managing Project**. Moreover, this module receives data from other modules to create baselines;

Participant's manager: Receives data about suppliers and project team members and stores them to create an agenda for contacts and quick access to information about the team members development and suppliers history. Moreover, this module receives contracts which are linked to project team members and suppliers, and sends them to the *Contracts controller*;

Activities manager: Is responsible for all operations related to the activities of the project. From data received from the users, this module executes the following functions: lists activities (identification of activities, starting and finishing dates, description); listing risks (identification, impact estimate, identification of contingency measures, selecting factors capable of producing an increase in the probability of risks happening); registering purchases. Moreover, this module offers the user the following: help to allot resources to activities from information in the *Participant's manager*; automatically monitor the *Participant's manager* and information about the progress of activities to update the probability of risks happening; prepare and automatically update the activity schedule; create a calendar for freeing resources; register contracts related to obtaining resources for the execution of activities; support to create a Work Breakdown Structure; detect activity delays;

Contracts controller: Stores contracts received from the *Participant's manager* and the *Activities manager*, allowing consultations, modifications and exclusions;

Learnt lessons recorder: Registers lessons learnt by the users on executing the project and allows consultations that return to the users the lessons learnt by them in all the projects with which they are or were linked, as he receives all the codes of all those projects from the *Project manager*;

Reports writer: Allows users to create or register records of auditing done. Moreover, receives data from the *Activities manager*, from the *Participant's manager* and from the *Integrated controller of changes* (baselines) and, from them, produces reports according to users options and sends them to the *Information distributer* for distribution;

Information distributer: Uses the data that it receives from the *Participant's manager* to contact stakeholders (via emails or cell phones). This module also receives information from the *Warnings generator* and the *Reports writer*, distributes that information to contacts selected by the user or predefined by the system and notifies the requesting module about the result of the distribution. Moreover, this module offers a chat service to subsidize the effectiveness of the online meetings, as well as an email service;

Warnings generator: Receives data from the *Activities manager* to produce warnings about delays or the probability of an increase of risks, or else from the users to produce a warning about a specific event. The warnings are registered and sent to the *Information distributer.*

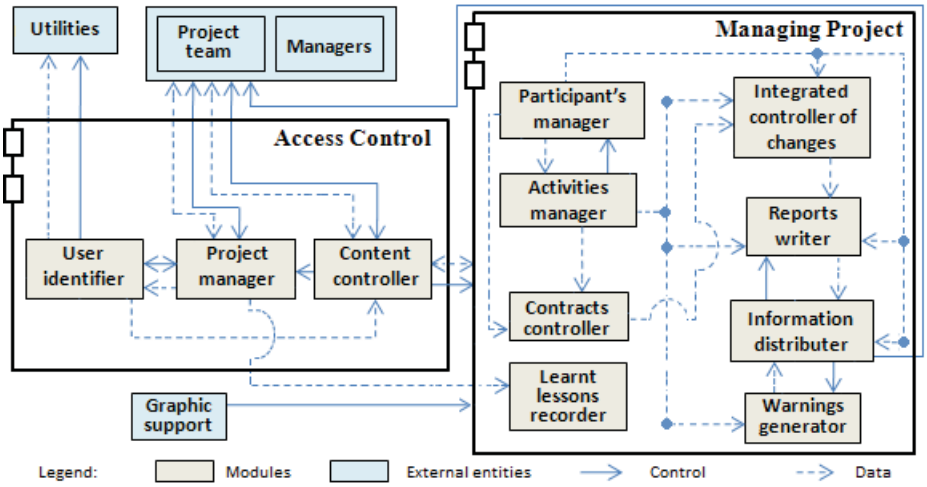


Figure 4. Architectural diagram

5. Final considerations and future works

The proposed software architecture for a SAPM will satisfy existing present day marketing necessities, especially for medium and small enterprises in Brazil. By identifying functional requirements through an empiric study based on guidelines and concepts established in the PMBoK [5] and in ISO/IEC 9126 [6] for the proposed architectural structure, it was possible to fill in the gap still existing in literature about web systems to aid project managing. A valuation of the web-based system quality characteristics such as portability, maintainability and efficiency, besides others, are being prioritized in the conception of SAPM as the objective is not only to construct quality architecture, but also the web system. Therefore, the future works agenda foresees the development of a SAPM prototype which will satisfy web-based systems quality criterions. The exploratory evolutionary model [23] was adopted so as to make viable an assessment by market professionals when each significant part of the system is concluded.

Acknowledgement

This research was partially supported by The State of São Paulo Research Foundation under project number 2008/07094-2.

References

[1] R.S. Pressman, *Software Engineering: A Practitioner's Approach*, 6th ed., Philadelphia: McGraw-Hill, 2005.

[2] D.I. Cleland and L.R. Ireland, *Project Manager's Portable Handbook*, New York: McGraw-Hill, 2002.

[3] A. Murphy and A. Ledwith, "Project management tools and techniques in high-technology SMEs," *Management Research News*, v. 30, n. 2, p. 153-166, 2007.

[4] International Standard Organization. ISO 10006: *Quality management - Guidelines to quality in project management*, s.1.p. ISO, 1997.

[5] PMBoK, *A Guide to the Project Management Body of Knowledge*, 3rd ed., Project Management Institute – PMI, 2004.

[6] International Standard Organization. ISO/IEC 9126: Software engineering – Product quality – Part 1: Quality model, 2001.

[7] A.M.N. Esteca, R.C.G. Souza, and A. B. Santos. "*Gerência de projetos: arquitetura de software concebida pela visão empírica das diretrizes do PMBoK e da NBR 9126", in Proc. of the SIMPEP 16th Simpósio de Engenharia de Produção,* Bauru, Brazil, 2009 (in portuguese).

[8] J.N. Buxton and B. Randell, *Software Engineering Techniques,* Technical report, Rome: NATO Science Committee, 1970.

[9] D. Garlan and M. Shaw, *Software Architecture: Perspectives on an Emerging Discipline,* New Jersey: Prentice-Hall, 1996.

[10] IEEE and ISO/IEC, "Systems and Software Engineering - Recommended Practice for Architectural Description of Software-Intensive Systems," *ISO/IEC 42010 IEEE Std 1471-2000 First edition*, c1–24, July 2007.

[11] L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice,* 2nd ed., Boston: Addison-Wesley Professional, 2003.

[12] I. Jacobson, M. Griss, and P. Jonsson, *Software Reuse: Architecture, Process and Organization for Business Success,* Utah: Addison-Wesley, 1997.

[13] C. Lung and M. Zaman, "Application of Design Combinatorial Theory to Scenario-Based Software Architecture Analysis", in *Proc. of the 17th International Conference on Software Engineering and Knowledge Engineering (SEKE)*, Republic of China, pp.418-423, 2005.

[14] M.D. Rice and S.B. Seidman, "An approach to architectural analysis and testing," in *Proc. of the third international workshop on Software architecture*, Orlando: Foundations of Software Engineering, pp. 121 – 123, 1998.

[15] V. Ambriola and A. Kmiecik, "Architectural transformations," in *Proc. of the 14th international conference on Software engineering and knowledge engineering (SEKE)*, Italy, vol. 27, pp. 275-278, 2002.

[16] S.J. Carrière, S. Woods, and R. Kazman, "Software Architectural Transformation," in *Sixth Working Conference on Reverse Engineering (WCRE)*, Georgia, pp.13, 1999.

[17] J. Bosch, *Design and Use of Software Architectures*, New York: Addison-Wesley, 2000.

[18] D. Garlan, "Software Architecture: A Roadmap", *The Future of Software Engineering*, A. Finkelstein, ed., ACM Press, 2000.

[19] J. Bosch, W.M. Gentleman, C. Hofmeister, J. Kuusela, "Software Architecture: System Design, Development and Maintenance," in *IFIP 17th World Computer Congress - TC2 Stream / 3rd IEEE/IFIP Conference on Software Architecture (WICSA3)*, Canada, 2002.

[20] S. Dumbrava, D. Panescu, and M. Costin, "A Three-layer Software Architecture for Manufacturing Activity Control in ERP Concept," in *Proc. International Conference on Computer Systems and Technologies and Workshop for PhD students in computing*, Bulgaria, 2005.

[21] W. Ma, D. Tran, and D.Sharma, "On Extendable Software Architecture for Spam Email Filtering," in *IAENG International Journal of Computer Science, Republic of China*, vol. 34, pp. 133-139, 2007.

[22] R.K. Ganti *et. al.* "Satire: a software architecture for smart attire," in *Proc. of the 4th International Conference on Mobile Systems, Applications and Services*, Suíça, pp. 110-123, 2006.

[23] I. Sommerville, *Software Enginnering,* 8th ed., Essex: Addison-Wesley, 2007.

Software Clustering Using Drools Tool

Nenad Rogulja[1], Branko Milosavljević[2], Zora Konjović[3]

Lanaco, *Banja Luka, Bosnia and Herzegovina*[1]
Faculty of technical sciences, *Novi Sad, Serbia*[2,3]
designment@gmail.com[1], {mbranko[2], ftn_zora[3]}@uns.ac.rs

Abstract

Maintaining the software systems is a complex task due to the complexity of the contemporary software systems caused by the size and frequent changes that are credibly expressed only by the source code. One solution for this problem is source code clustering, which makes it possible to generalize about some features which are cluttered in the source code. Earlier, there weren't any attempts to analyze those complexities by using rule engine alone. This paper evaluates possibility of using rule engine as a framework for various software engineering tasks. Specifically, the paper demonstrates that it is possible to perform software module clustering using rule engine.

1. Introduction

High level abstractions, various models – views through which we can look at software, helps us to cope with complexities of the systems. However, often we say that "the truth is in the code" [15], since high level abstractions aren't always synchronized with the code. By performing software clustering, we will regain more quality abstract views on the software.

Field of search based software engineering (SBSE) is in the focus in this paper. This is active area of the research in the computer science. And SBSE is used in many applications (module clusterization, code generation, project management…). According to M. Harman, software engineering (SE) is better suited for search based optimization then all other areas of engineering [18].

1.1. Motivation

Suppose we could create homogenous framework which will be able to perform static and dynamic software analysis, to propose abstractions to the software architect [3], which are needed for maintaining or upgrading software (e.g. to extract sequence diagrams [13]). Besides that, framework would be able to perform other very useful tasks: automatic (unit) testing, checking for security issues [22] etc. Even more, through the same framework, one would be able to perform refactoring, cleaning (e.g. code clone elimination [4], cycle detection) and other transformations (i.e. upgrading software to use newer library versions, to generate code, etc).

Framework should have ability to easily configure, tune; to perform (semi) automatic benchmarks of different configurations (varying parameters, algorithms, etc.).

Crucial problem here is clustering. Modern software systems are large and complex, which makes problems in understanding their structure. Main reason for this complexity lies in the fact that source code contains lots of entities (class, modules etc.), which are mutually dependent in interactive manner (procedure calls, variables referencing…). On the other side, these systems are repeatedly changed which could be faithfully identified only from source code (specification often doesn't follow source code changes). Clustering techniques applied to the software source code is an approach for solving this problem.

Main purpose of software clustering is to create view on software using clusters which contains minimal coupling between modules from any two different clusters, and maximum cohesion between modules inside a single cluster.

2. Related work

There is a very helpful paper for getting a big picture of vast applicability of those techniques [4], although the paper doesn't put rule engine into perspective.

Following Mancoridis et al. who first suggested the search based approach to module clustering [17], this paper takes this, search based, approach. They have used hill climbing, but also have tried using genetic algorithm and simulated annealing. And they got best results using hill climbing. However, it may be possible to define different crossover operation for use in genetic algorithm [19], and that may create feasible approach. Although, performing GA by exploiting rule engine is a problem. Therefore, we use local search technique. Search algorithms need metric which will direct the process, and we follow proven practice of using cohesion and coupling measurement.

For found software clusterization of *k* clusters and *n* modules, for measuring its clustering quality, we use Modularization Quality (MQ) [9]:

$$MQ = \sum_{i=1}^{k} CF_i \,, \qquad (1)$$

cluster factor CF_i is calculated with the following formula:

$$CF_i = \frac{\mu_i}{\mu_i + \frac{1}{2}\sum_{\substack{j=1 \\ j \neq i}}^{k} \left(\varepsilon_{i,j} + \varepsilon_{j,i}\right)} \qquad (2)$$

In the formula (2), μ_i is the total intr**a**-edges weight (connections between modules inside one cluster) of the i-th cluster, and $\varepsilon_{i,j}$ is the sum of int**e**r-edges weights (connections between modules from different clusters) between i-th and j-th cluster. In rare cases, when i-th cluster doesn't have any intr**a**-edges, CF_i equals zero.

It is possible to extract hierarchical clusters using the search based approach (by recursively clustering found clusters as modules). Other approach to achieve this is by using agglomerative approach [21]. However, that approach is harder to align with the rule engine requirements.

Instead of searching for specific clusterization, usually we need to allow user to prioritize some aspects (say relations between some modules) over others (which makes hypothesized entities), and then try to rebuild those aspects (by clustering) from the source code (i.e. concrete entities). This technique is called Software Reflexion [3], and can be considered as (semiautomatic) extension of a simpler clustering approach.

Recent paper, [5], shows that we can use rule engine for model completion, without a significant performance penalty, despite previous work arguing this would be impossible.

There are some advanced ideas about integrating whole SE analysis process into a single framework. One of the best examples are Symphony [1] – based on Bachaus tool [6], and DMS [8]. They might even use rule centric approach in their subcomponents, although it isn't used at higher levels. There are many works already done on creating tools and techniques to perform some kind of transformations that are specified with rules [8], [11] and [20].

Along using search clustering approach, there are other methods: hierarchical agglomerative clustering [2], [3], [21], or more analytical one, that uses spectral analysis [23].

Recently, we can see more works on including rule engines in solving specific software engineering problems [7].

3. Contribution

This paper brings a few contributions.

- In this paper we present that a modified Hill Climbing algorithm (HC) is suitable for use in rule-based clustering process. The algorithm has following features: a) heuristically determines initial solution; b) after reaching local optima, several random moves (escapes) are performed.
- This paper shows that it is possible to use rules engines for clustering (specifically, software module clustering).
- The ideas presented here, create base for building a framework for analysis of complex software and for comparing and evaluating various techniques in the search based software engineering field.

3.1. Change in the hill climbing algorithm

Hill climbing algorithm is very often used in search based clustering. For example, it can be used in the Bunch tool [9]. Local search, specifically hill climbing, often converge toward local optima, and the common solution to this problem is to start search from some random solution, and to perform it at least twice, and then to choose better one. [14]

We took another approach. We do not use random starting (because we believe that it will take considerable time even to begin to form meaningful clusters). Initial solution is heuristically created around modules with more incident edges. However, we do perform another randomness which (we hope) will prevent us from choosing local, instead global, optima. Surprisingly often, a random move from local optima toward its neighborhood will lead to a satisfactory algorithm (even if better techniques are known – but which could complicate rule engine implementation) [14].

So, we perform several random moves (escapes), when no better solution is found for some time. By looking at the search landscape in [24], we could notice that using MQ metric and MDG, there are usually at most two points around which the best found clusterization are grouped. We recognize the fact that there is a need for even more detailed search landscape analysis; although our testing shows that random escapes can overstep the local optimum surrounding.

Note that Drools can perform Tabu search, [14] (for example, so it does not repeat move with module which was recently used). Implemented search process, especially with this additional method, avoids to be trapped inside local optima. This approach, when we start from certain non-random solution, makes it possible to use once found solution as initial one for other, maybe in some way changed, search process.

3.2. Clustering with rule engine

Often, when we develop solutions for software engineering problems (reverse engineering, transforming one model to another), we tend to identify rules which we can apply to achieve desirable goal. Even if we have rules which can bring decisions of what is needed to be done (or how to transform one model to another), we usually need, as an input, some sort of a model (often with graph representation). For example, after we decide which viewpoint we need to use to create desirable view, we acquire a graph by clustering, and by extracting needed features from the source code. This was never done by the rule engine itself. This paper tries to bridge this gap by showing that it is possible to create rule engine solution that performs clustering algorithms.

Rule engines often implements Rete algorithm, for efficient large scale pattern matching, which is used in determining which rules are matched by their antecedent part [10]. Rete algorithm best works in situations where there is a small change in the facts, because it effectively caches previous values of the facts and distributes knowledge (in tree-like structures).

It is known that while performing local search algorithm [14], we evaluate (all) neighbour solutions, and choose one of them. Then we apply only the move which will transfer focus from the current to the locally chosen solution. That move is usually quite simple, that makes only small change (comparing to the complexity of the current solution state). Exactly this favouring of a small (incremental) change makes rules engine (with Rete algorithm) a good choice for executing clustering process.

This paper proposes using rule engine for performing graph clustering. In this case, we use JBoss Rules engine (Drools, [12]) for software clustering, and to perform analysis that might follow.

3.3. Framework

Now that we are capable of performing various analysis in a single environment, with the help of many rules organized in the various groups, we could create extendable framework for performing many software engineering related tasks.

What new will bring such a framework? First, many benefits are inherently brought from the very structure of a good rule engine:

- centralized rules repository,
- declarative programming,
- scalability,
- easy to change and configure…

Second, JBoss Rules extends this list even more providing components such as: Solver, Fusion – to perform Complex Event Processing (CEP, useful for dynamic analysis), diagramming and linking with process engine/workflow using Rule Flow, domain specific language (DSL), fuzzy logic… It could be said that we only specialize one general framework into a more specific one, software engineering framework.

4. Details

Clustering is done in two phases:

- heuristically create initial solution, and
- current solution is improved by searching for optimal one, regarding criterion of quality of a clusterization.

To be able to use rule engine to perform local search we use `Drools`, and its `Drools Solver` component.

In the Figure 1, we show process of software clustering (including code guidance). `Workbench`

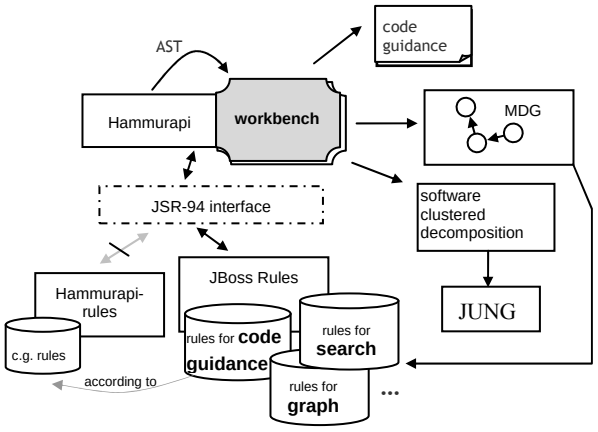


Figure 1. Framework for SBSE using rules engine

component is the developed application. We have replaced `Hammurapi rules` engine [16] with `JBoss rules`, because we need tight integration with `Solver`, but on the other side we want to use lexical analyses from `Hammurapi` (which uses `Mesopotamia` tool).

Main advantage of this approach is that we could add various rules for completely different types of analysis (along those necessary for clusterization itself, we could have rules for security, bad code practice - "code smell", code clone detection…). So effectively we combine all these analysis into one maintainable framework system. Analysis could be performed on different level of granularities. Some analysis could inspect found structural components (clusters), and relationships between them, some could focus on lower level – *abstract syntax tree* (AST) elements, or make hybrid approach. Another benefit is that it is possible to easily perform incremental analysis (say, to put end user into a loop of making decisions, pausing search process, as needed), than to continuously monitor system in its developing phase (but also when is in its exploitation). Most of needed requirements are by their nature declarative, based on event-condition-action (ECA) paradigm.

In this paper we, for coupling and cohesion, consider counting of dependencies (count of different method calls of referenced class, including constructors and calls of parent class methods when redefining them; than existence of local variables, parameters, import declarations etc. which are of the referenced class type) from one toward other module. What makes impact on dependency between two modules is guided with principal goal to create clusterization which will ease software maintenance and understanding complex software: changes in one module should easily propagate inside a cluster, and in the same time propagation out of a cluster should be minimal.

4.1. Model

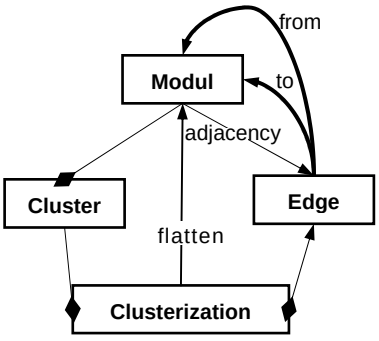


Figure 2. MDG representation model

For representing structure of software system (modules and interconnections between them) we use concept of module dependency graph (MDG) [9], [17]. From MDG concept, we define OO model of its representation. Its simplified schema is shown on the figure 2, *Module* is a class of Java program. *Edge* is representation of dependency of source module (*from*) to destination module (*to*). Dependency has quantitative measure, weight, that represents total count of different method called of the destination module, redefinition of methods of destination module inside source module, etc. *Cluster* contains modules. And *Clusterization* represents one solution of software clustering. Aggregations (diamond edges) represents associations which are persisted in the MDG XML file, while *flatten* association keep references on those modules which will be moved around while searching (after shrinking the search space). And *adjacency* association keeps all edges which are adjacent to a module (irrelevantly of its direction, mainly for creating initial clusterization).

By using Hammurapi tool for static code analysis we can get lexical tokens, AST elements, from the source code of the analyzed application as facts in the `Drools` rule based system. Than, we use rules that recognize dependency relation from the AST, so that we can generate arcs in the MDG (for example, by recognizing methods calls of some class).

4.2. Creating initial solution

For initial clusterization we use heuristic with basic idea to form initial clusters around module with large count of neighboring modules. Modules are sorted in descending order with total count of all adjacent edges. In this ordered list we specify border between big and small modules. Big modules make new clusters if there isn't an edge with greater weight than, with parameter specified, threshold, which is adjacent to some module that is already in a cluster. Small modules are usually added to some cluster in which it has adjacent module.

4.3. Search process

After forming an initial solution, we apply a local search procedure to try to continually improve it. While searching we keep track of the best solution found during the process.

In the case of software clustering, relation of adjacency (neighboring) of solutions is defined in this way: two solutions are adjacent if they differ only in belonging to the cluster of exactly one module. Having that in mind, move transfers a module from current cluster to another existing or a new cluster. It is important that move doesn't make big change in the set of facts, in other words, we must be able to isolate

influence of a single move on the set of evaluated rules and facts.

For search process we use `Drools Solver` (rule engine extension) as planner application (scheduler). Move in `Drools Solver`-tool is defined with Java class. As evaluation function is calculated for every move (which is described with `Drools` rules), for specific module (fact) we move, it must be very efficient. For an illustration, on listing 1 is shown a rule for calculating impact of a module on the total measure of clusterization quality.

```
declare Pair
  intra : double
  inter : double
end

rule "one module contribution"
  lock-on-active true
  when
    $m : Module( $c : cluster );
    $p : Pair() from accumulate(
      $e : Edge(eval(from==$m) || eval(to==$m),
                $w:weight, $to:to),
       init( Pair p = new Pair(); p.setIntra(0d);
             p.setInter(0d);),
       action(
        if($e.getFrom().getCluster()!=$c ||
           $to.getCluster()!=$c){
             p.setInter(p.getInter()+$w);
        }else{
          p.setIntra(p.getIntra()+$w);
        }),
       reverse(
        if($e.getFrom().getCluster()!=$c ||
           $to.getCluster()!=$c){
             p.setInter(p.getInter()-$w);
        }else{
          p.setIntra(p.getIntra()-$w);
        }),
       result( p )
      );
  then
   modify($m){
    setIntra($p.getIntra()),
    setInter($p.getInter());
   }
end
```

Listing 1. DRL rule for calculating one module change contribution to the cluster factor value

In this listing we see use of facts like `Module` and `Edge`, which are integral part of MDG representation. Also, for the purpose of calculating clusterization factors CF_i, we create `Pair` helper fact. It is created for each module which is moved for storing needed values. Every module makes contribution to its cluster intra and inter cluster weights (μ_i,$\varepsilon_{i,j}$ or $\varepsilon_{j,i}$ in formula 2) depending of its relations to other modules (in the same, and other cluster, respectively). Module's moves, change its associated cluster, what is here seen through change in `e.getFrom().getCluster()` and `e.getTo().getCluster()` for every edge related to the moved module. When contribution to the cluster factor value from a single module is calculated, it is stored in a module fact (class instance) itself. Then, with other rule, we calculate clustering factor based on those stored values (also using *accumulate* construct). For each move, only one module is changed, and all other values are "cached" in the Rete network of the rule engine, and only changes propagate through network, avoiding excessive operations.

This listing could be read as: "For each module which is moved to some cluster, calculate Pair values from all adjacent edges depending on: is the module adjacent over that edge in the same cluster or not. Store calculated pair values into the module for following cluster factor calculation."

4.4. Shrinking the search space

One efficient way to reduce search space size (which is bounded up with factorial of the number of modules) is to identify *omnipresent* modules [9]. Those are modules which are either quite often referred (*library, suppliers*), or which refers many others modules (*client, driver*). In addition, we also reduce search space size by eliminating moves which will transfer modules which are adjacent only to a single module (except *omnipresent* modules), as they always need to be in the same cluster as that (adjacent) module. When we transfer such module, we also transfer all modules that have this module as its only neighbour. We also can specify which classes won't be processed.

4.5. An example

As an example for demonstrating software clusterization we have chosen open source code of *Byte Code Engineering Library* (BCEL, used for analysis and manipulation of Java class files for *byte code enrichment*). This source contains about 300 modules.

Visualization is implemented using *Java Universal Network/Graph framework* (JUNG). We implement collapsing of cluster's nodes into a sub-graph, or group identified clusters into bunches. In graphical representation, weight of an edge represent intensity of a dependency between modules, and the colours of the circles represents cluster to which module belongs. On the figure 3, we have initial MDG graph for BCEL code; and figure 4 shows detected clusters. They are represented with a polygon; polygon is larger as it contains more modules inside (we could drill down into its content).

5. Conclusion

In this paper it is shown that it is possible to perform clustering with rule engine. Hereto, we did that,

primarily using `Drools Solver`, in a way that good performance can be achieved. When a module is moved from one cluster to another, it fires only rules that are directly affected by the movement and only for the changed facts.

This makes it possible to create homogenous rule based framework for various software engineering tasks.

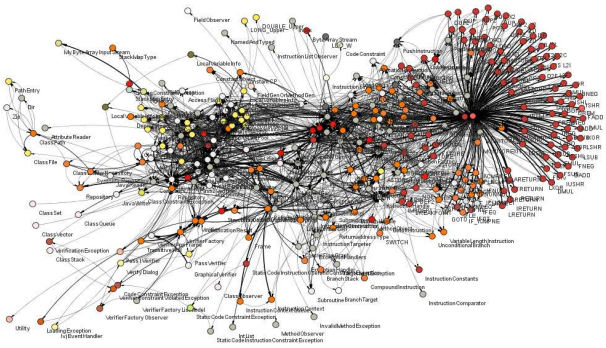


Figure 3. MDG for the BCEL source code library

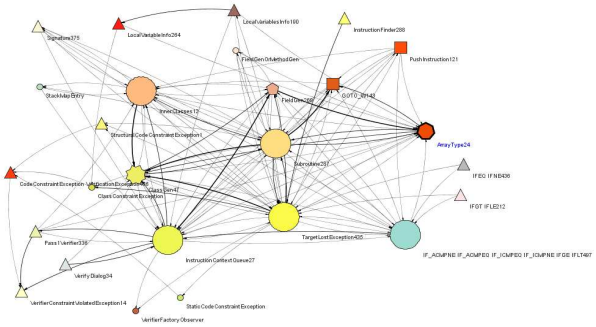


Figure 4. BCEL library clusterisation

References

[1] A. von Deursen, C. Hofmeister, R. Koschke, L. Moonen, C. Riva, Symphony: view-driven software architecture reconstruction,pp.122, Fourth Working IEEE/IFIP Conference on Software Architecture (WICSA'04), 2004

[2] N. Anquetil, C. Fourrier, T.C. Lethbridge, Experiments with clustering as a software remodularization method, Sixth Working Conference on Reverse Engineering, 1999

[3] A. Christl, R. Koschke, M.A. Storey, Equipping the reflexion method with automated clustering, wcre, pp.89-98, 12th Working Conference on Reverse Engineering, 2005

[4] G.C. Harman, M. Di Penta, New frontiers of reverse engineering, Future of Software Engineering, 2007

[5] J.M. Wright, J.B. Dietrich, Non-monotonic model completion in web application engineering, aswec, pp.45-54, 21st Australian Software Engineering Conference, 2010

[6] A.Raza, G.Vogel, E.Plodereder. Bauhaus–a tool suite for program analysis and reverse engineering. In Proceedings of the 11th Ada-Europe International Conference on Reliable Software Technologies, LNCS 4006, Portugal, 2006.

[7] M. Zygmunt, M. Budyn, Managing software complexity with business rules, pp.534-537, International Conference on Complex, Intelligent and Software Intensive Systems, 2010

[8] I.D. Baxter, DMS - program transformations for practical scalable software evolution, Proceedings of the International Workshop on Principles of Software Evolution, Orlando, 2002

[9] B.S. Mitchell, S.Mancoridis, On the automatic modularization of software systems using the Bunch tool, IEEE Transactions on Software Engineering, vol. 32, no. 3, pp. 193-208, Mar. 2006, doi:10.1109/TSE.2006.31

[10] C.L. Forgy, Rete: A fast algorithm for the many pattern/many object pattern match problem, Artificial Intelligence, vol. 19, pp. 17-37, 1982.

[11] E. Visser, A Survey of Strategies in Rule-Based Program Transformation Systems, Technical Report UU-CS-2005-022, Institute of Inf. and Computing Sc.,Utrecht University, 2005

[12] Michal Bali, Drools JBoss Rules 5.0 developer's guide, Packt Publishing, 2009

[13] P. Dugerdil, J. Repond, Automatic generation of abstract views for legacy software comprehension, Proceedings of the 3rd India software engineering conference, 2010

[14] H.R. Lourenco, O. Martin, T. Stutzle. Iterated local search. In Handbook of Metaheuristics, F. Glover and G. K., Eds. Int. Series in Operations Research & Management Science, vol. 57. Kluwer Academic Publishers, 2002

[15] R. Koschke, What architects should know about reverse engineering and rengineering, wicsa, pp.4-10, Fifth Working IEEE/IFIP Conference on Software Architecture (WICSA'05)

[16] *Hammurapi rules User Guide*, http://www.hammurapi.biz/products/hammurapi-rules/

[17] S. Mancoridis, B.S. Mitchell, C. Rorres, Y.F. Chen, and E. R. Gansner. Using automatic clustering to produce high-level system organizations of source code. In International Workshop on Program Comprehension (IWPC'98), pages 45–53, California, USA, 1998. IEEE Computer Society Press.

[18] Mark Harman, Why the virtual nature of software makes it ideal for search based optimization, In Fundamental approaches to software engineering, Springer, 2010

[19] M. Harman, R. Hierons, M. Proctor, A new representation and crossover operator for search-based optimization of software modularization, GECCO, 2002

[20] J.R. Cordy, T.R. Dean, A.J. Malton, K.A. Schneider, Software Engineering by source transformation-experience with TXL, scam, pp.0170, First IEEE International Workshop on Source Code Analysis and Manipulation, 2001

[21] O. Maqbool and H. Babri. Hierarchical clustering for software architecture recovery. IEEE Transactions on Software Engineering, 33(11):759–780, 2007.

[22] K. Sohr, B. Berger, Towards Architecture-Centric Security Analysis of Software, 2nd International Symposium on Engineering Secure Software and Systems, 2010., in press

[23] A. Shokoufandeh, S. Mancoridis, M. Maycock, Applying spectral methods to software clustering, wcre, pp.0003, Ninth Working Conference on Reverse Engineering, 2002

[24] B.S. Mitchell, S. Mancoridis, "Modeling the Search Landscape of Metaheuristic Software Clustering Algorithms," Proc. Genetic and Evolutionary Computation Conf., 2003.

A model proposal for program comprehension

Francesca Arcelli
arcelli@disco.unimib.it

Marco Zanoni
marco.zanoni
@disco.unimib.it

Riccardo Porrini
r.porrini@gmail.com

Mattia Vivanti
mattia.vivanti
@gmail.com

Università degli Studi di Milano Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione
Viale Sarca, 336 20126 Milano, Italy

Abstract

One of the principal task in the context of reverse engineering is related to program comprehension activity, that can be performed through different techniques, tools and strategies.

To better understand and exploit the knowledge necessary to comprehend an existing system different models and metamodels are necessary to represent the information extracted at different levels of abstraction. In this paper we describe the model we have defined in our Marple (Metrics and Architectures Reconstruction PLug-in for Eclipse) project, whose main objectives are related to design pattern detection and to software architecture reconstruction of an existing system to support program comprehension.

We introduce the architecture of the model we have proposed outlining its extensibility features and we describe the implementation of the model through the Eclipse Modeling Framework (EMF). We are interested in using this model as the only data repository for any step of elaboration on the system.

1 Introduction

A field of software engineering which is getting more and more importance is reverse engineering [6, 15]. One of its main purposes is to support program comprehension in order to achieve as much as possible information about the analyzed system. In this way it is possible to identify the fundamental components of the system and to heavily simplify each task that has to be performed on the system, from bug fixing to system modernization, since the system is not seen as a "black box" but as a "white box", made up of smaller components that interact with each other to achieve the program's goals.

In the literature many approaches for program comprehension have been proposed, some using only static analysis (as for example [13] [11], [17]) and others using both static and dynamic analysis (see for example [9], [18]), each one providing useful information to better perform the reverse engineering tasks. For these reasons it could be interesting and promising to be able to use and integrate the information gained through different sources and approaches.

To represent this useful information we need a model able to provide full abstraction of the underlying information that can easily be exported and exchanged between the different components or actors involved, using for example an XML serialization.

In our project called MARPLE (Metrics and Architectures Reconstruction PLug-in for Eclipse)[4], we implemented a model for the representation of a software system. In this project according to this model, we aim to perform design pattern detection (DPD) and software architecture reconstruction (SAR) of Java programs using different modules that interact with the underlying model in order to contribute to the given goals.

Through design pattern detection it is possible to gain a better comprehension of a software system and on what kinds of problems have been addressed during the development of the system itself. The presence of design patterns (DPs) can be considered as an indicator of a good software design, as DPs are reusable for their self definition. Moreover, they are very important during the re-documentation process, in particular when the documentation is very poor, incomplete or not up-to-date. In order to detect them we adopt an approach based on design pattern decomposition [2], by obtaining from the source code the sub-components that can be considered indicators of the presence of patterns in the source code. These sub-components (or micro-architectures), which we call *Basic Elements* are placed at an intermediate level of abstraction between source code

and design patterns and have to be found and used to describe and then to detect design patterns.

The main objective of SAR is to abstract from the analyzed system's details in order to obtain general views and diagrams with different metrics associated with them. The extraction of such data helps the engineers to have a global understanding of the system and of its architecture.

During the design of the model we considered as a principal requirement the possibility to use and integrate the multiple sources of information together, as for example those provided through SAR and DPD. In this way the model we have proposed is characterized by good extensibility features, that will permit us to add further information gained through the existing or future elaboration steps. For these reasons, we are going to use the model as the only data repository for any step of elaboration on the system. In this way we aim to provide a model driven approach for program comprehension to better support software evolution. Moreover, through the model extensibility features described in the paper, we are able to consider most of the object oriented languages.

The paper is organized in the following Sections. In Section 2 some related works are described. In Section 3 we introduce the architecture of the model. In Section 4 we describe the extensibility features of our model. In Section 5 we describe our implementation through the Eclipse Modeling Framework. Finally, in Section 6 we conclude and shortly outline some future directions of this research.

2 Related works

An interesting and well known model which provides a language independent representation of object oriented source code is Famix[8], used in the Famoos project (`http://www.iam.unibe.ch/~famoos/`) for the reengineering of object-oriented legacy systems. We have not used a model like this since it is too detailed for our purposes, so we decided to define our model, as simple as possible. For example a difference between the models is that Famix defines several entities representing different types of association, as method invocation, inheritance and variable access; into our model these types of information are covered by Basic Elements. Another difference is that in Famix the logical and physical entities are seen in a homogeneous way (belonging to the same hierarchy), and into our model we keep them separated. Finally, the entire Famix specification is tuned for the use with the CDIF serialization, and it is mainly available for the Smalltalk platform; we need instead a model that integrates with Eclipse, being available for Java, and supporting the EMF technology and XMI serialization.

Our work can be also compared to the Knowledge Discovery Meta-model, developed in the context of the ADM[16] project from OMG(`http://adm.omg.org`). The KDM is an extensive model representing a software system. The model recently had an implementation through the MoDisco Eclipse project (`http://www.eclipse.org/gmt/modisco/`), that aims at the integration of the ADM process into the Eclipse framework. Currently MoDisco is not fully available to be experimented. Our work is not directly related to KDM or MoDisco but we aim to integrate our knowlegde and plugins and cooperate with the Modisco team development group.

Other related works in this area can be found in [7, 14, 12, 10].

3 Model Architecture

The model has been designed in order to save, handle and manage the most relevant and necessary information for program comprehension; by analyzing a software project we can extract three fundamental types of information:

- the logical structure of the project under consideration: as we work on object-oriented languages, the entities are types, methods, attributes, etc and the relations between them;
- the physical structure of the project: for program comprehension we are interested in knowing the project's folders and files (containing code or not);
- the meta-data that we'll associate to logical and physical entities; in MARPLE we actually consider metrics and Basic Elements (described in the introduction).

For program comprehension purposes we need a model representing a software system that uses these three information types.

The structural requirement features of the model are the following:

flexibility: capacity to represent software systems written in different object oriented programming languages;

extensibility: capacity of the model to be easily adapted to new needs; for example in order to handle the definition of language plug-ins it must allow extensions able to manage language specific entities and properties; another example is the specification of new meta-data types, like the results of DPD;

optimization: capacity of the model implementation to provide an easier information recovery mechanism, since all the SAR and DPD operations will retrieve data from the model; some metrics will be calculated directly on the model (for example the number

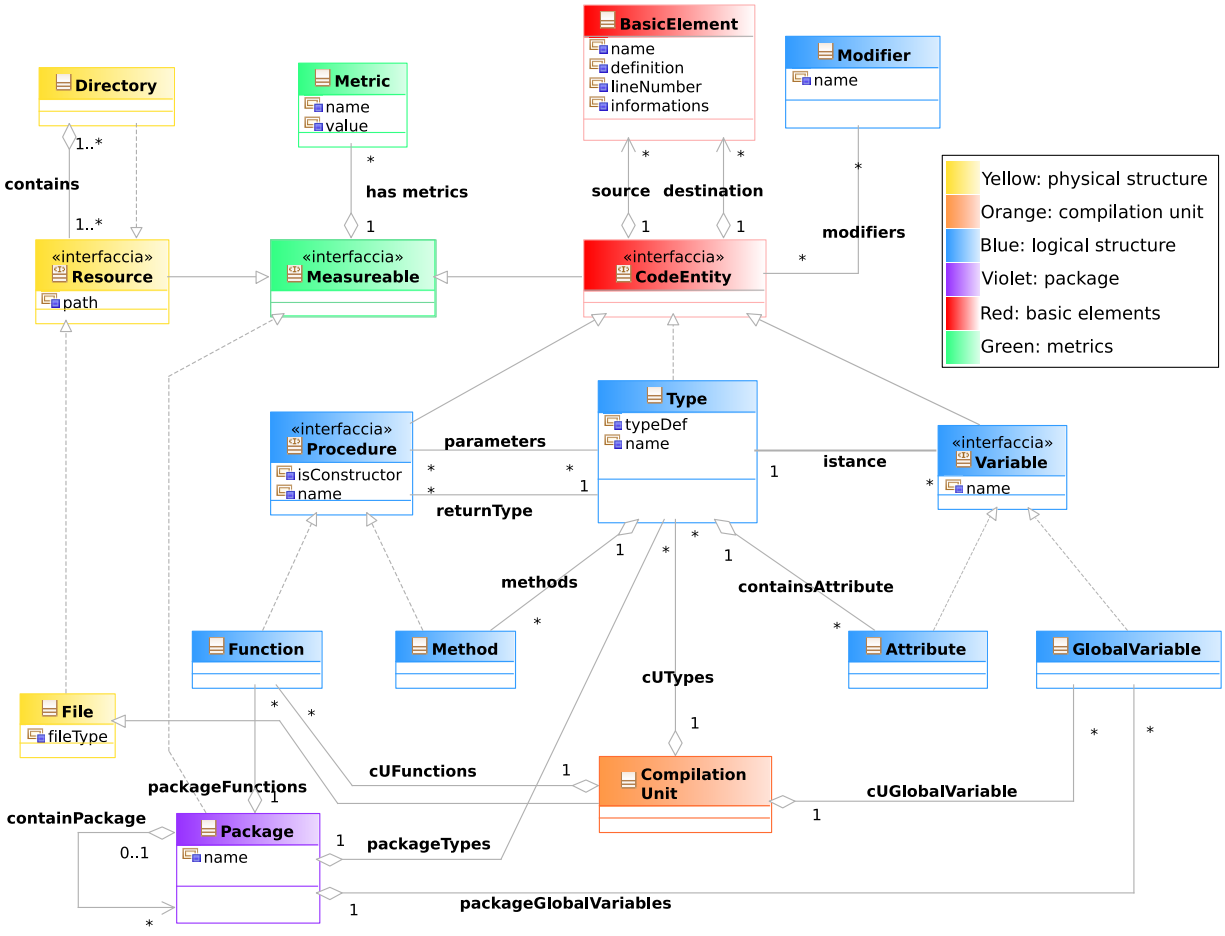


Figure 1. Model Architecture

of classes), so it will be used also by the metrics collector module of Marple.

In Figure 1 the UML definition of the model architecture is shown.

The entities have been colored in order to identify their nature.

The entities representing a project's physical structure are colored in yellow:

- `Resource` is an interface that is implemented by all the physical entities and is characterized by a path;
- `Directory` is defined as a `Resource` that can contain many `Resources`;
- `File` is a simple physical `Resource`.

`CompilationUnit` is a file containing source code; it is a container of logical entities of the project, so it represents the link between the physical and the logical parts of the project. For this reason it is colored in a different way (using orange).

The logical entities are colored in blue and represent the entities of a generic object oriented programming language: types, behavioural entities and structural entities and the relations between them. The following logical entities have been defined:

- `Type` represents the definition of a class, interface or abstract class in source code. The distinction between class and interface is made through its `typeDef` attribute.
- `Procedure` is an interface that represents the definition in source code of a behavioural abstraction, which is characterized by input parameters and by one return type; both are represented in the model with the two associations between `Procedure` and `Type`. `Procedure` is extended by:

  - `Method` represents a method of an object oriented language: it is a procedure belonging to a class, so is linked to a class using a composition;
  - `Function` represents a global function, like the ones found in C++.

- `Variable` is an interface that represents the definition in the source code of a structural entity; it is defined as an instance of a type, using an association between `Type` and `Variable`. It has two implementors:
  - `Attribute` represents a class attribute, that is a part of the type's state, and it is linked to its type using a composition;
  - `GlobalVariable` represents the definition in source code of a variable which is globally accessible and with a lifetime equal to the running system's lifetime.

`Type`, `Function` and `GlobalVariable` are contained in both a physical container, `CompilationUnit`, and in a logic one, `Package`. A `Package` represents a code entity collection defined in different source files, like Java packages and C++ namespaces. As `Package` is only a logical container, it is colored using violet.

All logical entities (`Type`, `Variable` and `Function`) implement an interface named `CodeEntity`, which represents all the source code entities associated with a modifier and all the source code entities associable to a Basic Element. We represented the modifiers using an external class and not an attribute of the interface because different metrics and Basic Elements may need to select a `CodeEntity` basing on its modifiers; through this architecture we can easily read the association with no need to check each entity attribute. As the modifiers belong to the logical structure of the project, `Modifier` is colored in blue. `CodeEntity` is instead colored in red, the same color used for `BasicElement`, because it represents the entities the Basic Elements refer to.

`BasicElement` models the instance of a Basic Element found in source code, which has a name and is mainly characterized by a source `CodeEntity` and a destination `CodeEntity`, both implemented through aggregations.

Finally metrics are represented by the `Metric` class and they are linked, through an aggregation, to the entities they belong to. All the entities associable to a metric implement the interface `Measurable`: the physical entities (`Resource`), the logical entities (`CodeEntity`) and `Package`. `Metric` and `Measurable`, that are the classes enabling the metrics measurement, in Figure 1 are colored in green.

4 Model Extensibility

The model has been defined not only considering its integration in MARPLE, but also to provide a basis for program analysis and comprehension. Thus we defined some model's entities to be generic: `Metric`, `BasicElement`. These entities are characterized by an attribute representing their name, i.e. "LOC" for the "lines of code" metric, or "FieldAccess" for the "Field Access" Basic Element. This design choice is an extensibility feature of the model, since it is possible to specify new entities without changing the model design.

We defined three interfaces providing top level abstractions to our model: `Resource`, `CodeEntity` and `Measurable`. They represent respectively physical entities, logical entities and entities which is possible to associate metrics to. These abstractions provide a further extensibility feature: it will be possible to integrate new metadata in the model simply by linking them to the appropriate interface using associations. This integration will not affect the model's architecture, but it will only extend its definition.

We also defined `Modifier` entities as general as possible, in order to delegate their definitions to the implementation. We defined `GlobalVariable` and `Function` entities in order to allow global existence respectively for variables and functions, as they exist for example in C++. `Package` has been defined with an aggregation on itself, allowing nested containers like C++ namespaces. These entities provide flexibility to our model, covering the definition of C++ source code and of most of the other object oriented languages.

We defined only bidirectional associations in our model. This characteristic allows to quickly and easily access information required by SAR and DPD tasks. Moreover the containment relationships define a hierarchy on model's entities. This allows recovering entities from the model simply by navigating their containers, and not necessarily the whole model. All these features make the model optimized for information recovery.

5 Model Implementation

The model has been implemented through the Eclipse Modeling Framework (EMF)[1] because it provides an interesting environment for model-driven development, available for the Eclipse Platform.

EMF is a framework suitable for the definition of a model since it allows to define the model through both graphical and XML interfaces. The framework provides an interesting feature that allows the developer to automatically generate the model source code; it provides also a way to perform queries against the generated model. Finally

EMF exposes an utility for the model serialization in a standard XML or XMI format, making the model exchangeable between research groups.

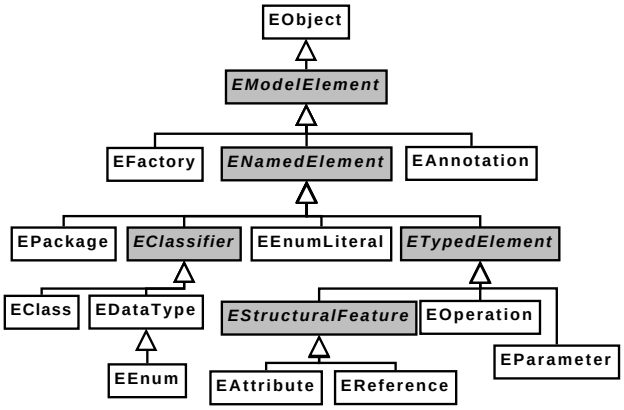


Figure 2. Ecore Hierarchy

The entire plug-in is based on the `Ecore` class hierarchy (see Figure 2) that our model extends. Through the `Ecore` class hierarchy it is possible to define all we need to implement our model (i.e. `EClass` represents an abstraction of a class, and `EOperation` represents an abstraction of a class' method).

Since the model architecture is defined in UML it is necessary to redefine it using the `Ecore` hierarchy. There are two main issues in the redefining process:

Aggregations: `Ecore` does not implement aggregations, but only containment relationships;

Associations: only monodirectional and not bidirectional associations can be set.

These issues can be easily overcome: the first one by implementing aggregations using compositions if it is possible or associations if not; the second one by defining two inverse associations for each bidirectional one.

The source code generation from an `Ecore` model is very simple to be performed; the resulting source code is perfectly integrated in the `Ecore` hierarchy (all classes descend from `EObject`). This integration allows to full control the run-time side with the reflective API; it gives also the possibility to run queries against model's instances and to serialize them in XML or XMI format.

Other than the model's classes, the generated source code contains other two important classes:

`ModelPackage`**:** it contains the java definitions of the model;

`ModelFactory`**:** it implements the *Abstract Factory* design pattern, providing methods for run-time creation of model's classes contained in the `ModelPackage`.

Using `ModelFactory` we can instantiate single model's classes; these classes are composed together in many EMF `Resource` classes to form the model instance. These resources are grouped in an EMF `ResourceSet` class: the model instance.

Each instance of `Resource` in a `ResourceSet` can be serialized on a distinct file; the format we chose is XMI as it is the standard for model interchange. We decided to split our model into these XMI files:

package: it contains the project logical structure (package, classes, etc);

root: it contains the project physical structure (directories and compilation units);

metrics: it contains all the metrics calculated on the project;

bes: it contains all the Basic Elements within the project;

modifiers: it contains all the source code modifiers.

The model has been integrated in MARPLE. In order to preserve model consistency and avoid information duplications, an interface layer has been created. This layer stands between the analysis module and the model, and decodes all the information retrieved, exploiting the *Factory Method* design pattern. Each information retrieved is decoded by passing it to a concrete handler class, which calls `ModelFactory` in order to create the right instance and to put it in the right `Resource`. For each kind of information a different *Factory Method* has been designed, keeping the layer's architecture as modular as possible, in order to allow new information detection by simply creating the handler family that will decode it.

6 Conclusions and future works

Program comprehension is an important field of reverse engineering. In order to better support program comprehension we need all the possible information about the system to be exposed to the user performing the reverse engineering tasks.

We think that the model we have defined is sufficiently simple and complete to fit these requirements. The model can be extended easily to contain the output of different analysis, because it keeps concerns separated but linked together.

Now our work will be concentrated on exploiting this extensibility feature in order to model and compare the results of design pattern detection and integrate them in the model (as we already proposed in [3]). In future works we will try to expand the capabilities of the model definition in order

to be able to represent the evolution of a software system, keeping track of different system versions.

Combining this aspect to the results of analysis modules like the DPD one, it will be possible to see also the evolution of any type of analysis, enabling new points of view of the comprehension of a software system.

Moreover we are interested to use our model based approach to unify software and data reverse engineering [5], considering the important software assets of databases.

Another perspective is to integrate the MoDisco framework into our tool, taking advantage of the already implemented features, and being able to use standardized and widely accepted exchangeble models.

References

[1] Eclipse modeling framework. Web site. `http://www.eclipse.org/emf/`.

[2] F. Arcelli, S. Masiero, C. Raibulet, and F. Tisato. A comparison of reverse engineering tools based on design pattern decomposition. In *ASWEC '05: Proceedings of the 2005 Australian conference on Software Engineering*, pages 262–269, Washington, DC, USA, 2005. IEEE Computer Society.

[3] F. Arcelli, C. Tosi, and M. Zanoni. A benchmark proposal for design pattern detection. In *FAMOOSr 2008: Proceedings of 2nd Workshop on FAMIX and Moose in Reengineering*, 2008.

[4] F. Arcelli, C. Tosi, M. Zanoni, and S. Maggioni. The marple project - a tool for design pattern detection and software architecture reconstruction. In *Proceedings of the WASDeTT Workshop, co-located event with ECOOP 2008 Conference*, Cyprus, 2008.

[5] F. Arcelli, G. Viscusi, and M. Zanoni. Unifying software and data reverse engineering. In *Proceedings of ICSOFT 2010, Athens*, July 2010.

[6] E. J. Chikofsky and J. H. C. II. Reverse engineering and design recovery: A taxonomy. *IEEE Software*, 7(1):13–17, 1990.

[7] T. R. Dean, A. J. Malton, and R. C. Holt. Union schemas as a basis for a C++ extractor. In *Proceedings of the Eighth Working Conference on Reverse Engineering*, Stuttgart, Germany, October 2001. IEEE Computer Society. www.cppx.com.

[8] S. Demeyer, E. Tichelaar, and P. Steyaert. Famix 2.0 - the famoos information exchange model. Technical report, University of Berne, 1999.

[9] L. Explorers. Codelogic. Web site. `http://www.logicexplorers.com/index.html`.

[10] R. Ferenc, A. Beszedes, M. Tarkiainen, and T. Gyimothy. Columbus - reverse engineering tool and schema for C++. In *Proceedings of the 18th International Conference on Software Maintenance*, pages 172–181, Montreal, Canada, October 2002.

[11] IBM. Sa4j: Structural analysis for java. Web site, 2004. `http://www.alphaworks.ibm.com/tech/sa4j`.

[12] N. A. Kraft, B. A. Malloy, and J. F. Power. An infrastructure to support interoperability in reverse engineering. *Information and Software Technology*, 49(3):292–307, 2007.

[13] M. Lanza, S. Ducasse, H. Gall, and M. Pinzger. Codecrawler: an information visualization tool for program comprehension. In *ICSE '05: Proceedings of the 27th international conference on Software engineering*, pages 672–673, New York, NY, USA, 2005. ACM.

[14] T. C. Lethbridge, S. Tichelaar, and E. Pl"odereder. The dagstuhl middle metamodel: A schema for reverse engineering. *Electronic Notes in Theoretical Computer Science*, 94:7–18, 2004.

[15] H. A. Müller, J. H. Jahnke, D. B. Smith, M.-A. Storey, S. R. Tilley, and K. Wong. Reverse engineering: a roadmap. In *ICSE '00: Proceedings of the Conference on The Future of Software Engineering*, pages 47–60, New York, NY, USA, 2000. ACM.

[16] P. Newcomb. Architecture-driven modernization (adm). In *WCRE '05: IEEE Proceedings of the 12th Working Conference on Reverse Engineering*, page 237, Washington, DC, USA, 2005.

[17] L. O'Brien, D. Smith, and G. Lewis. Supporting migration to services using software architecture reconstruction. *Software Technology and Engineering Practice, 2005. 13th IEEE International Workshop on*, pages 81–91, 24-25 Sept. 2005.

[18] A. Zaidman, A. Hamou-Lhadj, O. Greevy, and D. Röthlisberger. Workshop on program comprehension through dynamic analysis (pcoda'08). In *Working Conference on Reverse Engineering*, volume 0, pages 345–346, Antwerp, Belgium, 2008. IEEE Computer Society.

Software Process Reuse Based on Knowledge and Process Components

André Luiz Peron Martins Lanna[1], Carlos Alberto Marques Pietrobon[1,2]
PUC-MG – Pontificia Universidade Catolica de Minas Gerais[1]
UFOP – Universidade Federal de Ouro Preto[2]
E-mail: andre.lanna@gmail.com, capietro@pucminas.br

Abstract

Software process reuse can improve the software quality. This paper presents the use of concepts from Component-Based Development (CBD) on software processes as a mean to improve process reuse. This reuse method considers the technical, organizational and human aspects of a software process beyond the knowledge acquired on past executions.

Acknowledgement

Special thanks to FAPEMIG - Foundation for Research Support of Minas Gerais, for the support given to this work.

1 Introduction

Nowadays software product quality is an interest topic by academic area and software organizations. In the software reuse approach (typically) the artifacts previously developed, tested and sucessfully integrated in other applications are used as groundwork for the development of new software. Its use offers advantages like costs reduction and fast product releases [1, 2]. Other approaches focus on the development process[1]. On these approaches it's expected that processes having quality generate products with quality, what indicates a clear relation between them.[1].

However define a software process is a difficult task to be performed since the technical, organizational and human factors of a process should be regarded and stowed in a harmonic way.[3]. Due this difficulty the process definition is usually performed by experienced people, with deep knowledge of organizational processes in an attempt to increase the quality[4].

Typically both approaches currently occur separately and their results can be leveraged if they are considered together [5]. Osterweil in [6]indicates an intersection between software and processes when points that similar to software products it's possible to define requirements, design, implement and evolve software processes. This similarity allows the adaptation of techniques usually applied on software artifacts to process context, including those related to software reuse.

This work aims on reusing software processes (with their inherent knowledge) during new processes definition in order to increase their maturity, stability and quality. Based on the traditional reuse technique called Component-Based Development (CBD) it was defined a reusable structure to represent a software process in its different aspects (technical, organizational and human), at different details levels (granularity) beyond storing the knowledge acquired in its past executions. It allows new processes will be defined in minor time and costs, beyond they will obtain the benefits from improvements made in the reused processes, what impacts positively the quality of the developed software.

This paper is structured as follows: section 2 presents the background and related work. Section 3 describes the purpose of this work: it is defined a process model, the reusable structure for representing processes and storing their knowledge and finally how such structures are described in order to facilitate their reuse. Section 4 presents an example of reusing software processes. Conclusions about this work are carried out on section 5.

2 Background and related works

A software process is a set of policies, technologies and activities executed by different roles at an organization to produce or maintain a software product and artifacts[1, 7]. It has to be considered under technical, organizational and human aspects[3]. The technical aspect comprehends everything related to representing a software process like artifacts, technologies, milestones and so on. It consists of smaller units such as sub-processes, activities and tasks so that a unit inside this classification may be formed based on units that are one level below. The organizational aspect de-

[1]On this text the terms "software process", "software development process" and "process" have the same meaning.

scribes concepts defined by the organization such as roles played by stakeholders, constraints imposed to the project, organizational policies and goals. The human aspect is related to the human resources and skills that an organization disposes for a process. Despite it may vary substantially between organizations, it was initially defined on [3] as individual skills, technical knowledge and experiences.

Another important process's characteristic is the fact that during its execution it uses and produces much knowledge, which can be of different types like descriptions, cases, lessons learned, ideas, doubts among others[8]. The knowledge related to a process has great value for the organization because it can improves the software quality to the extent that members increase their understanding about the processes. The importance and benefits of using knowledge acquired on past processes during definition and tailoring a new software process are shown on [9] and [10].

Component-based development is a software reuse method on which the reusable structures are called software components. A software component is defined as a composition unit that represents pieces of software with requirements, interfaces and context dependencies well-defined[11, 12]. The term "unit of composition" indicates that components can be grouped in order to form bigger and more complex units. Their interfaces specify the services provided by the component and perform communications with other components through the resources exchange (required or results from services from other components). Lastly a software component does not exist by itself. It is necessary satisfy a set of conditions, fact known as context dependence, what allows a component be reused successfully in its totality (never partially).

Choose the adequate component to reuse is not always a simple task. Even if there are documentation and description related to the component, they can be insufficient for understanding the component's services. To solve this problem proposals like [13, 12] try to improve the expressiveness of a component's description using concepts that are well-defined and understood on a context. These works presents their own ontologies, each one defining their terms and meanings for representing components' semantic knowledge. Therefore the use of ontologies seems to be a viable solution to describe software components.

3 Process definition through process components reuse

To reuse software processes three steps are necessary: a) define a software process model, b)define a reusable structure to represent a software process according to the process model and c)describe this reusable structure on a high expressiveness level to facilitate its understanding and reuse.

3.1 Software process model

Although technical, organizational and human aspects describe different characteristics of a process, they must be considered as being dependents due the fact that decisions in one aspect may influence the others two. Moreover the technical aspect determines the detail level for representing a software process, because it can represents an entire process or parts of it (activities and tasks). Therefore when defining a software process these three aspects should be considered together in order to harmonize them.

Knowledge is also part of software process but it is considered separately from the process aspects. The knowledge about a software process is used to describe or register situations occured in any of its past executions, which may include one, two or all process aspects. Each knowledge record in a software process is called knowledge item and can be of different types. This work considers the knowledge types defined by [8], but allows the organizations define their own types of knowledge.

Based on these characteristics it was defined as a software process model the set of technical, organizational and human aspects, beyond all knowledge items, as figure 1 shows. The arrow indicates that knowledge items include the 3 aspects of a software process.

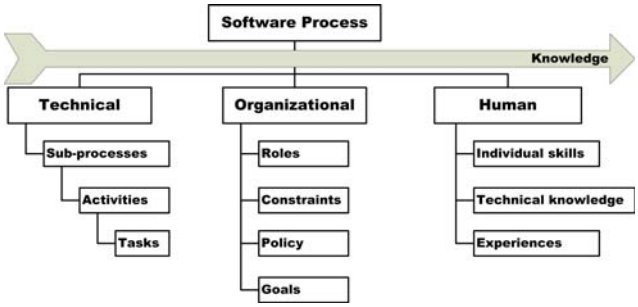


Figure 1. Software process model.

3.2 Software Process Component

A software process component is a reusable structure for representing a software process adherents to the process model. It is defined as being: a) units of composition with interfaces and context dependencies well defined, b) able to represent the three aspects of a software process independent from the details level (processes, activities or tasks) and c)able to store the knowledge items related to the process inside the proper component.

The term "units of composition" indicates that a new software process will be defined by grouping processes components. Their interfaces have the role of realizing the connections and communications between the components

by exchanging the artifacts produced and consumed by processes. Context dependencies are understood as the process requirements that have to be attended to allow the execution of services provided by process components. One process component has variable granularity depending on the details level for its process representation (if it is a process, activity or a task). Therefore the size of a process component is defined by the technical aspect of the software process.

Finally the process component must be able to store various knowledge items acquired over past executions of the process represented. As established by the process model, the process component stores the knowledge within the component but independent from the aspects.

For component representation it's used the XML. It is flexible enough to change or create the constituents of every aspect or knowledge, in order to suit the specifics characteristics of an organization. The four large constituents groups of a process are represented by the tags *TechnicalAspect*, *OrganizationalAspect*, *HumanAspect* and *KnowledgeSet*, each one consisting of tags for descriptions of that aspect. These tags has *type* as an attribute which assumes one value defined in the process model. Organizations can define their own elements for the aspects or knowledge through new values that the attribute *type* will receive during a process definition.

Interfaces of the processes components must be well defined to allow connections and communications between software processes, informing the services and the resources required and provided by them. Such services are the processes, activities and tasks defined for the component while the resources are the software artifacts produced and required during its execution. The XML structure of a process component contains specifics tags for representing such interfaces, services and artifacts.

3.3 Software process component description

In order to facilitate and improve the process reuse, it's necessary describe the process component and the elements that constitute the software process. The component description is realized by assigning keywords on 23 metadata (like component name, creation date among others) grouped on 5 categories. More details can be found on [14].

However the description of process aspects and knowledge items shall be conducted based on ontologies able to define meaningful concepts for the process context. Each software process can be reused in different environments, so the aspects and knowledge descriptions should not be restricted to only a certain ontology, but it must absorb the terms defined by any ontologies adopted by the process. For each item that composes the software process there is a description performed within the XML tag that represents the process item and it is filled with a text explaining its significance. This text uses the terms and elements defined by the ontologies adopted by the process.

```
<?xml version="1.0" encoding="UTF-8"?>

<ProcessComponent date="20/03/2010" author="Andre" collaborator="Pietrobon">
    <TechnicalAspect granularity="task">
        <TechnicalItem id="1" name="Identify_Requirements" type="task">
            The purpose of this task is to identify and capture functional and
            non-functional requirements for the system. It is performed by an
            Analyst
        </TechnicalItem>
        <TechnicalItem id="2" name="Glossary" type="Artifact">
            Record the terms used on software process.
        </TechnicalItem>
        <TechnicalItem id="7" name="Use-Case_Model" type="Artifact">
            This artifact presents an overview of the intended behavior of the
            system.
        </TechnicalItem>
    </TechnicalAspect>
    <OrganizationalAspect>
        <OrganizationItem id="5" name="Analyst" type="Role">
            Identify and understand problems and opportunities.
        </OrganizationItem>
    </OrganizationalAspect>
    <KnowledgeSet>
        <KnowledgeItem id="6" type="description">
            The analyst should gather information from domain, identify the
            system requirements, capture the actors and use cases scenarios.
        </KnowledgeItem>
    </KnowledgeSet>
    <interface id="10" name="Identify_Requirements" providedRequired="Provided">
        <service id="1" name="Identify_Requirements"
         numberInputArtifacts="1" numberOutputArtifacts="1">
            <artifact id="2" name="Glossary" inputOutput="input"/>
            <artifact id="7" name="Use-Case_Model" type="Artifact"/>
        </service>
    </interface>
</ProcessComponent>
```

Figure 2. *Identify Requirements* **task component.**

The description of elements that constitute the process component interface can also be performed through concepts defined by the ontology. Therefore these interfaces are no longer described based on keywords, but in well-defined and known concepts about the software process. This type of description facilitates and increases the understanding of the services provided by the component.

4 Sample process reuse

Considers two typical tasks of a software process: identify requirements and detailing use-cases scenarios. Together they can form an activity to detail software requirements. Considers also that the following terms are defined by the ontology adopted by the processes: a) *Use-Case model:* overview of the intended behavior of the system, b) *Requirements:* needs to be provided or opportunities to be realized and c) *Scenario:* context in which the use case will be performed.

The components "*identify requirements*" and "*detail use-case scenarios*" tasks are shown on figures 2 and 3. Note that the terms defined by the ontology are present in the items descriptions that compose the software processes, contemplating different aspects, knowledges and interfaces. For example the term *Requirements* appears in the description of techinical items in figures 3 and 4, while the term *scenario* appears in the description of an item of knowledge.

```
<?xml version="1.0" encoding="UTF-8"?>

<ProcessComponent date="15/02/2010" author="Celia" collaborator="Andre">
    <TechnicalAspect granularity="task">
        <TechnicalItem id="1" name="Detail_Use-Case_Scenarios" type="task">
            The purpose of this task is to describe use-case scenarios in
            sufficient detail to validate understanding of the requirements.
        </TechnicalItem>
        <TechnicalItem id="3" name="Use-Case_Model" type="Artifact">
            This artifact presents an overview of the intended behavior of the
            system.
        </TechnicalItem>
    </TechnicalAspect>
    <OrganizationalAspect>
        <OrganizationItem id="2" name="Analyst" type="Role">
            Identify and understand problems and opportunities.
        </OrganizationItem>
    </OrganizationalAspect>
    <HumamAspect>
        <humanItem id="4" name="Analyst_Skills" type="skill">
            Good communication skills, verbally and in writing. Ability to
            collaborate effectively with the team.
        </humanItem>
    </HumamAspect>
    <KnowledgeSet>
        <KnowledgeItem id="6" type="description">
            The analyst should gather information from domain, identify the
            system requirements, capture the actors and use cases scenarios.
        </KnowledgeItem>
        <KnowledgeItem id="4" type="decision">
            Andre will perform the ANALYST role in this process.
        </KnowledgeItem>
    </KnowledgeSet>
    <interface id="8" name="Detail_Use-Case_Scenarios" providedRequired="Required">
        <service id="5" name="Detail_use-case"
          numberInputArtifacts="1" numberOutputArtifacts="1">
            <artifact id="3" name="Use-Case_Model" inputOutput="input"/>
            <artifact id="3" name="Use-Case_Model" inputOutput="output"/>
        </service>
    </interface>
</ProcessComponent>
```

Figure 3. *Detail Use-Case Scenarios* **task component.**

In addition, in the interfaces of the components appears the term *Use Case model* as input and output artifacts. This allows that these processes can be described and retrieved through these terms, which have a specific meaning to the context of the software in question. Moreover, the use of the term *Use Case model* ensures that the artifacts present in the interfaces are exactly the same for both components.

Interfaces are used to connect and establish communications between components. So the interface *Identify requirementes* provides a service which output artifact is a *Use-case Model*. Moreover, the interface *Detail Use-Case Scenarios* has a service that requires the same artifact *Use-Case model* (as defined by the ontology). Based on these informations in the interfaces, the two components can be grouped to form the activity of *detail requirements*.

5 Conclusions

The use of concepts defined by ontologies in process descriptions and knowledge showed be a valid strategy for increasing software quality. Experiments *in-vitro* presented that the understanding of stakeholders about the software process has increased, as the processes used terms and concepts well defined and known by them. This led to errors reduction during the execution of software processes.

On the other hand the software reuse seems be effective for the definition of software processes. Reusing software processes helped the task of define a software process because it's based on processes known by the organization.

However the great advantage of this approach appeared when the two approaches were considered together. It was possible to define software processes in minor time and costs with higher quality and stability, due the knowledge provided a support to development staff. The next step of this work is to assess the degree of improvement achieved by the unification of the two proposals for software quality and compare the results when they are considered alone.

References

[1] I. Sommerville, *Software Engineering*. São Paulo, SP, Brazil: Pearson Addison-Wesley, 8 ed., 2007.

[2] J. Li, A. Gupta, J. Arvid, Borretzen, and R. Conradi, "The empirical studies on quality benefits of reusing software components," vol. 2, pp. 399 –402, july 2007.

[3] A. L. P. M. Lanna and C. A. M. Pietrobon, "Proposal for reusing software processes using knowledge and visualization," in *VI Brazilian Workshop of Thesis and Dissertations on Software Quality*, (Florianópolis, SC, Brazil), 2008.

[4] A. S. Barreto, L. G. P. Murta, and A. R. Rocha, "Software process definition: a reuse-based approach," in *XXXIV CLEI - Conferencia Latinoamericana de Informatica*, 2008.

[5] D. Lucrédio, K. dos Santos Brito, A. Alvaro, V. C. Garcia, E. S. de Almeida, R. P. de Mattos Fortes, and S. L. Meira, "Software reuse: The brazilian industry scenario," *Journal of Systems and Software*, vol. 81, no. 6, pp. 996 – 1013, 2008.

[6] L. J. Osterweil, "Software processes are software too, revisited: An invited talk on the most influential paper of icse 9," *International Conference on Software Engineering*, 1997.

[7] A. Fuggetta, "Software process: a roadmap," in *ICSE '00: Proceedings of the Conference on The Future of Software Engineering*, (New York, NY, USA), pp. 25–34, ACM, 2000.

[8] M. A. Montoni, "Knowledge acquisition: An application on development process," Master's thesis, Federal University of Rio de Janeiro, 2003.

[9] P. Xu, "Knowledge support in software process tailoring," (Big Island, HI, United states), pp. 87 –, 2005.

[10] L. d. M. S. Borges and R. d. A. Falbo, "Managing software process knowledge," in *CSITeA'02*, pp. 227–232, 2002.

[11] C. Szyperski, *Component Software: Beyond Object-Oriented Programming*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., 2002.

[12] H. H. Wang and J. Sun, "A semantic web environment for components," *The Knowledge Engineering Review*, 2009.

[13] P. Ramadour and C. Cauvet, "An ontology-based reuse approach for information systems engineering," *Signal Image Technology and Internet Based Systems*, 2008.

[14] A. L. P. M. Lanna, "Software process reuse based on process components and knowledge," Master's thesis, Pontificia Universidade Católica de Minas Gerais, 2009.

Ontological Filters for Slow Intelligence Systems

Shi-Kuo Chang [1], Emilio Zegarra[1], Chia-Chun Shih[2],
Ting-Chun Peng[2], Francesco Colace [3] and Massimo De Santo[3]

[1]Department of Computer Science, University of Pittsburgh, USA
[2]Institute for Information Industry, Taiwan
[3]Department of Information and Electrical Engineering, University of Salerno, Italy

{chang, ezegarra}@cs.pitt.edu, {chiachun, markpeng}@iii.org.tw, {fcolace, desanto}@unisa.it

Abstract: *Slow Intelligence Systems are general-purpose systems characterized by being able to improve performance over time through a process involving enumeration, propagation, adaptation, elimination and concentration. The transform functions of the building blocks for Slow Intelligence Systems are knowledge transforms. When the knowledge base is an ontology, the transforms are ontological transforms. A particularly important ontological transform is the ontological filter, which can be used both as the Eliminator and as the Concentrator. The Propagator can also use ontological filtering to selectively send messages to other Slow Intelligence Systems. In this paper the lightweight-plus ontology formalism will be introduced and adopted. We can apply the ontological filters to product and service customization, and to the detection of hot topics and trends on the Internet.*

1. Introduction to Slow Intelligence Systems

We will first introduce the concept of Slow Intelligence and present a general framework for designing and specifying Slow Intelligence Systems (SIS). We view SISs as general-purpose systems characterized by being able to improve performance over time through a process involving *enumeration*, *propagation*, *adaptation*, *elimination* and *concentration* [1]. An SIS continuously learns, searches for new solutions and propagates and shares its experience with other peers. An SIS differs from expert systems in that the learning is implicit and not always obvious. An SIS seems to be a slow learner because it analyzes the environmental changes and carefully and gradually absorbs that into its knowledge base while maintaining synergy with the environment.

We will start with an example. A rose plant in many ways behaves like an SIS. In the spring it will grow many branches (enumeration), which interact with the surrounding environment (propagation) to adjust their orientation (adaptation). In late spring some branches will die (elimination), and among the remaining ones only a handful will be provided with most of the resources to grow really tall (concentration). The rose plant not only manages to survive by reacting quickly to environmental changes (quick decision cycle) but also ensures the survival of its species by preserving the strongest branches (slow decision cycle). These characteristics are the trademarks of an SIS.

We can distinguish Natural SISs and Artificial SISs, in that Natural SISs such as ecosystems [2] are created and perfected by Nature over time, and Artificial SISs are intentionally created by Man, usually within a relatively short span of time. The characteristics of Slow Intelligence manifest themselves in both Natural SISs and Artificial SISs. There are a large number of intelligent systems, quasi-intelligent systems and semi-intelligent systems that are "slow". Sports training systems [3], [4], online learning systems [5], intelligent travel planning systems [6], intelligent health restoration systems [7], 3D reconstruction systems [8], classifier selection systems [9], [10], and so on all exhibit some characteristics of SISs.

2. Characteristics of Slow Intelligence Systems

An SIS is a system that *(i) solves problems by trying different solutions, (ii) is context-aware to adapt to different situations and to propagate knowledge, and (iii) may not perform well in the short run but continuously learns to improve its performance over time* [1].

The above definition is from the behavioral viewpoint (the black box approach). One of our goals is to understand the characteristics and structure of both natural and artificial SISs (the white box approach).

SISs typically exhibit the following characteristics:

Enumeration: In problem solving, different solutions are enumerated until the appropriate solution or solutions can be found.

Propagation: The system is aware of its environment and constantly exchanges information with the environment. Through this constant information exchange, one SIS can also propagate information to other (logically or physically adjacent) SISs.

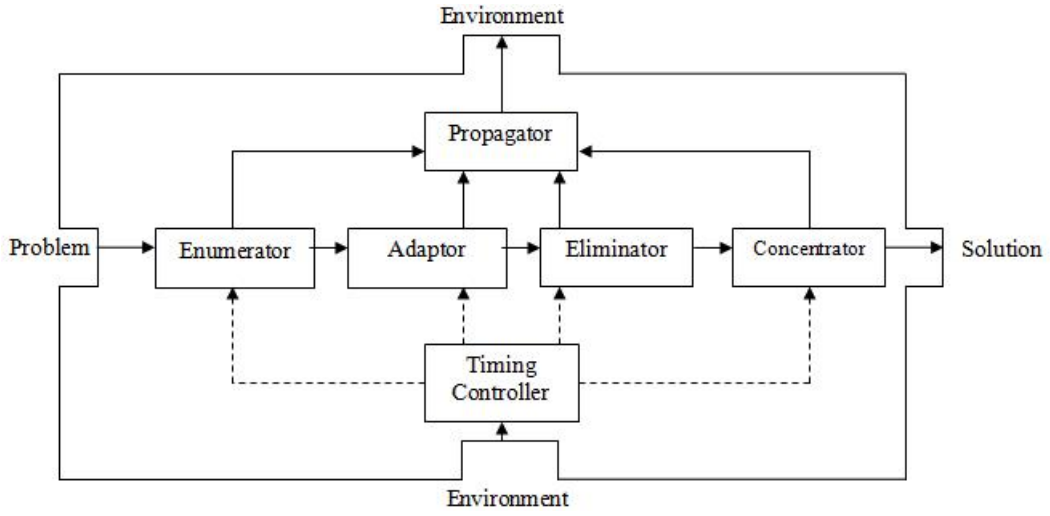


Figure 1 - The basic building block (BBB)

Adaptation: Solutions are enumerated and adapted to the environment. Sometimes adapted solutions are mutations that transcend enumerated solutions of the past.

Elimination: Unsuitable solutions are eliminated, so that only suitable solutions are further considered.

Concentration: Among the suitable solutions left, resources are further concentrated to only one (or at most a few) of the suitable solutions.

The above five characteristics are well known to the Darwinists [2]. The sixth one, on the other hand, is rather unique for SIS:

Slow decision cycle(s) to complement quick decision cycle(s): SIS possesses at least two decision cycles. The first one, defined as the quick decision cycle, provides an instantaneous response to the environment. The second one, defined as the slow decision cycle, tries to follow the gradual changes in the environment and analyze the information acquired by experts and past experiences. The two decision cycles enable the SIS to both cope with the environment and meet long-term goals. Sophisticated SIS may possess multiple slow decision cycles and multiple quick decision cycles. Most importantly, actions of slow decision cycle(s) may override actions of quick decision cycle(s), resulting in poorer performance in the short run but better performance in the long run.

Table 1 summarizes the characteristics of some SISs, with the number of asterisks (0 to 3) indicating the strength of a characteristic.

3. Structure of Slow Intelligence Systems

Now we can consider the structure of SIS by the introduction of the basic building block and advanced building block. Figure 1 illustrates the Basic Building Block (BBB).

Problem and solution are both functions of time, thus we can represent the time function for problem as $x(t)_{problem}$, and the time function for solution as $y(t)_{solution}$. The timing controller is also a time function timing-control(t). For the two-decision-cycle SIS, the basic building block BBB can be expressed as follows:

	Enumeration	Propagation	Adaptation	Elimination	Concentration	Multiple Decision Cycles
Intelligent Travel planning Systems [6]	***	***	***	***	***	*
Intelligent health restoration system [7]	**	**	**		**	*
3D Reconstruction Systems [8]	***	***	***	***	***	
Classifier Selection System[9, 10]	***	***		***	***	

Table 1- Characteristics of sample Slow Intelligence Systems

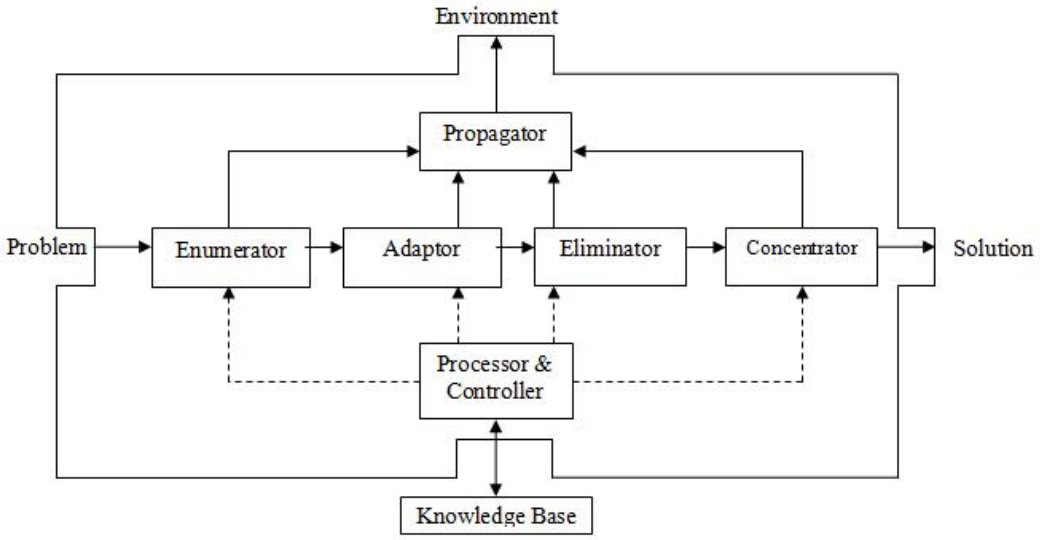


Figure 2 - The advanced building block (ABB)

```
if timing-control(t) == 'slow'
then /* timing-control(t) is 'slow' */
       y(t)solution =  gconcentrate (geliminate (gadapt
(genumerate(x(t)problem))))
else  /* timing-control(t) is not 'slow' */
       y(t)solution =  fconcentrate (feliminate (fadapt
(fenumerate(x(t)problem))))
```

where $g_{enumerate}$, g_{adapt}, $g_{eliminate}$, and $g_{concentrate}$ are the *transform functions* for enumeration, adaptation, elimination and concentration respectively during slow decision cycles, and $f_{enumerate}$, f_{adapt}, $f_{eliminate}$, and $f_{concentrate}$ are the transform functions for enumeration, adaptation, elimination and concentration respectively during quick decision cycles. Depending upon applications, the order of these transforms may be altered, and some transforms may be omitted. Section 6 provides an example where some of these transform functions are omitted during quick decision cycles.

An Advanced Building Block can be a stand-alone system as shown in Figure 2. The major difference between an ABB and a BBB is the inclusion of a knowledge base, further improving the SIS's problem solving abilities.

Since the transform functions of the ABB are influenced by the knowledge base, they are considered to be *knowledge transforms*. More specifically, when the knowledge base is represented by an ontology, the transforms are *ontological transforms*. A particular ontological transform emphasized in our research is the ontological filter. Ontological filters can be used both as Eliminators and as Concentrators. The Propagator can also use ontological filtering to selectively send messages to other SISs. For reasons to be explained later, an SIS works at its best with a special type of ontology called lightweight-plus ontology, which will be described in details in the next section.

4. Lightweight-plus Ontology

To define an appropriate ontology that can be used by SISs to classify tasks, represent knowledge, acquire information about the context, improve the interoperability among subsystems and other information sources and manage ontology evolution, we propose an intermediate ontology between the Lightweight Ontology [11] and the Heavyweight Ontology [12], called the *Lightweight-plus Ontology*. There have been many definitions of ontology. A good definition introduced by the ISO 18629-1 states that ontology is "a lexicon of specialized terminology along

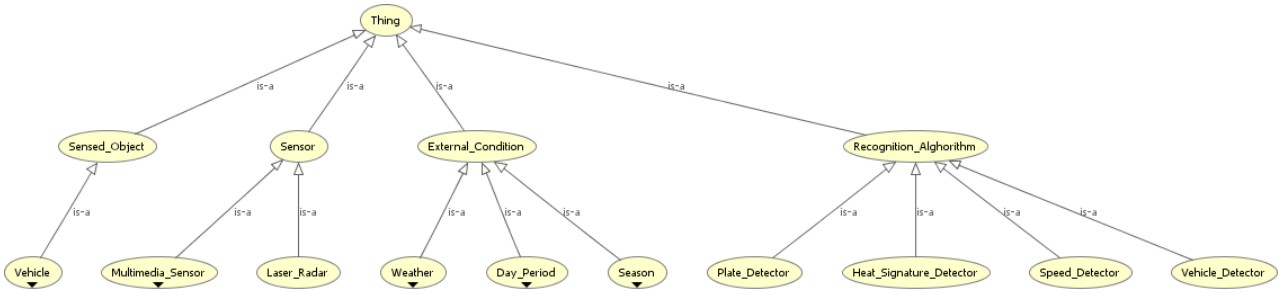


Figure 3 - Ontological Knowledge Base

with some specification of the meaning of terms in the lexicon". A lexicon is a set of symbols and terms. An ontology is "lightweight" when it includes concepts, concept taxonomies, simple relationships between concepts (for example "specialization", "is a") and properties that describes concepts. As an example, an SIS for sensor-based multimedia information processing can discover that some sensors are not providing information because they do not provide the complete view of the scene, and automatically choose to query the sensors that can provide the whole scene. In order to do so, the system should have the knowledge about the real world and the sensors, described by an ontological knowledge base according to the schema represented in Figure 3 [13].

As illustrated by Figure 3, this ontological knowledge base consists of four parts: the sensor part describing the sensors, the recognition algorithms, the external condition and the sensible objects part describing objects to be sensed. If there are no "non-hierarchical relations" (as in the example illustrated by Figure 3), the ontology can be called a *Lightweight Ontology*. An ontology is defined as O = {C, A, H, R_T, R} where:

- C is the concept set. $c \in C$ expresses one concept. In each ontology there is always a root concept marked as "Thing". For each $c \in C$ there exist a descendant_nodes set (C_{DN}) containing all its lower_layer concepts, and an ancestry_nodes set (C_{AN}) containing all upper_layer concepts.
- A is the concept attributes set. For $c \in C$ its attributes set is expressed as $A_C = \{a_1, \ldots, a_n\}$ where n is the number of attributes related to *c*.
- H expresses the concept hierarchy set. The formalism (c_i,c_j) means that c_i is the sub-concept of c_j. In other words this set contains *is_a* relations among the classes.
- R_T is the set of semantic relations type. $R_T = R_{TD} \bigcup R_{TU}$, where R_{TD} is the set of predefined relation type (*same_as*, *disjoint_with*, *equivalent*), and R_{TU} is the set of user defined relation type. The formalism (c_i,c_j, r) with $r \in R_T$ means that between c_i and c_j there is the r relation. The set $RelR_T(c_i,c_j)$ contains the relation r between c_i and c_j.
- R is the set of non-hierarchical relations. The formalism (c_i,c_j, r) with $r \in R$ means that between c_i and c_j there is the r relation. The set Rel(c_i,c_j) contains the relation r between c_i and c_j.

In the case of a Lightweight Ontology, there are no R_T or R sets. Therefore the Lightweight Ontology is defined as: O_L = {C, A, H}. The lightweight ontology is a simple and basic representation of a knowledge domain and contains only a general statement of a problem. The lightweight ontology can be regarded as a sort of taxonomy and a first approach to the resolution of various users' queries. Obviously it could be complicated by the insertion of new non-hierarchical relationships that enrich its semantic expressivity. In this case, a more mathematically rigorous representation is needed to provide greater confidence that the real meaning behind terms coming from different systems is the same. Heavyweight ontologies are extensively axiomatized and thus represent ontological commitment explicitly. Axioms help to exclude terminological and conceptual ambiguities due to unintended interpretations. Heavyweight ontologies can have a lightweight version. Many domain ontologies are heavyweight because they support heavy reasoning. Therefore, the heavyweight ontologies add axioms, well-formed formulas in a formal language, and constraints to lightweight ontologies in order to clarify the intended meaning of the terms gathered on the ontology. A heavyweight ontology is defined by O_H = {C, A, H, R_T, R, A_X} where A_X indicates the axioms that are in the ontology. Lightweight and heavyweight ontologies can not be considered as good solution for an SIS.

Lightweight ontologies propose a simple view of the domain and do not allow a detailed description of the interactions among all the components in the domain. Using a lightweight ontology, users or systems can only share a very small and simple description of a domain that might not be useful for the resolution of a problem. On the other hand, using a heavyweight ontology, users can create and share very complex domains. However, it is very easy to have problems defining axioms as well as managing the ontology. To address these problems, a solution could be the introduction of a *lightweight-plus ontology* defined as O_{L+} = {C, A, H, R_T, R}. By introducing non-hierarchical relations, a lightweight-plus ontology is more complex and semantically richer than the lightweight ontology, but it is not as complex as a heavyweight ontology because there are no axioms to consider.

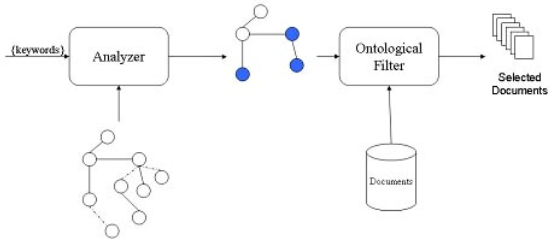


Figure 4 - Search based on Ontological Filtering. Topics related to keywords are nodes in gray color. The relations represented by the solid lines are "non-hierarchical" while the dotted lines represent "hierarchical" relations.

Some user requests can be supported by the use of one of these lightweight-plus ontologies. Although the use of heavyweight ontologies guarantees the full execution of a task or a request, it does not assure the optimal solution. Each of these ontologies enriches its semantic level by the introduction of new classes, relations, functions, formal

axioms and instances. Such enrichments usually lead to better satisfaction of user requests. The lightweight-plus ontology will be the starting point for the definition of an ontological filtering methodology, which will be explained in the next sections.

5. The Ontological Filter

When a user searches a web repository that refers to a well-defined domain, the user usually expresses the query by the use of keywords. The search is conducted using the keywords and by counting their occurrences (mutually exclusive or not) in the documents. This is a simple syntactic search.

With the ontological filter (Figure 4), the user still searches using keywords, but the keywords are considered as attributes of concepts in an ontology. A keyword can be associated with various concepts. For example, in an ontology related to sports, the keyword "ball" could be associated with various kinds of sports. In this way, by the use of keywords we can navigate the domain through the topics in an ontology which can be considered as a sort of sieve for the searching of documents. The system returns to the user only those documents related to certain concepts in this ontology. Other documents that do not contain keywords but are related to the concepts that contain them will also be selected. This kind of approach can be generalized to different problem domains. For example, in the case of a video surveillance system, each component can be part of some topics and can be identified by some keywords. Thus the requests of the user can be easily mapped as a set of topics belonging to an ontology.

A more formal definition of Ontology Filtering can be developed in the following way. First of all, the lightweight-plus ontology representing the problem domain is defined. This ontology can be obtained either from human experts or through automatic reasoning techniques to infer the description of the domain by the analysis of related data. In this way the O_{L+} = {C, A, H, R_T, R} representing the domain of interest is constructed.

The user introduces its query by the use of keywords K_U={k_{ui}}. The following strategy can be applied to build the lightweight-plus ontology:

Step 1: $\forall k_{ui} \in K_U$ and $\in A_{Ci}$ add to the set C' of the Ontology O'_{L+} = {C', A', H', R'_T, R'} the concept C_i belonging to an ontology O_{L+} = {C, A, H, R_T, R}. The set A_{Ci} contains the attributes related to the the concept C_i and is a subset of A.

Step 2: $\forall C_i \in C'$ add to the set C' of the Ontology O'_{L+} = {C', A', H', R'_T, R'} all the concepts C_j that share relations of kind H, R_T and R with the nodes C_i. These relations are inserted in the sets H', R'_T and R' and all the attributes of the various C_j

At the end of these two steps the O'_{L+} = {C', A', H', R'_T, R'} is the lightweight-plus ontology that expresses the domain according to the requests of the user. In this way the ontological filter can be developed: in fact all the "objects" that are in the complete domain represented by the O_{L+} can be filtered by the use of the ontology O'_{L+} obtaining only the objects that user needs (Figure 5).



Figure 5 - In this figure three layers are shown. The lower layer contains the items that are in the domain, which is represented by the lightweight-plus ontology in the middle layer. The upper layer contains the keywords that are related to the concepts in the lightweight-plus ontology. On the left, the full structure is illustrated. On the right, the structure is illustrated where the user introduced some keywords (upper layer) that are related to some topics (the nodes in gray color) in the lightweight-plus ontology (the middle layer), so that some "items" in the lower layer can be selected.

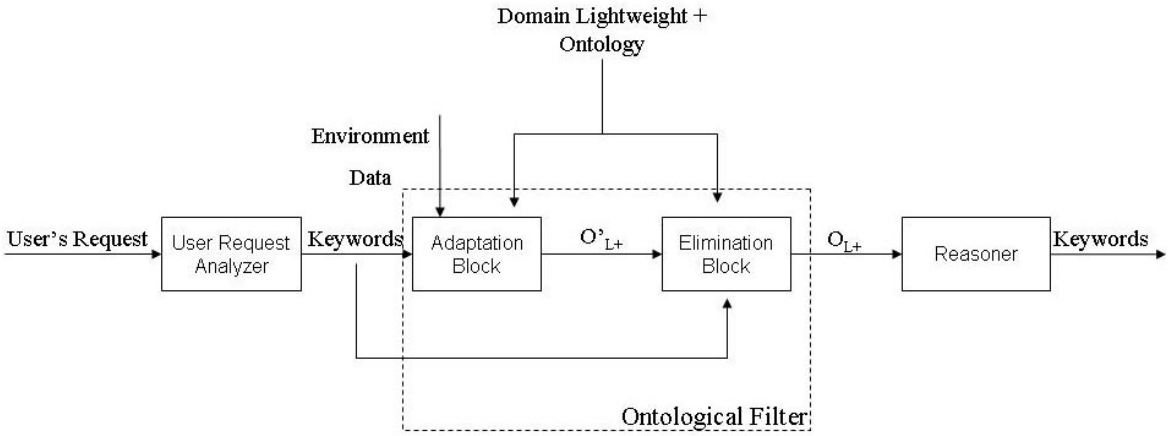


Figure 6 - Ontological Filter and the Slow Intelligent System

The ontological filtering approach can be applied to process the user's request by determining all the "items" needed for the resolution of a query. This approach can be combined with the SI approach and to implement the Adaptor and the Eliminator blocks (Figure 6). The Adaptor block aims to the adaptation of the system acquiring information on the context where it is working. The Eliminator selects the best methodologies and items to solve a problem according to the information acquired in the previous phases from experts and past experiences. The ontological filtering approach and in particular the previous described methodology can be used to implement the functionalities of these two blocks. In fact after the first stage of a Slow Intelligent System, namely the Enumeration Block, the full domain, related to the problem, can be explained by the use of a lightweight-plus ontology. In order to obtain a solution, the system has to adapt the domain to the environment where the problem is located and to choose one of the possible solutions. Ontological filtering works as follows: by analyzing the user requests the ontological filter can obtain information about the environment and choose only those objects in the domain that are useful for the resolution of the problem.

As an application to distributed multimedia systems, in a video surveillance system the request "send an alarm message each time a vehicle exceeds the speed of 60 mph in the evening period" can be managed in the following way: the system can extract the following keywords: "evening" "vehicle" "exceed" "speed" "60 mph". In this way a subset of the lightweight-plus ontology describing the full domain can be extracted. In particular, information on the environment can be extracted (for example the evening period) and also information on technologies and methodologies for the resolution of the user query (the system has to track a vehicle and measure its speed by the use of sensors). Thus the components, the work conditions and how they have to work are obtained using the relationships in the lightweight-plus ontology.

In addition to applications to distributed multimedia systems, another kind of application will be briefly described in the following section.
.

6. Application to Product & Service Customization

The ability to create customized product and service configurations fulfilling user's needs is a major goal sought after by many companies, especially in the design and implementation of web-based services that allows users to create and customize in a simple and intuitive way a preferred product and/or preferred services [14], [15]. To define a common vocabulary that opens possibilities such as the ability to express customer requirements correctly and to exchange knowledge, we can exploit the ontological formalism.

Following the Slow Intelligence approach, we can design an ontology-based *product and service customization system* that maximally satisfies users' needs starting from the desired requirements, the available components, the context information and previously obtained similar configurations. This system is built around the SIS component infrastructure, and finds the preferred candidate configuration through a process of enumeration, propagation, adaptation, elimination and concentration. Figure 7 illustrates the framework. This system is composed of SIS blocks that are *ontological transforms* to configure customized products and services. The following main SIS blocks are identified:

- **Enumerator Block**: This block has as inputs the user request and the product definition and as output the ontology O'_{AP}. It has two ontology builder components as well as comparison, merging and simplifier modules. As the output of this block, the ontology representing a first rough version of the customized product is obtained.

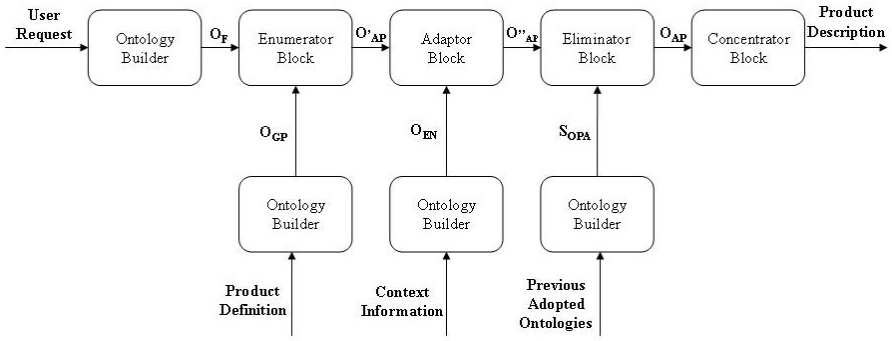


Figure 7 - Ontological transforms for product and service customization

- **Adaptor Block**: This block has as inputs the ontology O'_{AP}, representing a first rough version of the customized product, and the information about the context and as output the ontology O''_{AP}. It has as components an ontology builder as well as a comparison, merging and simplifier modules. As the output of this block, the ontology representing a context-adapted product is obtained.
- **Eliminator Block**: This block has as inputs the context-adapted ontology O''_{AP} and product ontologies developed in the past by other similar PCSs that work in other part of the system. The aim of this block is the tuning of the ontology O''_{AP} according to the previous configurations obtained in similar contexts. The output of this block is the ontology O_{AP} that represents the adapted product.
- **Concentrator Block**: this block has as inputs the ontology O_{AP} and as output the configuration of the product. The aim of this block is the generation of an XML file representing the ontology of the customized product.

Let U_R and U_P represent the user request and product definition respectively. The configuration problem CP is formulated as a composition of ontological transforms: $F_C(F_{EL}(F_A(F_{EN}(U_R, U_P))))$, where F_{EL} and F_C can be implemented as *ontological filters*. The proposed system can follow either a slow or a quick decision cycle. The previous formulation can be defined as the slow process, while the quick process can be defined as a simplified sequence of ontological transforms: $F_C(F_{EN}(U_R, U_P))$.

We have successfully employed this approach for configuring personal computers [16]. The selection task is accomplished according to the information furnished by the user (i.e. select the less expensive components), by the comparison of previous configuration furnished in the past and by the rules codified in the ontology (i.e. an Intel processor needs an Intel motherboard). At the end of the process, the system generates an XML file containing the final configuration of the personal computer. The graphical user interface presents this file to the customer, in a user-friendly manner, giving the user the opportunity to check the correctness of the configuration. The final configuration of the product can use the same XML file. Experimental results demonstrate the ability of the Product Configurator to acquire new knowledge over time yielding in an improved product configuration [16].

The above approach is also applicable to detecting emerging trends of hot topics on the web [17]. For example, if a new device in personal health care to painlessly measure blood sugar level appears on the market, it will be discussed a lot in the newspapers, TV/radio talk shows and on the web. The IDEAS advanced research projects group at the Institute for Information Industry in Taiwan is applying slow intelligence techniques to this data mining problem. Another application is in network management [18]. The SIS network manager acquires information from the various network hosts according to SNMP standard, and then applies slow intelligence techniques to solve network management problems. Initial experimental results confirm the feasibility of this approach [18].

7. Conclusions

In summary, the SISs are general-purpose systems characterized by being able to improve performance over time through a process involving enumeration, propagation, adaptation, elimination and concentration. The transform functions of the building blocks for SISs are knowledge transforms. When the knowledge base is an ontology, the transforms are ontological transforms. A particularly useful ontological transform is the ontological filter, which can be used to implement both the Eliminator and the Concentrator following the slow intelligence approach.

In this paper we presented in details lightweight-plus ontology and the ontological filter. A lightweight-plus ontology combines the hierarchical relationships of a lightweight ontology with the semantic relationships of heavyweight ontology, but without the complexities of the semantic reasoning provided by the axioms. The lightweight-plus ontology is thus easier to work with for representing knowledge bases and problem domains in SISs.

To support the transformation functions of SIS components, the paper presents the ontological filter. With ontological filters, SIS components are able to prune the ontologies by identifying the attributes in the concepts matching the required keywords. Using the semantic relationships provided by the lightweight-plus ontology, the pruning of the ontology is able to keep those concepts that either match the keywords directly as well as indirectly, by a semantic relationship.

This paper also described briefly conceptual models for using ontology filtering as well as lightweight-plus ontology to product and service customization, and to the data mining of hot topics on the Internet. Ontological filtering can be applied in other areas as well. Further work includes applying these concepts in multimedia query processing where data fusion of sensor data must be analyzed as well as its use as part of the Propagator in a SIS to selectively send messages to other SISs.

References:

[1] Shi-Kuo Chang, "A General Framework for Slow Intelligence Systems", International Journal of Software Engineering and Knowledge Engineering, Volume 20, Number 1, February 2010, pp. 1-15.

[2] Charles Darwin, "Origin of Species by Means of Natural Selection: Or the Preservation of Favored Races in the Struggle for Life", Publisher: P.F. Collier and Son, 1901.

[3] Jie Li, "Modern sports training system of self-organization theory and philosophic", Sports Science, Vol. 18, No. 6, 1998.

[4] Gottfried J. Mayer-Kress, "Complex systems as fundamental theory of sports coaching?" Keynote presentation to the 2001 International Sports Coaching Symposium of the Chinese Taipei University Sports Federation, Taichung, Taiwan.

[5] John Langford, Alexander J. Smola and Martin Zinkevich, "Slow Learners are Fast", Neural Information Processing Systems Foundation 2009.

[6] David Camacho, Daniel Borrajo and Jose M. Molina. "Intelligent travel planning: A multi-agent planning system to solve web problems in the e-Tourism Domain", Autonomous Agents and Multi-Agent Systems, Vol. 4, 2001, 387-392.

[7] O. D .Karaduman, A. M. Erkmen and N. Baykal, "Intelligent "Health Restoration System": Reinforcement learning feedback to diagnosis and treatment planning", 5th WSEAS International Conference on Telecommunications and informatics, Istanbul, Turkey, 2006, 463-468.

[8] Thomas Marill, "Emulating the human interpretations of Line Drawings as Three-Dimensional Objects", International Journal of Computer Vision, Vol. 6, No. 2, 1991, 147-161.

[9] Giorgio Giacinto and Fabio Roli and Giorgio Fumera, "Selection of Classifiers based on Multiple Classifier Behavior", Proc. of the Joint international workshops on Syntactical and Structural Pattern Recognition & Statistical Pattern Recognition, 2000.

[10] Giorgio Giacinto and Fabio Roli, "Adaptive Selection of Image Classifiers", Electronics Letters, Publisher: Springer Verlag Ed, 1997, 38-45.

[11] Davies, J., D. Fensel, et al., Towards the Semantic Web: Ontology-driven Knowledge Management, Wiley, 2002

[12] Fensel, D., Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, Springer-Verlag, Berlin, 2001.

[13] Shi-Kuo Chang, Gennaro Costagliola, Erland Jungert and Francesco Orciuoli, "Querying Distributed Multimedia Databases and Data Sources for Sensor Data Fusion", IEEE Trans. on Multimedia, Vol. 6, No. 5, October 2004, 687-702.

[14] X. Ding, "Product Configuration on the Semantic Web Using Multi-Agent", Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on, pp. 304 – 309, 2008.

[15] N. Franke, F. T. Piller, "Configuration Toolkits for Mass Customization: Setting a Research Agenda", International Journal of Entrepreneurship and Innovation Management, No. 1, 2003.

[16] Emilio Zegarra, Francesco Colace, Massimo De Santo and Shi-Kuo Chang, "An Ontology-based Configurator for Customized Product Information based upon the Slow Intelligence Systems Approach", Proceedings of Int'l Conference on Software Engineering and Knowledge Engineering (SEKE2010), San Francisco Bay, July 1-3, 2010.

[17] Chia-Chun Shih and Ting-Chun Peng, "Building Topic/Trend Detection System based on Slow Intelligence", Proc. of 2010 Int'l Conference on Distrubted Multimedia Systems, Chicago, Illinois, Oct 14-16, 2010.

[18] Francesco Colace, Shi-Kuo Chang and Massimo De Santo, "SINMS: A Slow Intelligence Network Manager based on SNMP Protocol", Proc. of 2010 Int'l Conference on Distrubted Multimedia Systems, Chicago, Illinois, Oct 14-16, 2010.

Modeling Human Intelligence as A Slow Intelligence System

Tiansi Dong
Department of Computer Science
University of Hagen, Germany
E-mail: tiansi.dong@fernuni-hagen.de

Abstract

Chang [2] argues that slow intelligence is a property shared by a surprisingly large number of intelligent systems, and presents examples in scenarios of multimedia data query and mission control in crisis management systems. This paper strengthens this argument by showing that human infant's object tracing can be understood as a slow intelligence and be simulated in this framework. We first introduce an experiment on infant's capability of object tracing, and present two rules in psychology which govern object tracing results under certain conditions, namely, the rule of minimal spatial transformation and the rule of categorical continuity. We review related works and reduce object tracing problems into the problem of object mapping. We show that the following capabilities can be simulated in the slow intelligence framework: the capability of acquiring knowledge about one scene, and the capability of mapping object between scenes. An experiment result is briefly outlined. Some discussions and outlooks are presented.

1 Introduction

Though human adults can play football, pick out fresh apples in grocery, and enjoy f lms, newly born infants have diff culty in tracking moving objects. Carey [1] presents an interesting example about object tracing as follows: at f rst you look at Panel A, then close your eyes for 5 minutes; when you open the eyes again, you see Panel B (instead of Panel A). In this case, you will understand that the bird has f own from upper-left to bottom-right and the rabbit has moved from bottom-left to upper-right. Imagine the situation as Carey [1, p.72-73] described as follows: *imagine that the center is new a fixation point, and Panels A and B are projected one after the other onto a screen while you maintain fixation on the common point* and *the timing of the stimuli supports apparent motion*. What would happen? You will see a bird moves from upper-left to upper-right and turns out to be a rabbit and a rabbit moves from bottom-left to bottom-right and turns out to be a bird, as illustrated in Figure 1. Two rules were drawn by psychologists for object tracing: one is that only objects in the same category are traced; the other is that objects are so traced that the spatial difference between the two scenes is minimal. One privileges spatial information leading to the object tracing result in Panel D, the other privileges the category (or property) information, leading to the object tracing result in Panel C. There are lots of research works carried out by psychologists to explore infants' object tracking problems, see [11] and [1] for a review. In this paper, we focus on the computational simulation of these two object-tracing rules in the framework of slow intelligence.

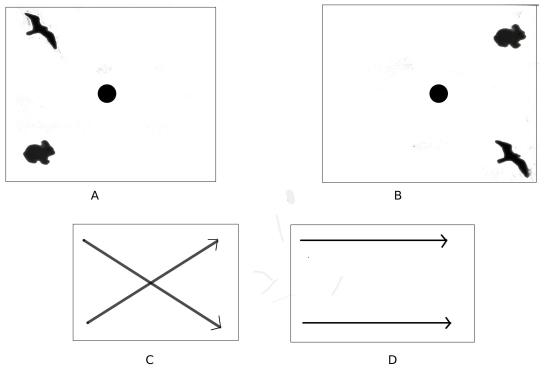


Figure 1. Two ways to observe the changing from Panel A to Panel B lead to different interpretations

The philosophy we follow is a developmental perspective, such that adults' intelligence system is developed from infants', and infants' intelligence system is developed from the state that they even cannot trace moving objects. This also explains why we are interested in the possibility to simulate object tracing – It is one of the early stages towards simulating human intelligence. Chang [2] observes that almost all of the intelligence systems in nature, including roses, have, in his terminology, *slow intelligence*. He

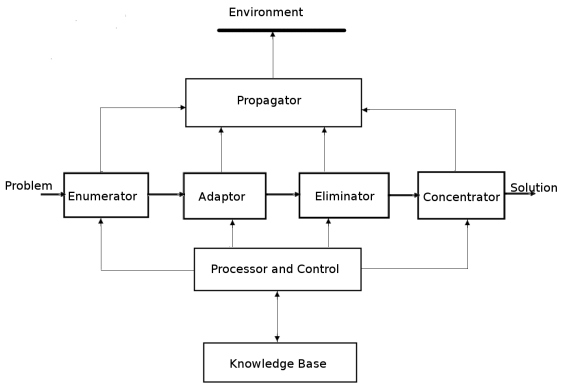


Figure 2. The framework of the advanced building block of a slow intelligence system

explains its functional definition as follows: it is the capability to solve problems by trying different solutions, to adapt to different situations and to propagate knowledge; with slow intelligence an agent may not perform well in the short run but continuously learns to improve its performance over time. A *slow intelligence system* (SIS) consists of function models as follows: enumeration, propagation, adaptation, elimination, and concentration. The framework of slow intelligence system is illustrated in Figure 2. If his argument is sound, the SIS frame shall be used to simulate human cognitive activities.

The aim of this paper is to show the possibility of simulating an object tracing problem in the framework of slow intelligence system (SIS). The rest of the paper is structured as follows: Section 2 analyzes the object tracing problem and briefly reviews some related works; section 3 and section 4 show that knowledge acquisition from static scene and between scenes can be both understood in the slow intelligence framework; section 5 outlines an experiment result which simulates the scenario of Figure 1.

2 Object tracing problem

> *A universe without objects . . . is a world in which space does not constitute a solid environment but is limited to structuring the subjects very acts; it is a world of pictures each one of which can be known and analyzed but which disappear and reappear capriciously.*
>
> *Piaget [12, pp.3-4]*

Human infants can not see constant objects when they are newly born. For a spatial change of an object from location A to B, they will believe two objects, one disappears in location A and one appears in location B [12]. The object tracing problem can be stated as follows: an agent perceives a sequence of snapshots of the scenes, how can he integrate them into a spatial-temporal whole, such that location changes of objects would be understood as objects' movements? A typical situation is to see a film: We have no difficulty in tracing objects and constructing continuous spatial-temporal changes among the motion pictures. In this paper we focus on the simplest situation in object tracing: Suppose an agent perceives two different pictures, e.g. Panel A and Panel B in Figure 1, how shall he map objects between the two pictures, so that mapped objects shall be recognized as the same instance and spatial differences between them shall be understood as spatial motion? For the example as illustrated in Figure 1, an agent might understand, *the bird has flown from upper-left to bottom-right*, or *the bird has flown from upper-left to upper-right and turned out to be a rabbit*. This suggests two sub-tasks for object tracing: one is object recognition, e.g. an agent recognizes birds and rabbits, the other is object mapping, either a bird is mapped to another bird located somewhere else, or a bird is mapped to a rabbit. Object recognition can be achieved with one scene, while object mapping is a reasoning process between two scenes. Therefore, we approach to the object tracing problem in two phases: the first phase is to retrieve knowledge from one scene, the second is to map knowledge between two scenes. Before showing that both problems can be approached in the framework of slow intelligence system, we review some related works.

2.1 Knowledge of objects

We look at scenes and recognize objects. Our visual system is capable of recognizing objects from stimuli. Psychological research shows that objects are recognized first at a particular level of abstraction. In particular Rosch and her colleagues [14] found that categories within taxonomies of objects are structured such that there is generally one level of abstraction at which humans find it easiest to name objects and recognize them the fastest, namely "basic level category". Basic level of abstraction is the level at which categories carry the most information, possess the highest cue validity, and are, thus, the most differentiated from others. Jolicoeur and his colleagues [8] found that every object has one particular level at which contact is made first with semantic memory. This level corresponds to the basic level in most cases. That is, the task of recognizing objects is a task of categorization.

2.2 Knowledge of static spatial relations

Knowledge of spatial relations has been researched with fruitful results. To name a few, Tolman [16] proved the exis-

tence of mental spatial representations by serial experiments with rats. Piaget and Inhelder [13] found that human babies acquire the three types of relations in a specif c order: topological relations, orientation relations, and distance relations. Carey [1] noted that infants f rst make a categorical distinction between contact and non-contact. Dong [4] formalized a qualitative spatial knowledge representation for scenes, in which the connection relation is primitive, and other qualitative spatial relations are developed.

2.3 Relations between knowledge of scenes

Object tracing problem can be addressed as an object-mapping problem between two scenes. Given knowledge representations of two scenes, how shall be objects in one scene mapped to objects in the other? Two rules adopted in this paper are that the rule of minimal spatial transformation (spatiotemporal information) and the rule of categorical continuity (feature information). With the two rules it is quite easy to explain the result shown in Panel C, Figure 1. To explain the result of apparent motion shown in Panel D, Figure 1, our hypothesis is that object recognition and spatial mapping are two separate cognitive routines, each corresponds to one rule: If they work in parallel, the routine of spatial mapping would be faster and deliver results before that of object recognition. Or, the routine of spatial mapping has the primacy over that of categorization, i.e. [15]. In either case, spatial mapping routine works with un-recognized objects, and later these un-recognized objects are labeled with names.

3 Acquiring spatial knowledge of a scene

In this section we show how to acquire spatial knowledge of a scene within the slow intelligent framework.

Dong [5] argued that three kinds of spatial relations can be developed by the connection relation with three axioms as follows.

1. $\forall x\ [\mathbf{C}(x,x)]$
2. $\forall xy\ [\mathbf{C}(x,y) \rightarrow \mathbf{C}(y,x)]$
3. $\forall x,y[\mathbf{C}(x,y) \rightarrow \forall \mathtt{z} \exists z[z \in \mathtt{z} \wedge \mathbf{C}(x,z) \wedge \mathbf{C}(z,y)]]$

where x, y, and z are regions in one scene, $\mathtt{z}$ is the category of region z. $\mathbf{C}(x,y)$ is read as 'region x and region y is connected', and means 'one region, x or y, and the closure of the other share a common point', i.e. [9].

The f rst two axioms are trivial. The third axiom is the characteristic property of the connection relation, which can be interpreted inside of the slow Intelligence framework as follows: the problem is whether x and y are connected. If an agent with slow intelligence f nds one category of region, such that for each region in this category it cannot connect with both of the two regions, he will conclude that x and y are disconnected from each other. That is, universal and existential quantif cations in the logical formula serves as the enumerating process. With $\forall \mathtt{z}$ a SIS agent will enumerate each region category $\mathtt{z}$ and with $\exists z[z \in \mathtt{z}]$ he might still enumerate all regions z in region category $\mathtt{z}$. The enumerating process generates values of variables for the predicates. Interpretations of predicates are determined within applications (the adaptation phase in SIS). For example, in general topology, two regions being connected is interpreted as *one region and the closure of the other share a common point*. Predicates f lled with variable values serve as the elimination phase, for an agent can make decisions on whether the enumerating process shall continue or the problem is solved. After the problem solving process the SIS agent will concentrate on a simple relation between x and y: being connected, or disconnected.

In appendix, an experiment is brief y outlined.

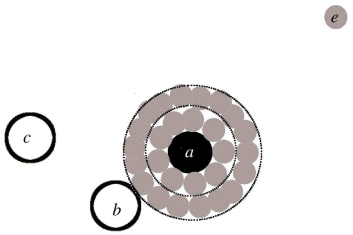


Figure 3. Knowledge that object C is nearer to Object A than to Object B can be acquired inside of the slow intelligence framework

The near extension of a by e, written as a^e, is def ned as the region z such that for every region w, z connects with w, if and only if there is a region e_0 in the same category as e such that e_0 connects both with x and w. This can also be interpreted inside of the slow intelligence framework: f rstly, enumerating all possible locations of e, then eliminating those that disconnect from a, at last concentrating on the space of a plus those regions in the category $\mathtt{e}$ which connect with a.

$$a^e \stackrel{\text{def}}{=} \iota z[\forall w[\mathbf{C}(w,z) \equiv \exists e_0[e_0 \in \mathtt{e} \wedge \mathbf{C}(x,e_0) \wedge \mathbf{C}(e_0,w)]]$$

In Figure 3, we know that object a is nearer to object b than to object c. This can be formally def ned as that *the near extension of a_1 by object e connects with object b while disconnects from object c, where a_1 is the near extension of object a by region e.* That is, $\mathbf{C}((a^e)^e, b) \wedge \neg\mathbf{C}((a^e)^e), c)$. Therefore, distance comparison can also be

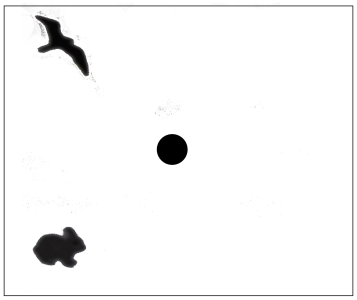


Figure 4. A slow intelligent system knows neither birds nor rabbits. When it perceives this scene, what it knows is two instances of the unknown object category, one is located upper-left, the other is located bottom-left

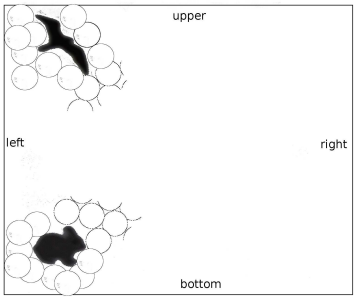


Figure 5. Location of the two objects can be acquired within the slow intelligence framework, i.e. the near extension procedure

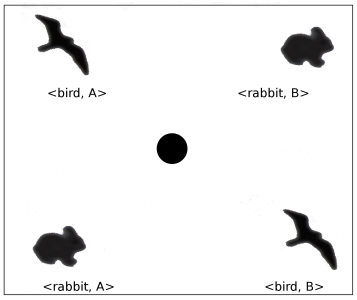


Figure 6. Project the two scenes in Figure 1 together, objects are named by its category, original scene name

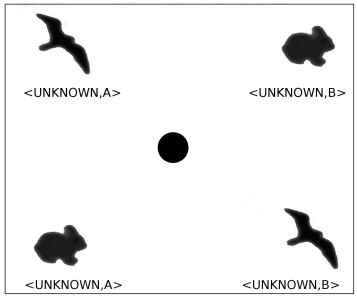


Figure 7. If a SIS agent can not distinguish birds from rabbits, it will believe that the bird in Panel A moves left and turns out to be a rabbit, and that the rabbit in Panel A moves left and turns out to be a bird

interpreted within the slow intelligence framework. Imagine that a slow intelligent agent (SIS) perceives Panel A in Figure 1, as illustrated in Figure 4, it can perform near extensions of the two objects, respectively, and this SIS agent would know that the bird is nearer to the upper-left side than to the bottom-right side, the rabbit is nearer to the bottom-left side than to the upper-right side, as illustrated in Figure 5.

As distance relations can be developed based on the connection relation, [3] and [6], and orientation relations can be understood as the distance comparison [7], we conclude that both distance relations and orientation relations can be interpreted in the slow intelligence framework.

4 Acquiring spatial knowledge between scenes

The spatial knowledge between scenes can be acquired by projecting the second scene onto the first one, and mapping objects in two scenes by minimizing the total cost of spatial transformations within the same object category. For example, we can project the two scenes in Figure 1 together, as illustrated in Figure 6. If we are capable of recognizing birds and rabbits, we will map the bird originally in the first scene to the bird originally in the second scene, and map the rabbit originally in the first scene to the rabbit originally in the second scene. To ease the object reference, we name each object by its category and its original scene name, e.g. <bird,A> refers to any bird from scene A, <rabbit,B> refers to any rabbit from scene B.

Imagine that a slow intelligent agent (SIS) is in a very early stage that it does not know what birds or rabbits look like. When this SIS perceives Panel A in Figure 1, as illustrated in Figure 4, it only know there are two unknown objects with the name: <UNKNOWN,A>. When Panel A

suddenly change to Panel B, it perceives two unknown objects located differently. Then with the condition of minimizing the total cost of spatial transformation, mapping two objects of the same category in Panel A to the two objects in Panel B would lead to the conclusion that the bird in Panel A is mapped to the rabbit in Panel B and that the rabbit in Panel A mapped to the bird in Panel B, as illustrated in Figure 7.

Let us make the scene-mapping problem clearer and show how it can be approached in the slow intelligence framework. From one scene, an agent acquires knowledge of objects in the scene and spatial relations among them. This knowledge, under certain conditions, can be represented in a graph structure $\mathcal{G} =< V, E >$, the node set $V = \{v_1, v_2, \ldots, v_n\}$ represents objects in the scene, each v_i has two kinds of information: its recognized category, v_i.`category`, and its location in this scene, v_i.`location`. The edge set $E = \{e_1, e_2, \ldots, e_m\}$ represents spatial relations among these nodes. The scene-mapping problem might turn out to be the graph mapping problem: Given two graphs $\mathcal{G}_1 =< V_1, E_1 >$ and $\mathcal{G}_2 =< V_2, E_2 >$, find a mapping $f : V_1 \rightarrow V_2$ such that v and $f(v)$ are of the same category and that the sum of $trans(v_i, f(v_i))$ reaches minimum, $trans(u, w)$ represents the spatial transformation between u and w. The minimum-finding task can be approached in slow intelligence framework as the procedures of enumerating, adapting, eliminating, and concentrating. That is, enumerating a possible mapping f_i, adapting to real environment, eliminating f_i if its total transformation is not minimum, and concentrating on the $f_{minimum}$. Based on $f_{minimum}$, a SIS will interpret location difference between mapped objects as movements.

5 Experiment result

We implemented a simple slow intelligence system which simulates two object tracing results in Figure 1. To simulate the first result in Figure 1, we create scene one with the first object `BIRD` located at `(1,5)` and the second object `RABBIT` located at `(1,1)`, and scene two with the first `RABBIT` located at `(5,1)` and the second `BIRD` located at `(5,1)`. The object-tracing result is illustrated in Figure 8. If the slow intelligence system can not distinguish birds from rabbits, that is, we set all the object categories into `UNKNOWN`, the object-tracing result is illustrated in Figure 9.

The computational complexity of this implementation is $\mathcal{O}(n!)$ with regards to the maximum object numbers n in scenes, therefore, it is indeed a *slow* system.

In this simulation, the knowledge representation of a single scene is simplified. For example, object locations and distances between objects are only represented in quantitative forms, though, spatial knowledge acquired through

```
Scene 1:
BIRD with Index 0 located at (1,5)
RABBIT with Index 1 located at (1,1)

Scene 2:
RABBIT with Index 0 located at (5,5)
BIRD with Index 1 located at (5,1)

 In Scene 1:
***BIRD with Index 0 located at (1,5)
 moved to the object with index 1 in the next scene
***RABBIT with Index 1 located at (1,1)
 moved to the object with index 0 in the next scene

 In Scene 2:
This is the last scene.
```

Figure 8. A slow intelligence system with the capability of recognizing birds and rabbits traces objects' moving in Figure 1

```
Scene 1:
UNKNOWN with Index 0 located at (1,5)
UNKNOWN with Index 1 located at (1,1)

Scene 2:
UNKNOWN with Index 0 located at (5,5)
UNKNOWN with Index 1 located at (5,1)

 In Scene 1:
***UNKNOWN with Index 0 located at (1,5)
 moved to the object with index 0 in the next scene
***UNKNOWN with Index 1 located at (1,1)
 moved to the object with index 1 in the next scene

 In Scene 2:
This is the last scene.
```

Figure 9. A slow intelligence system with no capability of recognizing birds and rabbits traces objects' moving in Figure 1

perception is neither metrical nor complete, rather qualitative and with distortions. Qualitative or uncertain spatial knowledge shall be considered in the future.

6 Conclusions and outlooks

Object tracing is one of the most important intelligent capabilities which infants developed in the early stage. We show that slow intelligence system (SIS) framework can be applied to simulate the decision process of object tracing. We argue that SIS framework can be applied to simulate human cognitive system.

Future work includes using SIS framework to simulate more core human cognition [1], and to connect with existing works in multimedia information processing, e.g. query on spatial-temporal data [10].

Acknowledge

I am indebted to S. K. Chang for a number of comments on the early draft of this paper, to I. Glöckner for his critical comments, and to three anonymous reviewers for their insightful critiques.

References

[1] S. Carey. *The Origin of Concepts*. Oxford Univesity Press, 2009.

[2] S. K. Chang. A general framework for slow intelligence systems. *International Journal of Software Engineering and Knowledge Engineering*, 20(1):1–18, 2010.

[3] T. de Laguna. Point, line and surface as sets of solids. *The Journal of Philosophy*, 19:449–461, 1922.

[4] T. Dong. SNAPVis and SPANVis: Ontologies for Recognizing Variable Vista Spatial Environments. In C. Freksa, M. Knauff, B. Krieg-Brückner, and T. Barkowsky, editors, *Proceedings of Spatial Cognition 2004*, Lecture Notes in Artif cal Intelligence, pages 344–365. Springer, Berlin, 2005.

[5] T. Dong. Towards a Spatial Representation for the Meta Cognitive Process Layer of Cognitive Informatics. In *Proceedings of the 6th IEEE International Conference on Cogntive Informatics*, pages 52–61. IEEE CS Press, Lake Tahoe, California, USA, 2007.

[6] T. Dong. A Comment on RCC: from RCC to RCC^{++}. *Journal of Philosophical Logic*, 37(4):319–352, 2008.

[7] T. Dong and H. W. Guesgen. A Uniform Framework for Orientation Relation based on Distance Comparison. In *Proceedings of the 7th IEEE International Conference on Cogntive Informatics*, pages 75 – 82. IEEE CS Press, Stanford University, California, USA, 2008.

[8] P. Jolicoeur, M. A. Gluck, and S. M. Kosslyn. From pictures to words: Making the connection. *Cognitive Psychology*, 16:243–275, 1984.

[9] J. K. Kelley. *General Topology*. Springer, New York, 1955.

[10] X. Li and S. K. Chang. An Interactive Visual Query Interface for Spatial/Temporal Data. In *Proceedings of International Conference on Distributed Multimedia Systems*, pages 257–262, San Francisco Bay, CA, 2004.

[11] K. Nakayama, Z. He, and S. Shimojo. Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn and D. Osherson, editors, *Visual cognition: An invitation to cognitive science*, pages 1–70. MIT Press, Cambridge, MA, 1995.

[12] J. Piaget. *The Construction of Reality in the Child*. Routledge & Kegan Paul Ltd, 1954.

[13] J. Piaget and B. Inhelder. *La représentation de l'espace chez l'enfant*. Bibliothèque de Philosophie Contemporaine, Paris: PUF, 1948. English translation by F. J. Langdon and J. L. Lunzer in 1956.

[14] E. Rosch, C. B. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[15] B. J. Scholl, Z. W. Pylyshyn, and S. Franconeri. When are spatiotemporal and featural properties encoded as a result of attentional allocation? *Investigative Ophtalmology and Visual Science*, 40(797), 1999.

[16] E. C. Tolman. Cognitive Maps in Rats and Men. *The Psychological Review*, 55(4):189–208, 1948.

SINMS: A Slow Intelligence Network Manager based on SNMP Protocol

Francesco Colace[1], Shi-Kuo Chang[2] and Massimo De Santo[1]

[1]Department of Information and Electrical Engineering, University of Salerno, Italy
[2]Department of Computer Science, University of Pittsburgh, USA
{fcolace, desanto}@unisa.it, chang@cs.pitt.edu

Abstract

Networks and distributed computing systems are becoming increasingly important and at the same time, more and more critical to the world of Information Technology. This rash spread, however, resulted in increased difficulty in configuring and managing networks. In particular the tasks of configuration management for IP network devices are becoming more and more difficult and the Simple Network Management Protocol is not able to manage very complex scenarios. Over the past years much efforts has been given to improve the lack of Simple Network Management Protocol and a promising approach involves the use of Ontology. This paper introduces a novel approach to the network management based on the use of the Slow Intelligence System methodologies and ontology. Slow Intelligence Systems is a general-purpose systems characterized by being able to improve performance over time through a process involving enumeration, propagation, adaptation, elimination and concentration. A Slow Intelligence System continuously learns, searches for new solutions and propagates and shares its experience with other peers. So the proposed approach aims to develop a system able to acquire, according to an SNMP standard, information from the various hosts that are in the managed networks and apply solutions in order to solve problems. To check the feasibility of this model first experimental results in a possible scenario are showed.

1. Introduction

Networks and distributed computing systems are becoming increasingly important and at the same time, more and more critical to the world of Information Technology. This rash spread, however, resulted in increased difficulty in configuring and managing networks. In fact there is an emergence of diverse network devices and it has become greatly difficult to configure those multifarious network devices with a manual work. The concept of network management is quite articulated. It involves activities such as the identification and management of various devices, monitoring their performance and much more. So efficient and intelligent configuration management techniques are urgently needed to configure these devices with automation or semi-automation [1]. A solution for this problem can be the adoption of the Simple Network Management Protocol (SNMP). The SNMP manages network hosts such as workstations or servers, routers, bridges and hubs to a central computer that runs the network management software. SNMP performs management services through a distributed architecture of systems and management agents. Since network management is critical for both the control for managing resources, SNMP can be used to:

- Configure remote devices: the management system can send configuration information to each host on the network.
- Monitor network performance: you can track the processing speed, network speed and outcome data transmissions.
- Detect network failures or unauthorized access: the management system can detect the occurrence of certain events on specific network devices. Activating an alarm, the device management system sends a message indicating the event occurred. Common types of alarms are activated and start down a device when an error connecting to a router and in case of unauthorized access.
- Check the network: the management system can monitor both the entire network to identify users or groups that access, use and specific types of devices and network services.

Since its introduction in the late 1980s the SNMP showed good performance in monitoring for fault and performance, but it is very hard to use in managing large networks. In fact SNMP structure of management information (SMI) or Next Generation Structure of Management Information (SMIng) is insufficient to represent hierarchical network configuration data. SNMP, besides, supports with difficulties several high-level management operations required by network configuration tasks. On the other hand as previously said the network management is a hot topic and there is a real interest in the development of an effective methodology. In literature ontology is considered a good way for supporting the network management and many papers deal with ontology based methodologies for network management. In particular they propose ontology as a layer able to improve the interoperability among devices and operators. In this sense [2] proposes an ontology driven approach to the semantic interoperability problem in the management of enterprise service. Another interesting approach is in [3], which proposes an improvement of the current network management methods

with the application of formal ontologies techniques. In particular it introduces a management information meta-model integrating all the information that currently belongs to different management model used to interoperate with the managed resource. Another advantage related to this approach is the ability to include basic semantic behavior for a manager to monitor and control these resources. This paper introduces a novel approach to the network management based on the use of the Slow Intelligence System methodologies [4] and ontology. The proposed approach aims to develop a system able to acquire, according to an SNMP standard, information from the various hosts that are in the managed networks and apply solutions in order to solve problems. In particular the proposed system can handle multiple networks and adopt solutions that have proved successful in some other context. By the use of ontologies the system will be able to choose the right action to take when some hosts send SNMP alerts. The use of the Slow Intelligence System approach will allow the system to automatically infer the actions to take. This paper is organized as follows. The next section introduces the slow intelligence systems approach. The second section describes the ontology. The third section explains why a Slow Intelligence System needs Ontology to work at his best while the third section describes the proposed system. The last section introduces the first experimental results.

2. What is a Slow Intelligence Systems

We will first introduce the concept of Slow Intelligence and present a general framework for designing and specifying Slow Intelligence Systems (SIS). We view Slow Intelligence Systems as general-purpose systems characterized by being able to improve performance over time through a process involving enumeration, propagation, adaptation, elimination and concentration [4]. A Slow Intelligence System continuously learns, searches for new solutions and propagates and shares its experience with other peers. A Slow Intelligence System differs from expert systems in that the learning is implicit and not always obvious. A Slow Intelligence System seems to be a slow learner because it analyzes the environmental changes and carefully and gradually absorbs that into its knowledge base while maintaining synergy with the environment. A slow intelligence system is a system that (i) solves problems by trying different solutions, (ii) is context-aware to adapt to different situations and to propagate knowledge, and (iii) may not perform well in the short run but continuously learns to improve its performance over time [4]. Slow Intelligence Systems typically exhibit the following characteristics:

Enumeration: In problem solving, different solutions are enumerated until the appropriate solution or solutions can be found.

Propagation: The system is aware of its environment and constantly exchanges information with the environment. Through this constant information exchange, one SIS may propagate information to other (logically or physically adjacent) SISs.

Adaptation: Solutions are enumerated and adapted to the environment. Sometimes adapted solutions are mutations that transcend enumerated solutions of the past.

Elimination: Unsuitable solutions are eliminated, so that only suitable solutions are further considered.

Concentration: Among the suitable solutions left, resources are further concentrated to only one (or at most a few) of the suitable solutions.

The sixth one, on the other hand, is rather unique for SIS:

Slow decision cycle(s) to complement quick decision cycle(s): SIS possesses at least two decision cycles. The first one, defined as the quick decision cycle, provides an instantaneous response to the environment. The second one, defined as the slow decision cycle, tries to follow the gradual changes in the environment and analyze the information acquired by experts and past experiences. The two decision cycles enable the SIS to both cope with the environment and meet long-term goals. Sophisticated SIS may possess multiple slow decision cycles and multiple quick decision cycles. Most importantly, actions of slow decision cycle(s) may override actions of quick decision cycle(s), resulting in poorer performance in the short run but better performance in the long run. Now we can consider the structure of SIS by the introduction of the basic building block and advanced building block. Figure 1 illustrates the Basic Building Block (BBB).

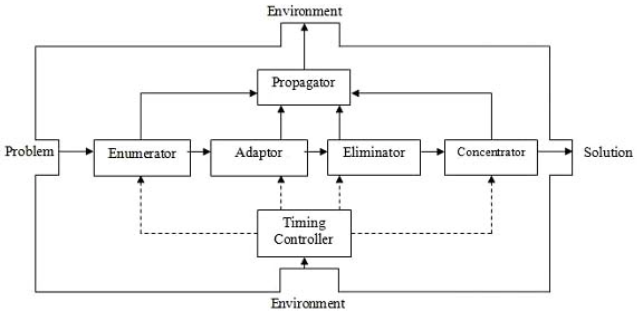


Figure 1 The basic building block BBB

Problem and solution are both functions of time, thus we can represent the time function for problem as $x(t)_{problem}$, and the time function for solution as $y(t)_{solution}$. The timing controller is also a time function timing-control(t). For the two-decision-cycle SIS, the basic building block BBB can be expressed as follows:

```
if timing-control(t) == 'slow'
then /* timing-control(t) is 'slow' */
       y(t)solution = gconcentrate (geliminate (gadapt
(genumerate(x(t)problem))))
else  /* timing-control(t) is not 'slow' */
       y(t)solution = fconcentrate (feliminate (fadapt
(fenumerate(x(t)problem))))
```

where $g_{enumerate}$, g_{adapt}, $g_{eliminate}$, and $g_{concentrate}$ are the *transform functions* for enumeration, adaptation, elimination and concentration respectively during slow decision cycles, and $f_{enumerate}$, f_{adapt}, $f_{eliminate}$, and $f_{concentrate}$ are the transform functions for enumeration, adaptation, elimination and concentration respectively during quick decision cycles. An Advanced Building Block can be a stand-alone system as shown in Figure 2. The major difference between an ABB and a BBB is the inclusion of a knowledge base, further improving the SIS's problem solving abilities.

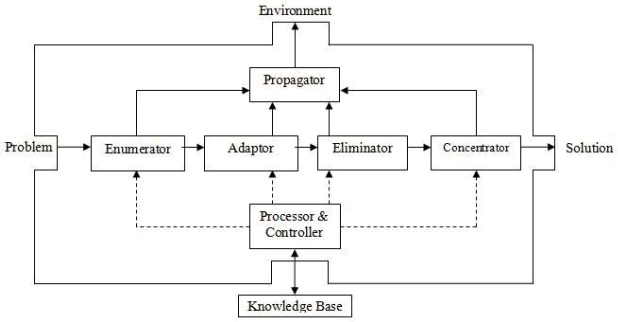


Figure 2 The advanced building block BBB

3. Why a Slow Intelligence System needs Ontology?

The definition of ontology is still a challenging task [5]. The term 'ontology' has as its origin in the Greek word 'ontos', which means 'being'. So in this sense ontology could be defined as a branch of philosophy dealing with the order and structure of reality. In the 1970s ontology came to be of interest in the computer science field. In particular the artificial intelligence community started to use the concept in order to create a domain of knowledge and establish formal relationships among the elements of knowledge in that domain for performing some processes of automated reasoning, especially as a means for establishing explicit formal vocabulary to be shared among applications. The term 'ontology' was first used in the computer science field by Gruber who used the term to refer to an explicit specification of a conceptualization [6]. The use of this term is rapidly growing due to the significant role it plays in information systems, semantic web and knowledge-based systems, where the term 'ontology' refers to "the representation of meaning of terms in vocabularies and the relationships between those terms" [7]. Also this kind of definition is still satisfactory for each field where ontology can be applied and so perhaps a good practical definition would be this: "an ontology is a method of representing items of knowledge (ideas, facts, things) in a way that defines the relationships and classification of concepts within a specified domain of knowledge" [5]. Following this point of view, ontologies are "content theories", since their principal contribution lies in identifying specific classes of objects and the relations that exist in some knowledge domains [8]. Ontologies can be classified into lightweight and heavyweight ontologies [9]. Lightweight ontologies include concepts, concept taxonomies, simple relationships between concepts (such as specialization "is_a") and properties that describes concepts. Heavyweight ontologies add axioms and constraints to lightweight ontologies. Axioms and constraints clarify the intended meaning of the terms gathered in the ontology. Commonly ontology is defined as $O = \{C, A, H, R_T, R\}$ where:

- C is the concept set. $c \in C$ expresses one concept and in each ontology there is ever a root concept marked as "Thing". In particular for each $c \in C$ there exist a descendant nodes set (C_{DN}) containing all its under layer concepts and an ancestry nodes set (C_{AN}) containing all upper layer concepts
- A is the concept attributes set. For $c \in C$ its attributes set is expressed as $A_C = \{a_1, \ldots, a_n\}$ where n expresses the number of attributes related to *c*
- H expresses the concept hierarchy set. The formalism (c_i,c_j) means that c_i is the sub-concept of c_j. In other words this set contains the *is_a* relations among the classes.
- R_T is the set of semantic relations type. $R_T = R_{TD} \, Y \, R_{TU}$. R_{TD} means the set of predefined relation (*same_as, disjoint_with, equivalent*) while R_{TU} means the set of user defined relation type. The formalism (c_i,c_j, r) with $r \in R_T$ means that between c_i and c_j there is the r relation. The set $RelR_T(c_i,c_j)$ contains the relation r between c_i and c_j
- R is the set of non-hierarchical relations. The formalism (c_i,c_j, r) with $r \in R$ means that between c_i and c_j there is the r relation. The set $Rel(c_i,c_j)$ contains the relation r between c_i and c_j

In the case of a Lightweight Ontology there are no R_T or R sets, therefore, the Lightweight Ontology is defined as: $O_L = \{C, A, H\}$. Therefore, the lightweight ontology is a very simple and basic representation of a knowledge domain and contains only a very general statement of a problem. The lightweight ontology can be viewed as a sort of taxonomy of the scenario and a first approach to the resolution of various users' queries. Obviously, it could be complicated by the insertion of new non-hierarchical relationships that enrich its semantic expressivity. In this case, a more mathematically rigorous representation is needed to provide greater confidence that the real meaning behind terms coming from different systems is the same. Heavyweight ontologies are extensively axiomatized and thus represent ontological commitment explicitly. Axioms help to exclude terminological and conceptual ambiguities due to unintended interpretations. Heavyweight ontologies can have a lightweight version. Many domain ontologies are heavyweight because they support heavy reasoning. Therefore, the heavyweight ontologies add axioms, well-

formed formulas in a formal language, and constraints to lightweight ontologies in order to clarify the intended meaning of the terms gathered on the ontology. Therefore, a heavyweight ontology is defined by $O_H = \{C, A, H, R_T, R, A_X\}$ where A_X indicates the axioms that are in the ontology. Lightweight and heavyweight ontologies can not be considered as good solution for a Slow Intelligent System.

Lightweight ontologies propose a very simple view of the domain and do not allow a detailed description of the interactions among all the components in the domain. Using a lightweight ontology, users or systems can only share a very small and simple description of a domain that might not be useful for the resolution of a problem. On the other hand, using a heavyweight ontology, users can create and share very complex domains. However, it is very easy to have problems defining axioms as well as managing the ontology. To address these problems, a solution could be the introduction of a lightweight plus ontology defined as $O_{L+} = \{C, A, H, R_T, R\}$. By introducing non-hierarchical relations, a lightweight plus ontology is more complex and semantically richer than the lightweight ontology, but it is not complex as a heavyweight ontology because there are no axioms to consider. The starting point of this approach is the idea that it is reasonable to think that between the lightweight and the heavyweight ontology there are a series of ontologies that can be defined as lightweight-plus ontologies.

Some user requests can be supported by the use of one of these lighthweight plus ontologies. Although the use of heavyweight ontologies guarantees the full execution of a task or a request, it does not assure the optimal solution. Each of these ontologies enriches its semantic level by the introduction of new classes, relations, functions, formal axioms and instances. These enrichments allow a better configuration of the system and the satisfaction of user requests. The lightweight plus ontology will be the starting point for the definition of a management network system.

4. A Slow Intelligence Network Manager based on SNMP Protocol

As previously said the aim of this paper is the introduction of a LAN-based management system based on SNMP protocol and the Slow Intelligence approach. Suppose to have M different LANs to which may belong to N different types of hosts that have to be managed. Each of these LANs is dynamic and therefore allows the introduction of new hosts and the disappearance of some of them. The local servers are in principle able to solve the main problems in the LAN management, but thanks to the dynamism of the LANS may be faced with unexpected situations. The environmental conditions in which the LAN operates can influence the performance of various hosts and must be taken into account. In this scenario a fundamental role is played by ontologies. In particular it is necessary to introduce and define the following ontologies:

- $O_{SNMP} = \{C_{SNMP}, A_{SNMP}, H_{SNMP}, R_{TSNMP}, R_{SNMP}\}$. This ontology aims to define the entire structure of SNMP protocol by analyzing the various messages and the relations between them
- $O_{Fault} = \{C_{Fault}, A_{Fault}, H_{Fault}, R_{TFault}, R_{Fault}\}$. This ontology describes each kind of possible errors that can occur within a LAN
- $O_{Cause} = \{C_{Cause}, A_{Cause}, H_{Cause}, R_{TCause}, R_{Cause}\}$. This ontology defines the causes of the faults that may occur in a LAN
- $O_{Solution} = \{C_{Solution}, A_{Solution}, H_{Solution}, R_{TSolution}, R_{Solution}\}$. This ontology defines the solutions that can be taken to recover from fault situations which occurred within a LAN
- $O_{Action} = \{C_{Action}, A_{Action}, H_{Action}, R_{TAction}, R_{Action}\}$. This ontology aims to identify the actions to be taken in order to recover from fault situations
- $O_{Component} = \{C_{Component}, A_{Component}, H_{Component}, R_{hComponent}, R_{Action}\}$. This ontology describes the components that may be present within a LAN
- $O_{Environment} = \{C_{Environment}, A_{Environment}, H_{Environment}, R_{hEnvironment}, R_{Environment}\}$. This ontology describes the environment where the LAN works

In order to allow the communication among the various hosts and servers that are in the various LAN the following messages have to be introduced:

M_{CSI}(SNMP, ID_Components) = this is the SNMP message that the client sends to the local server when an error has occurred. The ID_Componente used to identify the type of component that launched the message.
M_{SIC}({Action}) = this message, sent by local server, contains the actions that the client have to implement for the resolution of the highlighted fault.
The local server has to implement the following functions:

$O'_{Fault} = f(M_{CSI}(SNMP), O'_{SNMP})$ = this function aims to build the ontology of faults from the analysis of received SNMP messages and SNMP ontology within the local server. It is important to underline how the SNMP ontology on the local server is only a part of that present in the central server and is built from time to time following the faults that occur within the LAN.
$O'_{Cause} = g(M_{CSI}(SNMP), O'_{SNMP})$ = this function aims to obtain the ontology of the causes that generated the received SNMP messages.
$O'_{Solution} = h(O'_{Fault}, O'_{Cause})$ = this function calculates the ontology of possible solutions that the local server can find for the solution of the fault situation
$O'_{Action} = k(O'_{Solutiom})$ = this function calculates the ontology of possible solutions that the system can identify error to resolve the situation highlighted by the SNMP

These functions can be considered as the enumeration phase of the Slow Intelligent. After the determination of these functions the system can adopt the Action to apply in the LAN by the use of the following function:

{Action} = t(O'$_{Action}$, O'$_{Component}$, O'$_{Environment}$) = this is the set of actions that the client, or the host involved in the fault, must implement in order to solve the problem identified by the SNMP message. In practice, this involves defining, from ontologies of actions and components, the instances of actions to implement to resolve the faults that occurred. This function implements the Adaptation, Elimination and Concentration phases of a Slow Intelligence System. All these operations are carried out by involving the local server and hosts on the managed LAN. It is obviously the local server cannot always perform operations that are asked, because it does not know the full SNMP ontology. In fact the managed LAN can change: for example new components can be added. So new messages, functions and actions have to be expected among local servers and central server. The messages are so defined:

M_{CSlj}(SNMP, ID_Component) = this message contains the SNMP signal, sent by a host, that the local server is unable to manage and that it sends to the central network. The central server sends this message to the other local servers local in order to obtain information on the management of the SNMP signal.

M_{SljC}(O'$_{SNMP-i}$, O'$_{Cause-i}$, O'$_{Solution-i}$, O'$_{Action-i}$, {Action$_i$}) = this message contains the information obtained from local servers about the SNMP signal management. downstream of question to which they have undergone. This message can be empty when no local servers ever managed in the past this kind of SNMP signal.

Related to these messages there are the following functions:

O'$_{SNMP-i}$ = F(M_{SliSlj}(SNMP), $_{SNMP-j}$) = this function expresses the subset of the SNMP ontology built in the local server j needed by the local server i.

O'$_{Cause-i}$ = G(O'$_{SNMP-i}$) = this function expresses the ontology representing the causes of the fault. This ontology is built in the j-th local server and can be empty when this server never faced this problem.

O'$_{Solution-i}$ = H(O'$_{Cause-i}$) = this function gives the ontology of the solutions that can be adopted in order to solve the fault related to the SNMP signal. This ontology is built in the j-th local server and can be empty when this server never faced this problem.

O'$_{Action-i}$ = K(O'$_{Solution-i}$) = this function gives the ontology representing the actions that can be adopted for the solutions of the faults related to the SNMP signal. This ontology is built in the j-th local server and can be empty when this server never faced this problem.

The central server collects all the ontologies, obtained in the various local servers and previously described, and selects one of them according to an analysis based on ontology similarity. After this phase the central server can determine the action that have to be applied in the i-th LAN in order to solve the fault. So these actions can be sent to the i-th local server. In this way the following function can be introduced: {Action$_i$} = T(O'$_{Action-i}$, O'$_{Component-j}$) = this function calculates the set of actions that the client must adopt in order to solve the problem identified by the SNMP signal. The set of possible actions can of course be zero. In this case the support of an expert is needed. The previous messages and functions implement the propagation phase of the slow intelligence system approach.

The operational workflow is the following:

- Step 1: a SNMP messages generated by the Client as a result of a fault and sent to the local server
- Step 2: The local server receives the SNMP message and tries to identify the problem through analysis of various ontologies.
- Step 3 If the local server can identify the problem it generates the solutions and the actions that the various hosts in the LAN have to be apply.
- Step 3. 1 The hosts get the actions and put them into practice
- Step 4 If the local server does not identify the problem sends the report to the central server.
- Step 5 The central server sends to all local servers received the message
- Step 5.1 Other local servers after receiving the message attempts to determine the possible actions and then send everything to the central server.
- Step 6 If the central server has received the possible actions by local servers then sends them to the local server that has requested it. If no action is received, however, the central server, based on the received message and its general ontologies determines the actions to be sent to the local server.
- Step 7, the local server sends the actions to the various hosts that are in the LAN
- Step 7. 1 The hosts get the actions and put them into practice

5. Experimental results

In order to test the performance of the proposed system an experimental campaign has been designed. First of all the working scenario has been settled (figure 3).

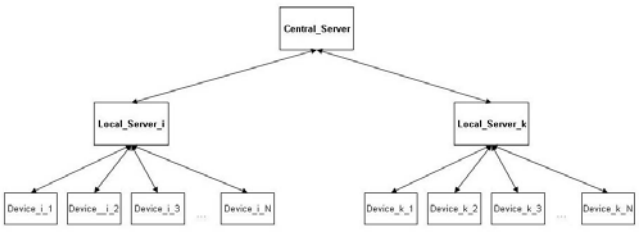


Figure 3: the operative scenario

In this scenario the central controller has to manage two different LANs. The first one is composed by a Cisco switch and 30 personal computers equipped with Microsoft

as Operative System and Microsoft Office as applicative software. The second LAN is composed by a Nortel switch, 30 personal computers equipped with various operative systems (Microsoft Windows XP, Linux Red Hat and Apple SnowLeopard) and a HP network printer. Each local server has SNMP ontology able to cover the 80% of the SNMP messages that the hosts in the LAN can launch. The experimental phase aimed to evaluate the following system's parameters:

- The system's ability to identify the correct management actions to apply in the LAN after a SNMP signal. This parameter, named CA, is so defined:

$$CA = \frac{\#Correct_Action}{\#Correct_Action + \#Wrong_Action}$$

- The system's ability to select in a LAN a viable solution that was previously adopted in a similar case in an other LAN. This parameter, named IS, is so defined:

$$IS = \frac{\#Correct_Inferred_Action}{\#Correct_Inferred_Action + \#Wrong_Inferred_Action}$$

- The system's ability to manage the introduction of a new component in a LAN. In particular the system has to recognize components that were previously managed in other LANs. This parameter, named KC, is so defined:

$$KC = \frac{\#Correct_Action_NC}{\#Correct_Action_NC + \#Wrong_Action_NC}$$

The previous indexes were calculated in the following way:

- The CA index: this index was calculated after 10, 20, 30, 40 and 50 SNMP signals. In this case there was not variations in the LANs
- The IS index: this index was calculated forcing some SNMP events in the LAN not expected in its SNMP reference ontology. This index was evaluated after 10, 20, 30, 40, 50 SNMP signal not expected.
- The KC index was estimated after the introduction of new components in a LAN. In particular for five times a component belonging to a LAN has been shifted in the other LAN and the index was evaluated after 10, 20, 30, 40, 50 SNMP signal launched from the host.

In the next table the obtained results are showed:

Index	10	20 30 40 50			
CA	90,00%	95,00%	93,33%	92,50%	92,00%
IS	50,00%	60,00%	66,67%	70,00%	74,00%
KC	60,00%	70,00%	76,67%	80,00%	82,00%

Table 1: Obtained Results. The KC has to be considered as average value

The indexes show the good performances of the system. In particular the CA index, that expresses the ability of the system in the recognition of the correct actions in the LAN after a SNMP signal, is very good. The IS index witnesses how the system uses at the best the SIS approach. In fact the system improves its performances sharing knowledge among the various local servers. At the beginning the index is very low but it increases after few iterations. In addition, the KC index shows a similar trend.

Conclusions

In this paper a novel method for network management has been introduced. This method is based on, SNMP; Ontology and Slow Intelligence System approach. It has been tested in an operative scenario and the first experimental seems to be good. The future works aim to improve the system by the use of new and effective methodologies for the ontology management and the use of other network management approaches.

Acknowledgements

The authors would like to thank Maria Lorusso e Attilio Sangiovanni for their collaboration and contribution to the work.

References

[1] Hui Xu, Debao Xiao, "A Common Ontology-based Intelligent Configuration Management Model for IP Network Devices", Proceedings of the First International Conference on Innovative Computing, Information and Control

[2] A. K. Yiu Wong, P. Ray, N. Parameswaran, J. Strassner, "Ontology Mapping for the Interoperability Problem in Network Management", IEEE Journal on Selected Areas in Communication, Vol. 23, No. 10, 2005

[3] Jorge E. López de Vergara, Antonio Guerrero, Víctor A. Villagrá, Julio Berrocal, "Ontology-Based Network Management: Study Cases and Lessons Learned", J. Network Syst. Manage. 17(3): 234-254, 2009

[4] Shi-Kuo Chang, "A General Framework for Slow Intelligence Systems", International Journal of Software Engineering and Knowledge Engineering, Volume 20, Number 1, February 2010, pp. 1-15.

[5] Jepsen, T., Just What Is an Ontology, Any way?, IT Professional, vol. 11, no. 5, pp. 22-27, Sep./Oct. 2009.

[6] Gruber, T.R, Translation approach to portable ontology specification, Knowledge Acquisition, v. 5, pp. 199-220, 93

[7]"OWL Web Ontology Overview", W3C Recommendation, 10 february 2004, http://www.w3.org/TR/2004/REC-owl-features-20040210/

[8] Maedche A., Staab S., Ontology Learning for the Semantic Web, IEEE Intelligent Systems, vol. 16 no. 2, Mar/Apr 2001, Page(s): 72-79.

[9] Corcho, O., A Layered Declarative Approach to Ontology Translation with Knowledge Preservation, Volume 116 Frontiers in Artificial Intelligence and Applications, 2005.

Building Topic/Trend Detection System based on Slow Intelligence

Chia-Chun Shih
Institute for Information Industry
Taipei, Taiwan
chiachun@iii.org.tw

Ting-Chun Peng
Institute for Information Industry
Taipei, Taiwan
markpeng@iii.org.tw

***Abstract*—It becomes an interesting research topic to detect trend in the Internet era, where millions of data are posted online everyday. As social media, for example, blogs, forums, and micro-blogs, are prevailing, many offline events are discussed online. The discussion data, which reflects what people are interested in, is useful for detecting trend. This research proposes a design of online topic/trend detection system with the advantages of Slow Intelligence. Unlike traditional Topic Detection and Tracking (TDT) tasks, which source data from offline news articles, the proposed system attempts to collect and analyze huge amount of up-to-date data from many heterogeneous websites on Internet. The Internet data complicates the system in two aspects: 1) it needs careful resource allocation to collect huge amount of up-to-date data based on limited computing resources; 2) it needs mechanisms to automatically or semi-automatically adapt data processing algorithms to handle varieties of data. This research adopts Slow Intelligence, which provides a framework for systems with insufficient computing resources to gradually adapt to environments, to handle these complexities.**

I. INTRODUCTION

It becomes an interesting research topic to detect trend in the Internet era, where millions of data are posted online everyday. As social media, for example, blogs, forums, and micro-blogs, are prevailing, almost all offline events are discussed online. The discussion data, which reflects what people are interested in, is useful for detecting trend. For example, HP Labs have demonstrated that social media can be effective indicators for predicting movie revenues [2].

Information Retrieval researchers have been working on Topic Detecting and Tracking (TDT) tasks for decades [1]. Traditional TDT tasks, which source data from on a stream of offline news stories, have been studied for decades; however, the Internet brings new challenges on designing online trend detection systems. The Internet-based online topic/trend detection systems need to collect and process huge amount of up-to-date data from many heterogeneous websites, which bring complexities.

This research adopts Slow Intelligence [7], which provides a framework to build adaptable systems by adopting an iterative process involving enumeration, propagation, adaptation, elimination, and concentration. In the beginning of the process, many alternatives are enumerated. However, as the process continues, alternatives that can not adapt to environment are eliminated, and finally, limited resources are concentrated on prospect solutions.

In this paper, we propose a design of online topic/trend detection system with the advantages of Slow Intelligence. Four complexities of designing online topic/trend detection systems are identified, along with corresponding Slow Intelligence solutions. The remainder of this paper is organized as follows. Section II shows the design of topic/trend detection system without Slow Intelligence. Section III identifies four complexities of designing trend detection systems, along with corresponding Slow Intelligence solutions. Conclusions are finally drawn in Section IV.

II. TOPIC/TREND DETECTION SYSTEM

The objective of the proposed online topic/trend detection system is to detect current hot topics and to predict future hot topics based on data collected from the Internet. Since it is unlikely to collect all data on the Internet, the system requires users to provide their information needs, including their concerned keywords and their concerned websites. Furthermore, because hot topics change quickly, the system requires periodical updates in hourly or daily intervals.

The system first collects latest data from Internet based on users' information needs by *Crawler & Extractor*, then adopt TDT techniques to discover current hot topics by *Topic Extractor*, and finally apply trend estimation algorithms [4] to predict hot topics by *Trend Detector*.

- ***Crawler & Extractor*** The responsibility of *Crawler* is to collect web pages from Internet. *Crawler* needs to be selective, that is, only collect web pages that satisfy predefined requirements. We plan to implement a focused crawler [5], which selectively collects web pages that are relevant to a pre-defined set of topics. The responsibility of *Extractor* is to extract information from web pages. Since a web page is mixed with information and noisy content (advertisements, navigation stuff, and so on), information extraction algorithms [6] are applied to extract desired information of web pages. The extraction algorithms rely on visual/textual/HTML-syntax features, such as CSS/tag attributes, visual alignments and layouts, for learning the template of web pages and further extracting desired information from identified informative blocks [10, 15, 16]. Information extracted in a single web page is combined into *text documents*, and stored in *Web data DB*. Figure 1 illustrates the workflow of this component.

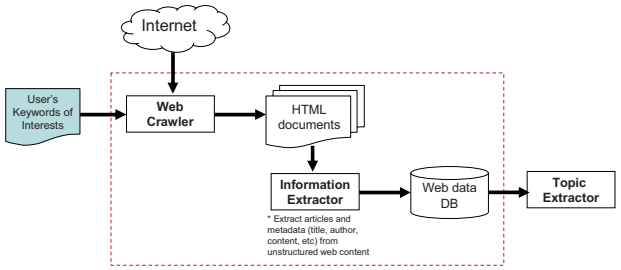


Figure 1: Crawler & Extractor

- ***Topic Extractor*** The responsibility of *Topic Extractor* is to detect hot topics from a set of *text documents*. The process of topic detection can be divided into the following steps, which adopt some state-of-the-art techniques: 1) *topic word extraction*: TF-IDF [13] scheme is applied to measure the importance of terms in a given text document and generates top-N topic word candidates for each text document. 2) *topic word clustering*: single-pass clustering [1, 14], a popular topic detection approach, is adopted to cluster related documents into associated topic groups. The centroid topic word of cluster with highest weighting score is treated as the representative name of each generated cluster, which represents an extracted "topic". 3) *extract hot topics*: hot topics derive from hot events in a particular timeline [1, 9]. We apply the Aging Theory [8] to model the life cycle of an event in a topic, which assumes that an event has a life cycle with four stages: birth, growth, decay and death. To measure the hotness/popularity of each topic, we summarize the energy of related events by considering pervasiveness and topicality of topic words in the events [9] combined with the level of user participation using social data. Figure 2 illustrates the workflow of this component.

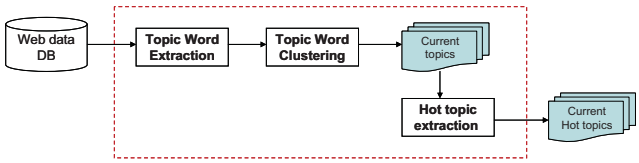


Figure 2: Topic Extractor

- ***Trend Detector*** The responsibility of *trend detector* is to detect trends (future hot topics) based on currently available data. Figure 3 illustrates the workflow of this component.

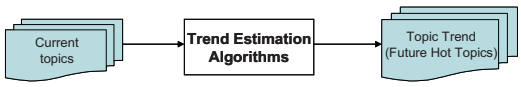


Figure 3: Trend Detector

III. TOPIC/TREND DETECTION SYSTEM WITH SLOW INTELLIGENCE.

In this section, we bring Slow Intelligence into the topic/trend detection system proposed in Section 2. The system needs to collect and process huge amount of up-to-date data from various heterogeneous websites on Internet, hence brings complexities:

1. It is unlikely to collect all web data based on limited amount of computing resources. The system needs to develop data collection strategies which can concentrate limited resources on collecting important web data.
2. Many computation methods are available for estimating trends [4]. If parameter settings are also taken into account, there are too many combinations to choose. Furthermore, Internet is a changing environment, which means current best solution may not perform well in the future. The system needs to automatically find best solution from many alternatives in a changing environment.
3. The crawler needs to revisit websites to collect up-to-date data in hourly or daily intervals. Each site has different amount of to-be-update data and different policy to restrict frequent access, which are unknown beforehand. The system needs to find feasible data collection schedule based on past experience.
4. Any changes in web pages may disrupt *Extractors*. It needs automatic repair mechanism if the many websites are being monitored. The repair mechanism needs to detect errors of *Extractors*, find alternatives, and choose the best solution from alternatives to fix the disrupted *Extractors*.

To conclude, these complexities require the system to: 1) adapt to changing environments; 2) choose from many alternative solutions; and 3) learn from past experiences.

Few technologies can satisfy these requirements. For example, genetic algorithms [12] can enumerate many alternatives and gradually find better solutions, however, past experiences are not considered. Context-aware approaches [11] can only adapt to changing environments while miss the other two requirements.

Fortunately, Slow Intelligence can meet all three requirements. Slow Intelligence provides a framework to build adaptable systems by adopting a iterative process involving enumeration, propagation, adaptation, elimination, and concentration. In the beginning of the process, many alternatives are enumerated. However, as the process continues, alternatives that can not adapt to environment are eliminated, and finally, limited resources are concentrated on prospect solutions. Knowledge gained in the process is stored in the knowledge bases, which help find better solution in next iterations.

Flexibility is another benefit of adopting Slow Intelligence. Slow Intelligence only provides frameworks, which implies developer can freely implement their own domain algorithms and thus Slow Intelligence can be applied to many domain applications..

In the remaining of this section, we introduce four Slow Intelligence subsystems. Each subsystem accordingly targets a complexity discusses in this section. The mappings between the subsystems and the topic/trend detection system are shown in Figure 4.

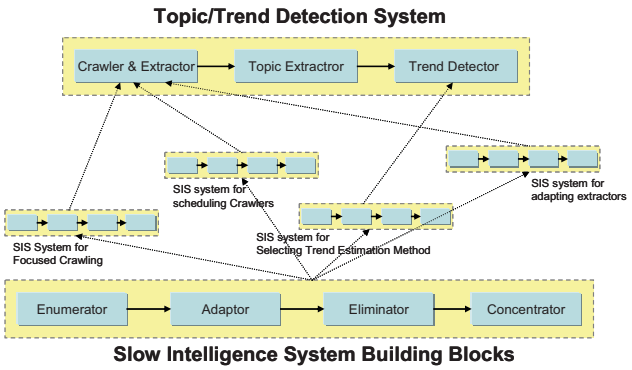


Figure 4: Topic/Trend Detection System with Slow Intelligence

A. *Slow Intelligence Subsystem for Focused Crawling*

Figure 5 illustrates the design of Slow Intelligence subsystem (the area with dotted border) to help concentrate limited computing resources on collecting important data. Data is important if

1. It is related to current hot topics or future hot topics.
2. It matches users' interests.

- **Enumerator**: The enumerator generates candidates of future hot topics. The enumerator takes current topics as candidates of future hot topics.
- **Adaptor**: The adaptor adapts candidates to meet users' interests. Information retrieval techniques [3] can be used to evaluate if candidates meet uses' interests.
- **Eliminator**: The eliminator evaluates candidates based on the selected trend detection algorithms, and eliminates candidates with low probabilities to become hot topics. (We'll discuss how trend detection algorithms are selected and adapted in the next subsystem)
- **Concentrator**: Only few candidates are selected as potential future hot topics. Keywords of these topics are fed into *Crawler*, so that *Crawler* can concentrate on collecting data related future hot topics that users may be interested in.

Two knowledge bases are involved in this subsystem:

- The first knowledge base is a typical SIS knowledge base, which provides essential domain knowledge for SIS. In this subsystem, essential knowledge includes users' interests and available algorithms for detecting trend.
- The second knowledge base stores historical records, including actual hot topics and predicted hot topics. Historical records could help select and adapt trend detection algorithms.

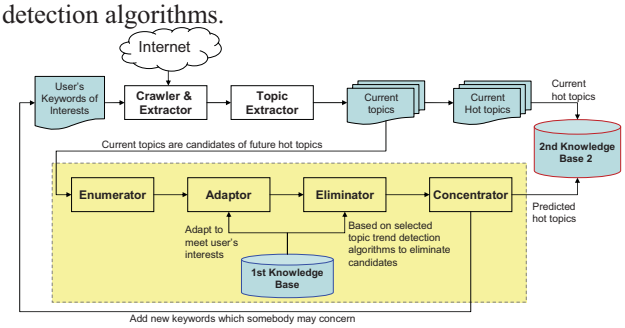


Figure 5: SIS for Focused Crawling

B. *Slow Intelligence Subsystem for Selecting Trend Estimation Method*

Figure 6 illustrates the design of Slow Intelligence subsystem (the area with dotted border) to help select and adapt trend detection algorithms for estimating trends. Historical data, including predicted future hot topics and actual current hot topics, are utilized to select and adapt trend detection algorithms.

- **Enumerator**: The first knowledge base provides the enumerator knowledge about algorithms for estimating trends, including available algorithms and known best parameter settings for algorithms. The enumerator generates candidates of algorithm/parameter combinations for further evaluation.
- **Eliminator**: Historical records (including actual hot topics and predicted hot topics), which are stored in the second knowledge base, are useful for measuring performance of candidates. A sample of historical records is used to evaluate performance of candidates. Poorly-performed candidates are eliminated.
- **Concentrator**: Extensive evaluations are conducted on candidates, and the evaluation results are preserved in the first knowledge base.

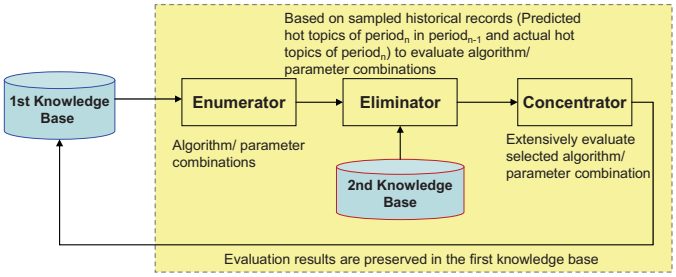


Figure 6: SIS for Selecting Trend Estimation Method

C. *Slow Intelligence Subsystem for Schedulging Crawlers*

In practice, the crawling schedule for each data source varies depending on characteristics of each website, for example, blocking policy of each website and data amount to be collected in each website.

In the cases of collecting data from websites with severe blocking policies, the crawling behavior and schedule of crawlers should be designed as similar as humans' access behavior to prevent being blocked. For instance, a target website may block accesses which send more than one requests in two second; therefore it would be smarter for crawlers to set a two-to-three-second interval between requests.

Furthermore, contradictory conditions exist. For instance, in the cases of collecting data from websites with frequent updates, *Crawler* needs to visit these websites more often than others, which may also cause *Crawler* being blocked.

Notably, characteristics of websites may change over time. However, it is impossible to manually configure efficient schedules for large-scale systems involving many targeting websites. A knowledge base with basic scheduling rules can aid the system to find acceptable scheduling plans in short-term cycles. The initial building of this knowledge base may need human involvement; however, feedbacks from *Crawler*

can further enrich the knowledge base for developing scheduling rules in long-term cycles.

Figure 7 illustrates the design of Slow Intelligence subsystem to help *Crawler* arrange schedules to efficiently collect up-to-date data while preventing being blocked by websites.

- **Enumerator**: The enumerator generates candidate schedules of data collection following scheduling rules in the knowledge base. The scheduling rules include general rules, learned heuristic rules, and user-defined rules.
- **Adaptor**: The adaptor adapts candidate schedules based on the observed abnormal behaviors of crawlers, for example, being blocked by websites or unexpected delay.
- **Concentrator**: The concentrator concentrates on the most efficient candidate schedule. The efficiency of candidate schedule is evaluated based on the estimated time to complete the task.

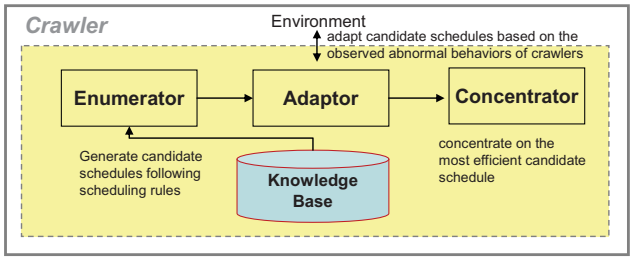


Figure 7: SIS for Scheduling Crawlers

D. *Slow Intelligence Subsystem for Adapting Extractors*

Extractor applies extraction algorithms, which use parameterized features to learn the template of web pages, to extract desired information. Once web pages change template, adequate features and optimal parameters have to be investigated and revised.

Figure 8 illustrates the design of Slow Intelligence subsystem to help *Extractor* adapt to changes in web pages.

- **Enumerator**: The enumerator generates feature/parameter settings that may fit the change.
- **Eliminator**: The eliminator eliminates candidates that can not successfully extract data based on trials in a few example pages.
- **Concentrator:** The concentrator concentrates on the selected candidates on a large scale of pages.

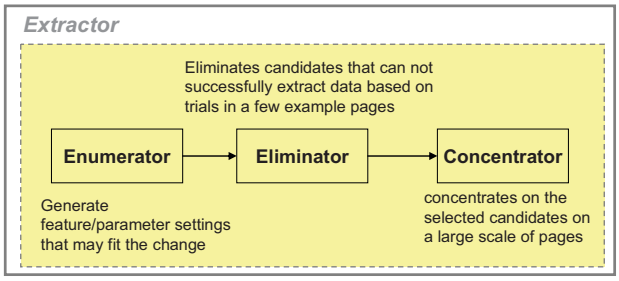


Figure 8: SIS for Adapting Extractors

IV. CONCLUSIONS

Although Internet provides abundant data for topic/trend detection, it incurs challenges on designing online topic/trend detection system. An online trend detection system requires careful resource allocation and automatic algorithm adaptation to process huge size of heterogeneous data. This research adopts Slow Intelligence, which provides a framework for systems with insufficient computing resources to gradually adapt to environments, to response the challenges. Four Slow Intelligence subsystems are proposed, and each subsystem targets a challenge in designing online topic/trend detection systems. In the future, we plan to implement the topic/trend detection system to empirically verify the four proposed Slow Intelligence subsystems. We also plan to include multimedia data into the topic/trend detection system.

ACKNOWLEDGMENT

This study is conducted under the "Digital Convergence Service Open Platform" project of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

REFERENCES

[1] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y., "Topic detection and tracking pilot study: Final report," *Proceedings of the DARPA broadcast news transcription and understanding workshop*, 1998.

[2] Asur, S. and Huberman, B., "Predicting the Future With Social Media," to appear in *Proceedings of WI-IAT,* 2010.

[3] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern information retrieval*: Addison Wesley, 1999.

[4] Bianchi, M., Boyle, M., and Hollingsworth, D., "A comparison of methods for trend estimation," *Applied Economics Letters,* vol. 6, pp. 103-109, 1999.

[5] Chakrabarti, S., Van den Berg, M., and Dom, B., "Focused crawling: a new approach to topic-specific Web resource discovery," *Computer Networks,* vol. 31, pp. 1623-1640, 1999.

[6] Chang, C. H., Kayed, M., Girgis, M. R., and Shaalan, K. F., "A survey of web information extraction systems," *IEEE TKDE,* vol. 18, pp. 1411-1428, 2006.

[7] Chang, S. K., "A General Framework for Slow Intelligence Systems," *IJSEKE,* vol. 20, pp. 1-15, 2010.

[8] Chen, C. C., Chen, Y. T., and Chen, M. C., "An aging theory for event life-cycle modeling," *IEEE Transactions on Systems, Man and Cybernetics, Part A,* vol. 37, pp. 237-248, 2007.

[9] Chen, K. Y., Luesukprasert, L., and Chou, S., "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," *IEEE TKDE,* vol. 19, pp. 1016-1025, 2007.

[10] Deng, C. A. I., Shipeng, Y. U., Ji-Rong, W., and Wei-Ying, M., "VIPS: a vision-based page segmentation algorithm."

[11] Dey, A. K., Abowd, G. D., and Salber, D., "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," *Human-Computer Interaction,* vol. 16, pp. 97-166, 2001.

[12] Mitchell, M., "An Introduction to Genetic Algorithms," MIT Press, 1998.

[13] Salton, G. and Yang, C. S., "On the specification of term values in automatic indexing," *Journal of documentation,* vol. 29, pp. 351-372, 1973.

[14] Smith, D. A., "Detecting and browsing events in unstructured text," *Proceedings of SIGIR*, 2002, pp. 73-80.

[15] Song, R., Liu, H., Wen, J. R., and Ma, W. Y., "Learning block importance models for web pages," *Proceedings of WWW*, 2004, p. 211.

[16] Zhai, Y. and Liu, B., "Web data extraction based on partial tree alignment," *Proceedings of WWW*, 2005, p. 85.

A Generic Architecture for Surveillance Systems

Erland Jungert, Christina Grönwall, Niklas Hallberg, Birgitta Kylesten, Fredrik Lantz, Lars Eriksson
FOI (Swedish Defence Research Agency)
Box 1165, S-581 11 Linköping, Sweden
{jungert, stina, nikha, birgitta.kylesten, flantz}@foi.se

Abstract: Surveillance systems are generally intended for the protection of physical objects such as stationary installations and transport vehicles. Development of surveillance systems is an extensive development process, which involves considerations of a large number of design aspects that relates to a variety of disciplines, e.g. sensors systems, communications, decision-support functionality, multimedia usage and design and user interaction. Furthermore, there is an extensive need for a comprehensive understanding of the usage domain. Needs and requirements analyses are of utmost importance when developing surveillance systems. The objective of the work presented is a generic architecture for surveillance systems. The outline of the architecture is based on needs assessment and on analysis of required capabilities. The architecture comprehensively describes how surveillance capabilities are achieved by processes, actors and systems support.

Keywords: Surveillance systems, surveillance systems capabilities, emergency management

1. INTRODUCTION

Systems for surveillance and protection of facilities and critical infrastructures, such as plants, power lines and railways, must possess a number of different capabilities. These capabilities must include *detection* of events deviating from normal conditions, but also measures to *respond* to different events by *warn* the operators and *prevent* or at least *delay* antagonist to reach their aim. Of utmost importance is that the system support detection of threats as early as possible. If threats are realized the responsible staff and relevant external actors must be alerted. Development of this type of systems requires a thorough analysis and design of the system structure.

In this paper a generic architecture of surveillance systems, given the required capabilities mentioned above, is presented and discussed. Of specific importance when designing architectures of this type is not only to allow for early detection of the threats but also to support surveillance at all times and light conditions. To make this possible relevant information must be collected and the surveillance systems must allow the users to work under a minimum of stress so that appropriate decisions can be taken [1]. In this work, the aim is not just the surveillance capabilities for protection of the facilities but also to determine the capabilities that support the handling of the various types of occurred incidents, to secure the facilities prior to and during an incident. For this reason, we are using the term physical security for this purpose.

According to Jordan et al. [2], information that need be collected must include type and number of intruders, their possible purposes, physical ability and location, but also the context information. Other works that relate to this work is, e.g., made by Williams [3], who has studied aspects of powerful decision support techniques for surveillance systems. Torbin [4] discussed approaches to surveillance systems for transports while Jungert, Hallberg and Hunstad [5] have studied aspects of service oriented architectures for command and control systems.

2. METHOD

The work was performed in six activities: (1) Data collection, (2) Needs assessment, (3) Workshop with domain expertise, (4) Modeling capabilities, (5) Modeling actors and processes, and (6) Modeling systems and functionality. The activities (4) to (6) were carried out iteratively.

2.1. Data collection

The first step in this activity was to determine which stakeholder and user representatives who should be interviewed. Seven responders were selected for interviews, representing (1) responsible for the systems at surveillance centers, (2) head of transport, (3) security managers, and (4) expert on the aspects of human-technology-organization. The interviews were performed to obtain information used as input for the needs assessment. For each role, slightly adapted interview questions were used regarding: (1) the respondent's role and mission, (2) actual management of security and safety, (3) routines for security and safety, (4) security systems, (5) information they utilization, and (6) what is working well and what can be improved regarding security and safety.

2.2. Needs assessment

The needs assessment was carried out in three steps. First, the interview notes were analyzed to identify statements that could be used to reveal needs. Second, the statements were carefully analyzed to interpret what needs the statements express [6][7]. Finally, the identified needs were categorized and structured in a hierarchic order. In this step, duplicates were removed and formulations were uniformed.

2.3. Workshop with domain expertise

A workshop with domain experts was performed to (1) clarify uncertainties that had been found during the needs assessment, (2) validate the identified needs, and (3) to establish which of those needs that should be acknowledged in the further work. During the workshop, the identified needs were presented to the domain experts, who were asked to comment on each need regarding if it was true need or not. If it was found to be a true need, they were asked that determine if the need should be used in the further development, by prioritize the importance of meeting each need on a scale of "not relevant", "relevant" or "highly relevant".

2.4. Modeling capabilities

Based on the given needs the highest priority, corresponding capabilities were identified, i.e., capabilities that the surveillance system should have. The identified capabilities were structured hierarchically. The modeling was influenced by the Ministry of Defense Architecture Framework (MODAF) [8].

2.5. Modeling actors and processes

Based on the capabilities and the interview notes, the actors and their information exchange were identified and modeled. Further, the surveillance processes were also determined and modeled.

2.6. Modeling systems and functionality

Based on the modeled capabilities and processes, systems components and their functions were identified and modeled. Hence, these models describe the functions that the system provides for the performance of the surveillance.

3. THE ARCHITECTURE

This section presents the architecture for surveillance systems, including the *Strategic Views*, *Operational Views*, *and Systems Views* as defined in MODAF including the mapping between capabilities and systems functions. The Strategic Views includes the capabilities that the surveillance should have. The Operational Views include actors, their information exchange, a general process of the surveillance, and handling of episodes. The Systems Views include systems components and their functionality.

3.1 Capabilities

Capabilities correspond to the functionality and support that systems provide, without expressing how this is accomplished. The capability models show how capabilities are divided into underlying capabilities, i.e., the capabilities that realize the higher level capabilities.

The physical security system consists of protection of personnel, property and information, but not the ordinary management/operation and maintenance. In the description of the capabilities, we assume that the system contains operators, software, computers, communication HW and SW, sensors and databases. A capability can be something that is performed by an operator alone, something that is performed on an operator's command, or something that is performed in hardware and software automatically. An example of capability is *Manage a fire alarm*.

The capabilities defined in this work originate from the needs assessment. Three main capabilities were identified (Figure 1):

1. *Manage normal events* that occur several times a normal day. This includes taking care of visitors, normal patrolling and checks of the facilities.
2. *Manage deviating events* need extra attention and investigation. This includes alarms, sensor signaling, and strange behavior from visitors. A deviating event will be redefined either as a false alarm or an incident.
3. *Manage incidents* correspond to handling events that need actions from an operator. Some incidents can be taken care of by an operator himself, like double check that a room is empty, while others like fires and trespassing, need support by external personnel.

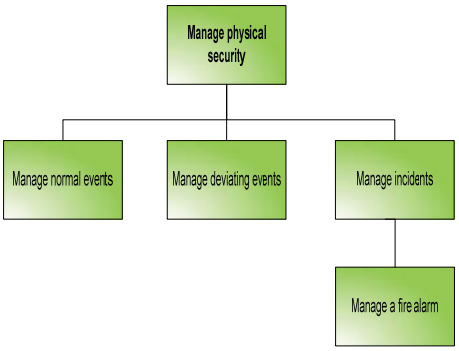


Figure 1. The top-level capabilities to manage *physical security*.

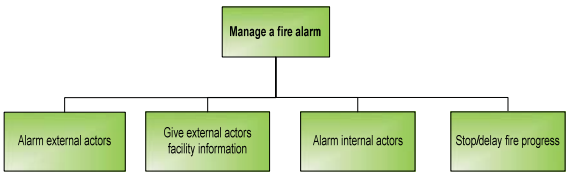


Figure 2. Capabilities needed to manage a fire alarm.

As an example, the capability *Manage a fire alarm* is broken down into sub-capabilities in Figure 2. The lowest capability level is still general enough to fit for several types of facilities within the secured area. The analysis of needs for a specific facility is taken care of in the service-oriented analysis.

3.2 Actors, process and episodes

This section presents the involved actors and their information exchange. Further, it also presents the surveillance process and the episode concept.

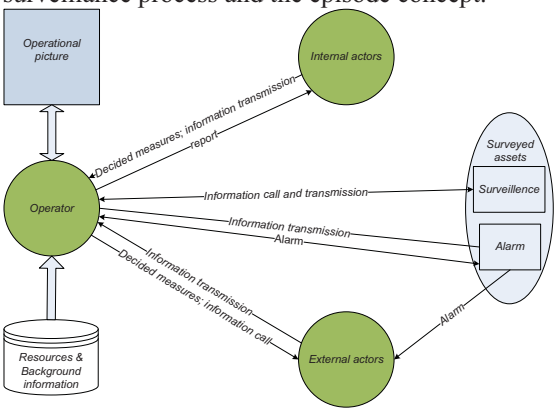


Figure 3. The conceptual description of actors of the surveillance system and the facilities subject to surveillance.

3.2.1 Actors

Actors involved in the protection of the facilities in focus include operator, internal actors and external actors (Figure 3). The operator of the surveillance system is generally responsible for the execution of the on-going activities and controls these activities through various communication activities with the internal and external actors. External actors are primarily the fire brigades, police forces or medical responders. The internal actors belong to the own organization set up by the management of the enterprise responsible for the facilities subject to surveillance. For example in case of an incident, the internal actors include personnel in command of the operations. The internal actors have mainly strategic responsibilities while the operator has the tactical responsibilities. The external actors have strictly operational responsibilities in what concerns the hands on management of an incident.

Information available to the operator includes an operational picture, information corresponding to available physical resources and all types of background information. The operator can request further information from the other actors but various sensors can also contribute.

3.2.2 The general process

The main target of the surveillance process is to determine events that deviate from the normal and if such a deviation is detected determine whether it is an incident or a false alarm (Figure 4). A false alarm should be managed, including documentation and report of the event. After this, the operator will resume the surveillance responsibilities. In the former case the operator starts to handle the incident following the instructions available in the *Handbook* [9], which can be seen as an on-line and knowledge based manual. The management of the incident will continue until the incident has come to an end after which the incident is reported. More than one incident must be possible to handle in parallel; besides this the surveillance must also go on during the management of incidents.

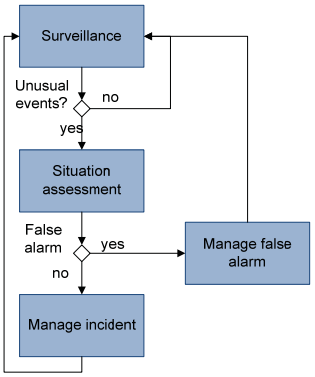


Figure 4. The general model of the surveillance process.

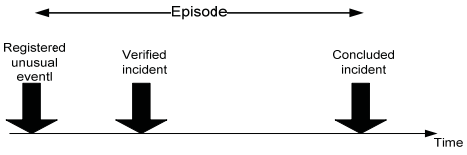


Figure 5. The extension of an episode over time.

3.2.3 Episodes

Deviating events must be verified and handled accordingly. The activities following an event will go on for a certain period in time. In this paper, this type of period is called *episode*. An episode describes the flow of events and activities that transpire during the concurrency of an incident. An episode starts once a deviating event is registered. In case of a false alarm it may end fairly quickly. If the episode corresponds to an incident it will go on until the incident can be

concluded. Figure 5 illustrates the extension of an episode over time.

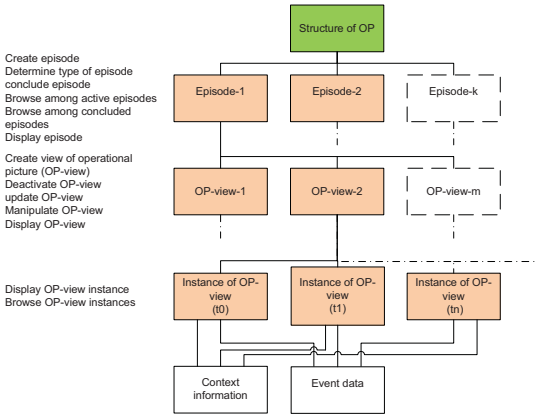


Figure 6. Structure of the operational picture, its information bases (at the bottom) and the corresponding system functions.

Surveillance systems must be able to handle episodes, both solitary as well as contemporaneously. Consequently, there must be functions available for manipulation and visualization of the information related to the episodes. Furthermore, as the episode information concerns operational information with given coordinates the visualization of this information should be visualized in the operational picture (OP) (Figure 6). Hence, the OP describes the ongoing episode during its existence by means of information collected by the surveillance system and participating actors.

The OP support must thus be able to handle a set of parallel episodes and for each episode a set of subordinate OP views. There each view present information concerning various aspects of the episode using different media, e.g. maps, tables or images but also voice or text messages. Over time, for each view, new instances of the different views are generated. A view concerns the incident information to be presented and should not be confused with the views in as used in MODAF. To support the visualization of the OP context information and additional event related data must be accessible (Figure 6).

3.3 Systems views

The architecture of the surveillance system can be viewed either as a hierarchical structure including decision support tools and other system functions or structured as a system control loop [10]. Both aspects will be described subsequently together with the visual user interface of the system.

3.3.1 System function hierarchy

The system includes a set of functions, which in turn includes a set of decision support tools. The decision support tools in the system are:

- Visualization of the OP,
- Handbook,
- Communication tool,
- Information search and threat analysis,
- Support for reporting actions and orders.

.

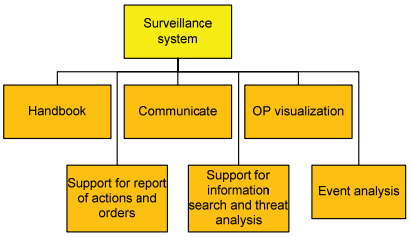


Figure 7. The top level of the system function structure.

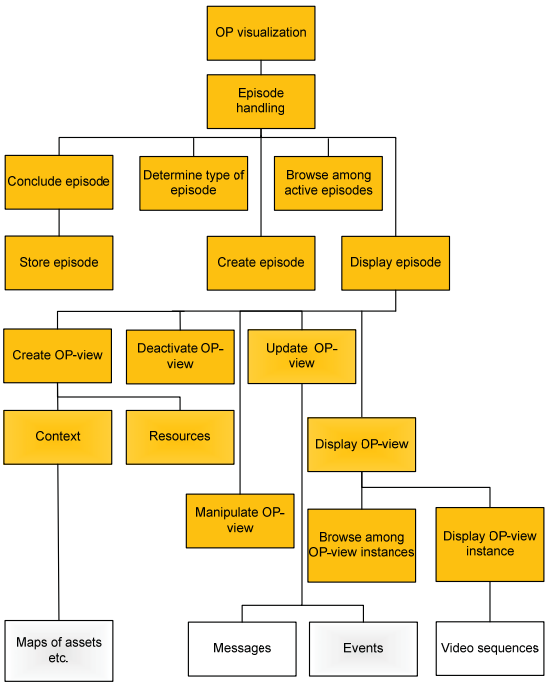


Figure 8. The hierarchy of the OP visualization functions and below the accessible information.

Figure 7 illustrates the top level of the system function hierarchy with its functions and decision support tools. Figure 8 presents the lower levels of the OP visualization structure including the required data sets for the generation of episode oriented views and view instances. The given functions correspond to the set of functions, as in Figure 6, operating on the different levels of the episode structure

3.3.2 System control loop

The control loop structure of the surveillance system can be seen in Figure 9. Logically, this structure is built up by two main modules, i.e., the system support module and the decision support module. Integrated to these two modules is the control loop. The system support module handles the basic incoming

information and the various local data sources required to support the work. The decision support module corresponds to the user interaction part of the system, intended to support the users.

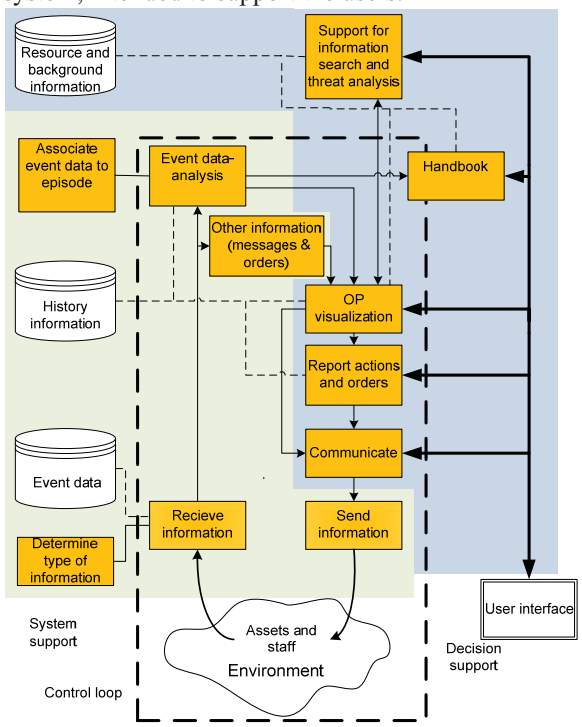


Figure 9. The surveillance system.

3.3.3 User environment

In this section the user environment with its visual user interface will be described and illustrated. The user environment is one of the main parts of the decision support module. The given illustration can also be seen as part of a simple scenario.

To reach the objective of an effective and efficient user environment in a surveillance system an iterative user-centered approach for several design cycles is required. The following description of the operator environment and the system functions should be considered as possible starting points for the development of surveillance systems. This description is based on the capabilities and functions that the present work identified as necessary for the monitoring of the system.

Figure 10 shows the operator environment with a large screen that presents important and processed information for a good overview. This overview is useful both for the operator and other actors. The monitors closest to the operator are used for presentation of information that the operator classifies, verifies and uses to perform tasks related to a specific incident.

The example in Figure 10 shows four incidents taking place simultaneously. They are illustrated in the operator's information presentations with yellow (indicating lower priority), recently occurred events are presented in orange, and red is used for currently prioritized events. Sensors indicate an apparent deviation from the normal conditions. The discrepancy in the form of an alert is displayed on a large screen and on one of the smaller displays. Incident management is handled by the operator who selects the sensors related to indicate deviations, and the operator can e.g., zoom a camera and also receive and make telephone calls from the staff. The operator alerts relevant external and internal actors. The operator steps through the reporting framework, video sequences and other sensor information, to update the current state of events, to update the corresponding episode instance. Communication is carried out via telephone and radio as well as by means of computerized media for the passing of messages, pictures, maps, and video sequences.

When a deviation from normal conditions is detected and categorized, the Handbook will automatically show the section describing how to handle the incident type as classified by the operator.

The operator must provide a checklist for actions and considerations that must be taken and can also browse the Handbook for additional information. After external and internal actors have been informed the operator processes and analyzes the situation and the generated overall picture (represented by the overview images). When the operator has specified the type of episode or situation that is at hand the system couples this information automatically to the Handbook, the actions checklist, map and other relevant documents. The review of the action checklist for the incident can be used at the handover to another operator as well. The overall picture is created for the use of from maps and/or photographs of the surroundings and the construction of the actual facility, including architectural drawings, other accompanying messages and notes. Information on available resources (e.g. available personnel, equipment) etc, is gradually updated and forwarded to requesting actors.

While the incident is ongoing reports of various types are sent and received; concerning such aspects as m minutes on operations, delegations of incoming orders, and measures to be taken. The checklist is followed with each action registered in terms of point in time and type of action performed.

For each episode its episode information is stored and a report is automatically filled in, by means of a template, and finally checked by the operator. The operator also reports incidents that should be especially considered and that may form a basis for the revision of the manual, the checklist, or for training purposes etc.

3.4 Mapping capabilities to functions

The development of the surveillance system architecture requires mapping of the capabilities into sets of functions. The mapping of the capabilities into sets of system functions is illustrated in Figure 11 that shows the result of the mapping of the *survey* capability into its corresponding function set including OP visualization, Communicate and Event analysis. The work process in this case is "survey". Other similar mappings have been identified as well, for example:

- surveillance (of facilities)
- manage deviating events
- management and surveillance (of incidents)
- attend incidents

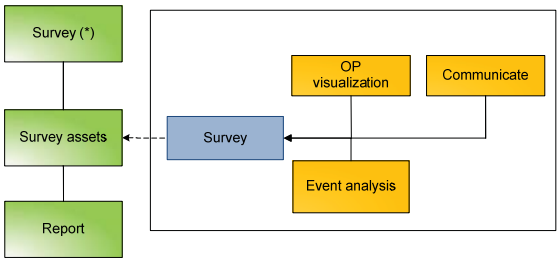


Figure 11. The surveillance capability and its mapping to its corresponding system functions (marked with a box) via the surveillance process.

4. CONCLUSIONS

This works demonstrates the development of a surveillance system architecture based on identified needs and capabilities. The architecture involves a set of system functions, decision support tools and a visual user interface. The work also shows the involvement of the various categories of actors that depend on the system during occurring incidents. Additional to this, it is also shown that the step from capabilities to the visual user interface (VUI) is tractable, that is it will be possible to trace the capabilities from the system functions in the VUI and the other way around. Furthermore, we believe that the proposed VUI is easy to use and understand, even under stressful situations. The motivation for this is based on the capability that allows for a separation between on-going incidents in the user interface, i.e. the operator can arbitrarily switch between incidents and concentrate the work on one incident at a time.

However, further work is needed to finalize the architecture and for that reason the development of a demonstrator is proposed.

ACKNOWLEDGEMENT

This work has been carried out as a research project on commission of SKB (Swedish Nuclear Fuel and Waste Management Company).

REFERENCES

1. Garcia, M. L., The design and evaluation of physical protection systems, Elsevier, Burlington, MA, 2001.
2. Jordan, S. E., Snell, M. K., and Madsen, M. M., Discrete-Event Simulation for the Design and Evaluation of Physical Protection Systems, proceedings of the 1998 Winter Simulation Conference, pp. 899–904.
3. Williams, J. D., Physical protection system design and evaluation, proceedings of the International Conference on Physical Protection of Nuclear Materials: Experience in Regulation, Implementation and operations, Vienna 10-12 November, 1997.
4. Torbin, R., Urban infrastructure security, Journal of Technology in Society, vol. 25, pp 549–552.
5. Jungert, E., Hallberg, N., and Hunstad, A., A Service-based Command and Control Systems Architecture for Crisis Management, The International Journal of Emergency Management, vol. 3, no. 2, 2006, pp 131-148.
6. Kano, N., Upsizing the Organization by Attractive Quality Creation, proceedings of the First World Congress on Total Quality Management, Champman & Hall, London, 1995, pp 60-72.
7. Hallberg, N., Timpka, T. and Eriksson, H., The Medical Software Quality Deployment Method. Methods of Information in Medicine, vol. 38, no. 1, 1999, pp 66-73.
8. Ministry of Defense Architecture Framework (MODAF), http://www.mod.uk/modaf (2010-03-17).
9. Chang, S.-K. and Jungert, E., A Self-Organizing Approach to Mission Initialization and Control in Emergency Management, proceedings of the International Conference on Distributed Multi-media Systems, San Francisco, CA, September 6-8, 2007.
10. Hollnagel, E. and Woods, D. D., Joint Cognitive Systems: Foundations of Cognitive Systems Engineering, CRC Press, New York, 2005.

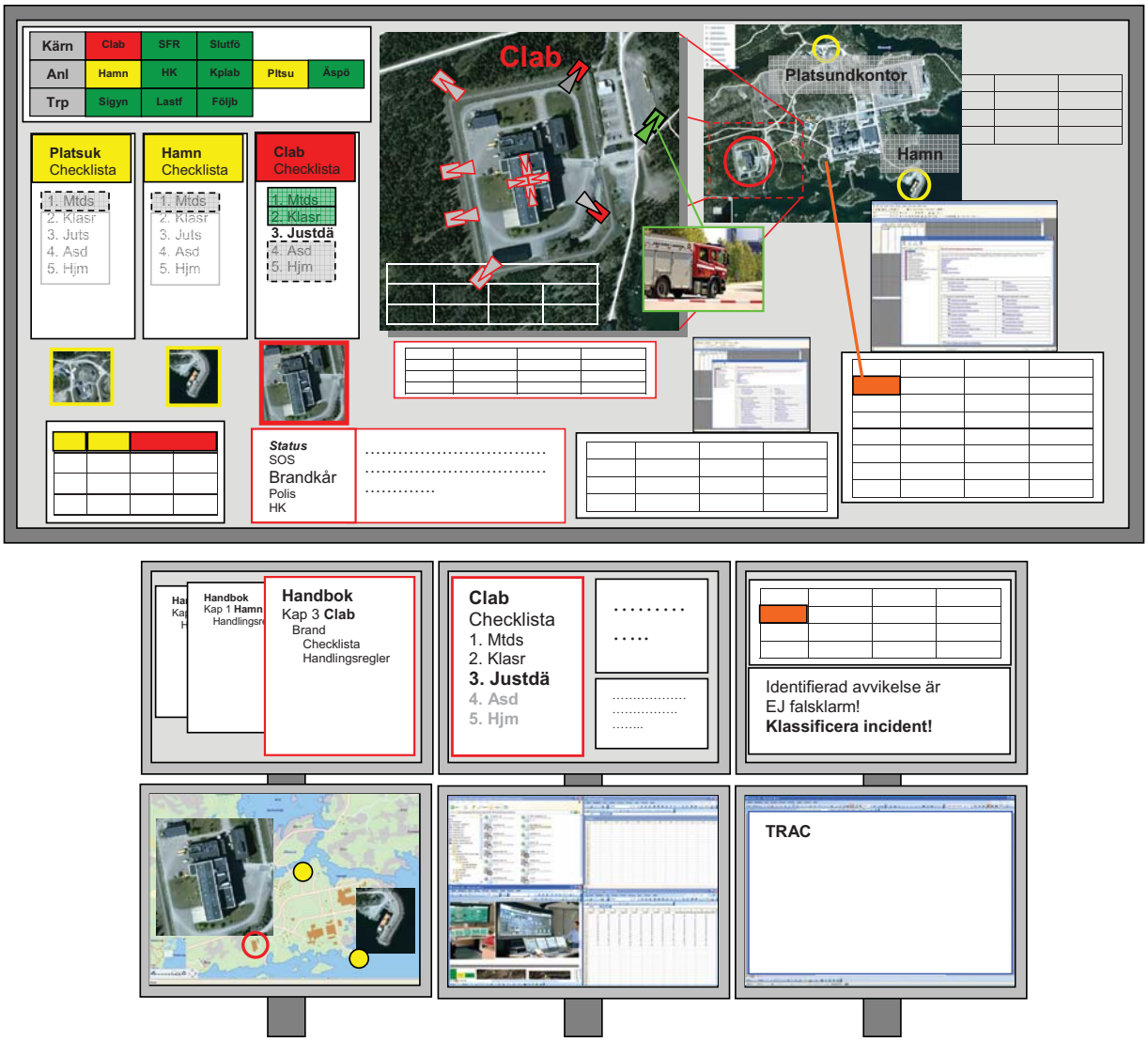


Figure 10. An example of a configuration of the operator environment and an overview of the suggested information presentations.

Privacy Enforcement by Identity Management in Smart Surveillance Systems

H. Vagts*,†, E. Krempel* and J. Beyerer *,†
*Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Fraunhoferstr. 1, Karlsruhe, Germany
†KIT – Institute for Anthropomatics , Adenauerring 4, 76131 Karlsruhe, Germany
Email: {Hauke.Vagts; Erik.Krempel; Juergen.Beyerer}@iosb.fraunhofer.de

***Abstract*—Surveillance Systems have become increasingly powerful. Conventional camera-based systems are extended with different types of (wireless) sensors, the number of data sources increased and hardware as well as algorithms improved. Such smart systems do not only threaten privacy, but technological improvements can also increase privacy of observed objects. Previous work to maintain privacy have focused on video surveillance only, which is not sufficient for modern multi-sensor systems. Hence, privacy must be enforced on a higher level of abstraction. In this work, a task- and service-oriented architecture is described, which hosts an *Object-Oriented World Model (OOWM)* to store information from different sources on a high level. To ensure privacy, access to the OOWM is only granted via a *Privacy Manager*, Identity Management (IdM) beeing one of its most important functionalities. Recently, privacy and Identity Management have been in the focus of research and many approaches for e-commerce have been proposed. IdM in the area of surveillance is different and no sufficient solutions exist by now. This work identifies the privacy issues that must be addressed by Identity Management solutions in smart surveillance systems. A new approach for privacy-aware IdM is proposed that copes with the new requirements and shows its integration in a task-oriented surveillance system.**

***Index Terms*—privacy and security, surveillance, Identity Management, distributed multimedia models and systems, web-services and multi-agent systems, emergency and safety**

I. Introduction

The term "identity" is used with different meanings, throughout this work the definition of Cameron [1] is taken as basis: "[. . .] defining a digital identity as a set of claims made by one digital subject about itself or another digital subject." As shown in Figure 1, every digital subject has multiple digital identities and every identity is specified by a number of claims. In the area of smart surveillance, a claim can be any measurable attribute, e. g., hair color, height, name or position. In the shown architecture, every identity has a unique attribute ID. These IDs are used to differentiate objects in the system.

The general objective of *Identity Management (IdM)* is the assignment of identities to entities and vice versa. This is done for a large variety of tasks. Accordingly, *Identity Management Systems (IMS)* take, for instance, care of the following tasks [2]: Management of different identities, protection of identities, authentication, anonymization, data synchronization between multiple IMS (parties) and filtering of identities. However, the requirements for IMS are heterogeneous. Three different types can be distinguished [3]:

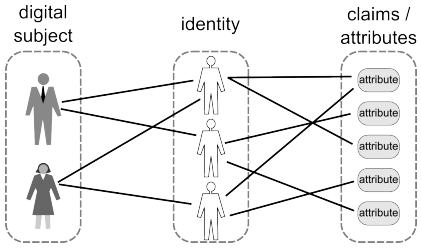


Fig. 1. Digital Subject, Digital identity and Attributes

Type 1: IMS for account management
Type 2: IMS for profiling of user data for an organization
Type 3: IMS for user-controlled context-dependent role and pseudonym management

Systems of type 1 or 2 are mostly used by large organizations and users have few control over their data. Systems of type 3 support the user, focusing on data protection. In the area of surveillance IdM and IMS, respectively, are neither of the named types rather IMS is a mixture of them. On the one hand, operators want to collect as much data as possible about subjects (profiling), using information from existing identities (e. g., data bases of employees). On the other hand, surveillance systems need to maximize privacy and data protection to be compliant with privacy regulations and to achieve user acceptance. Hence, a new concept for privacy-aware Identity Management that meets these requirements is needed.

In this work, a system architecture developed for the requirements of the smart surveillance is proposed. Then, new privacy challenges in smart surveillance systems are highlighted. Subsequently, a new concept of Identity Management is shown that can handle these challenges. In conclusion, a demonstration scenario is presented that has been implemented as a proof of concept.

A. Smart surveillance

When using conventional surveillance systems, human operators mostly have to perform tasks manually. They might have to take care of hundreds of cameras to track suspicious persons, detect unattended objects or the similar. Human operators can only observe a limited number of monitors and have decreasing awareness after a short time. These limitations lead to a new approach in surveillance. Smart surveillance systems

aim to support human operators by performing surveillance tasks on their own. The system captures data, processes them and helps the operators to detect important events. Thus, the operators can keep track of the whole site, while the surveillance assists them collecting and abstracting data. Another benefit of smart surveillance systems is the integration of further sensors, e. g., GPS, RFID or acoustic sensors. While the raw data of these sensors are mostly useless for human operators, an automatic system can evaluate and use it to perform certain tasks.

B. Privacy Issues

Smart surveillance systems supervise a huge amount of sensors and the information is evaluated autonomously. Hence, smart surveillance leads to new privacy challenges [4]. Surveillance systems must adhere to legal restrictions and must be compliant with data protection requirements. However, the legal situation is different throughout the world, thus the Fair Information Practices Principles (FIP, [5]) that have been published by the OECD can still be seen as the minimum requirements for privacy. These eight principles should be adhered to by every surveillance system.

Since existing approaches for privacy have focused on video surveillance, new methods for privacy enforcement in smart surveillance are required.

C. Related Work

Privacy in surveillance is a recent area of research and new solutions are required. In the area of video surveillance, some approaches exist to ensure privacy and security, but most of the approaches blur *regions of interest* (RoI) that might imperil privacy. In [6], Senior et al. propose a "privacy-preserving console" for video surveillance. The console rerenders the video stream and hides sensitive details detected by video analysis. Depending on the authorization level, access is granted to rerendered videos (e. g. with blurred faces or even enriched with additional information) or the raw video stream. They also purpose a "privacy cam", which processes the video sources and transmits encrypted information streams. In [7], Chattopadhyay and Bould also present a privacy cam, which is implemented on a Blackfin DSP and blurs RoI based on PICO [8]. Another scrambling approach is presented in [9]. In [10], Schiff et al. propose a respectful camera, which reacts on visual markers worn by the subjects. In [11], the usage of "talking cameras" is reported, if a camera detects motion, it sends an acoustic message to a subject. Even if such cameras should prevent vandalism, they can ensure privacy as well, e. g., a camera can count down vocalized, before it starts recording. Fleck's approach to privacy [12] is based on smart cameras, which transmit events instead of video data. Fidaleo et al. present in [13] a privacy-enhanced software architecture with a centralized server that hosts a privacy buffer, which can remove private or identifiable information from the stream. In [14], Vagts et al. propose a framework for privacy enforcement in smart surveillance systems.

In the area of Identity Management, multiple different approaches and commercial solutions exist. One example of an IMS of type 1 is OpenSSO[1]. The software, developed by SUN, realizes Single-Sign-On and also provides account management. Popular for profiling (type 2) are loyal reward shopping programs, e. g., Payback[2], but profiling is also used in other areas, e. g., to determine money laundering[3] or financing of terrorists[4]. An example for an IMS of type 3 is the iManager [15]. The IMS is especially developed for mobile devices and enables the user to manage different digital identities.

II. Smart Surveillance

To cope with the requirements of smart surveillance, a task-based approach is very promising. In the following, an architecture is sketched that implements it. More details about the architecture can be found in [16].

A. A Task-Oriented Approach for Surveillance

Most surveillance systems still perform a sensor-oriented approach, i. e., as many data as possible is collected and stored. Intelligent algorithms are then used to extract the relevant and required information. When using a huge amount of sensors, real-time processing is not possible. Following the task-orient approach any usage of a resource and each processing step are assigned to a concrete surveillance task. This leads to two great advantages: On the one hand, resources can be used more efficiently, e. g., if one specific person should be tracked in a central station. A non task-oriented approach examines the entire scene including the requested person. A task-oriented approach monitors only the relevant person and ignores the others.

One the other hand, the approach offers great possibilities for enhancement of privacy. It is required by law and the FIP that the purpose of a surveillance task is specified before the task is executed. If a task is specified strictly according to the purpose, a task-oriented System can ensure best possible privacy and data protection for the user subjects. As processing is task-related, person-related data can be isolated in case of multiple surveillance tasks and privacy protection mechanism can be established very granularly according to the requirements of the task. Hence, a task-oriented system is efficient and privacy-aware [14].

B. An Event-Driven SOA Architecture for Surveillance

For implementation of a task-oriented surveillance, the service-oriented approach shows potential. Surveillance tasks can be specified in workflows and they operate on a high level of abstraction. Figure 2 shows a demonstration scenario of the SOA-based surveillance System NEST (Network Enabled Surveillance and Tracking) [16]. In the demonstration scenario, one person is tracked while walking from a reception

[1] http://www.oracle.com/technology/products/id_mgmt/index.html
[2] http://www.payback.de
[3] http://www.antimoneylaundering.net
[4] http://www.riskvalues.com

desk to a specific destination. To fulfill the task, multiple services are used, e. g., for planning possible routes (route planing service), for the tracking itself (person tracking service) or for determining that the observed person has lost its way (route surveillance service). If the latter is happening, an event is generated, which signals the operator that something important has happened. As it can be seen in Figure 3, information about objects is stored in the OOWM and is provided to the services. Besides the services, which specify the surveillance task, other services are used to encapsulate sensors and to provide a standardized interface (sensor services). In the following, such sensor services are just called sensors. Due to the flexible SOA design, different sensor types can easily be integrated and a huge amount of data sources can be linked into the system. The fundamental idea is that different surveillance tasks can be described as automated processes with a minimum of human interaction involved. The operator choses the suitable services (to build a task) and the system performs the necessary operations automatically.

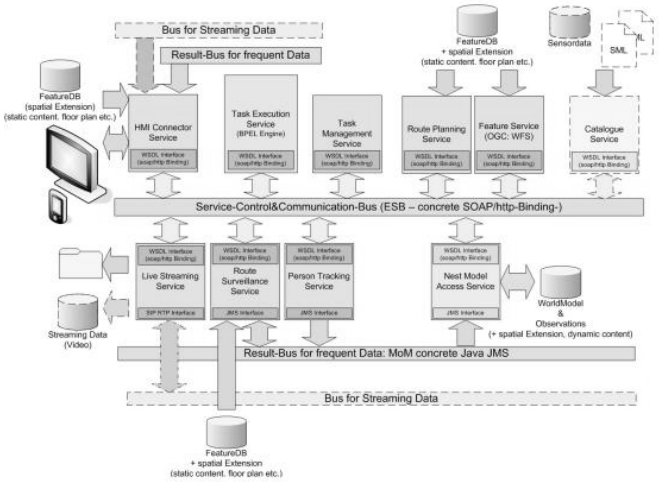


Fig. 2. NEST Architecture

C. Object-Oriented World Model

An Oject-Oriented World Model, as part of a surveillance system, is meant to transform object observations collected from multiple and heterogeneous sensor systems into a consistent object representation, in order to provide a unified information source for application-level software (services). Sensor systems in this context are defined as any combination of physical sensor deployments and corresponding signal-processing software, able to extract object information from raw sensor data (e. g., person tracking, face recognition, etc.). In order to serve its objective, an OOWM has to perform several tasks, which are closely linked to high-level data fusion:

- Information Representation and Distribution: An application-independent representation of objects, its features and uncertainty about their assessment have to be developed. To distribute object information to high level applications, a unified access and query mechanism has to be established.
- Data Association: For each new object observation, it has to be decided if it corresponds to a new object or represents updated information about a previously observed object.
- Data Fusion and Tracking: Updated information from new observations has to be fused with previously assessed information.
- Information Aging and Management: First, it has to be managed, whether objects which have not been observed for a longer time period can be removed from the object representation. Secondly, it has to be decided if enough observations have been collected to support the existence of a new object.

The main benefit of the use of an OOWM is the establishment of a generalized object representation, which is designed to be independent of application and signal-processing level. Detailed explanations about the technical implementation of an OOWM can be found in [17].

III. Identity Management

By developing a modern form of Identity Management, it is possible to integrate more privacy protection in smart surveillance systems. In this chapter, first the challenges for privacy in IdM are shown. Then, it is demonstrated how to solve these challenges by combining IdM with automated privacy policy enforcement.

A. Privacy Challenges

Smart surveillance systems have great potential. By integrating new sensors and task-based surveillance, operators are relieved of surveillance and crime prevention is increased. Unfortunately, new privacy challenges result from these new systems. In this section, three urging privacy challenges are highlighted and used to define the privacy protection goals of this work.

1) services can use Information from Foreign Tasks: In a task-oriented surveillance architecture, different tasks have different access rights to sensors. A critical task, e.g., for protecting access to a secret area, can make use of higher-performance sensors, while a non-critical task, e. g., counting persons using the elevator, can only use less-performance sensors. The services inherit this access right and every service can only access attributes collected for its task.

In a running smart surveillance system, a lot of different surveillance tasks exist and all objects and their attributes get stored in the OOWM. To prevent that the same attribute is stored multiple times, the OOWM merges all attributes from the same person in the real world in the same digital identity. Thus, attributes collected by different sensors get merged into one object. This is a great challenge for privacy, because now it is difficult to prevent services from using information collected for foreign tasks.

2) Privacy Levels: Smart surveillance systems can monitor a lot of different people simultaneously. Normally, these people require different levels of surveillance. For instance, in a mental institution: A mentally ill patient needs to be monitored differently than the hospital nurses or medics.

3) Malicious Services Merge Information: Services get their information from the OOWM. To enable them to differentiate between objects, every object has a unique ID. Malicious services can use these IDs to merge their information, just by merging attributes from objects with identical IDs. Therefore, it is not sufficient to just prevent services from accessing certain attributes, but furthermore to prevent services from using IDs to merge information.

For each of the three challenges above, a privacy goal emerges:

Privacy Goal 1: Services should only be able to use information collected by sensors assigned to their task.

Privacy Goal 2: It must be possible to monitor some people very strictly without violating the privacy of others.

Privacy Goal 3: Services must not be allowed to merge information by using the IDs.

B. Privacy Manager

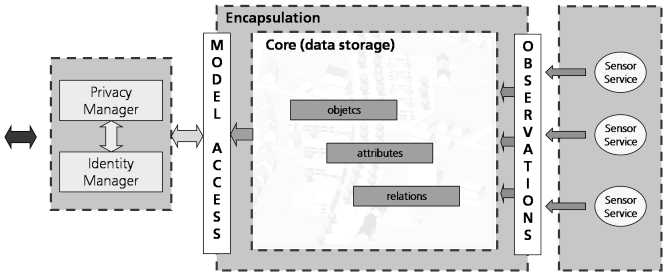


Fig. 3. Privacy Manager as information proxy

A first step to improve privacy protection is to block direct access to the OOWM. In the new approach, shown in Figure 3, services have to acquire their information from a special module, the *Privacy Manager* (PM). By this, it is possible to concentrate most of the privacy protection techniques in one place.

The Privacy Manager acts as an information proxy and is able to check every attribute leaving the OOWM. Further, the OOWM keeps track of which attribute is collected for which task. So the Privacy Manager can prevent services from obtaining information collected for other tasks. Thereby, *Privacy Goal 1* can be achieved by the Privacy Manager blocking access to foreign information.

Another achievement of this approach is the easy enforcement of privacy policies. Special policies, for every task and corresponding services, allow fine-tuned access rights. Not only is it possible to completely prohibit certain attributes for individual services, but it is also possible to edit them. Thus, privacy protection is increased, while still allowing the service to fulfill its task (see Section IV). In addition, group- or person-specific privacy policies can be enforced by the system and so it is possible to have different levels of surveillance for individuals or groups. So *Privacy Goal 2* is achieved by the Privacy Manager through removing or editing attributes in conflict with privacy policies (not discussed in this work, see e. g., [18]).

C. Multi-Layer Privacy Management

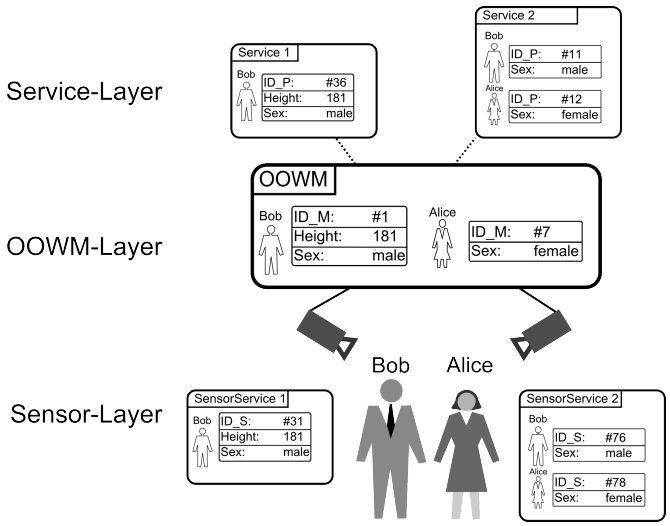


Fig. 4. Multi-Layer Identity Management

Without solving *Privacy Goal 3*, the effect of the Privacy Manager is limited. Malicious services are still capable of exchanging information and thereby evading its work. So a new approach, the multi-layer IdM, is used. This approach generates three different layers of IDs, representing the three different layers of identities available in the system (Figure 4).

On the lowest layer, the Sensor-Layer, ID_S are used. Every time a sensor spots a new object, it generates a new unique ID. Whenever a Sensor sends information about a known object, it reuses the corresponding ID. So the OOWM can differentiate between objects observed by various sensors.

The next layer is the OOWM-Layer. All information concerning the same object in the surveillance area gets merged into a single data object. For these objects, the OOWM generates a new ID from the set ID_M. In this layer, attributes detected by different Sensors with different ID_{S_x} are fused into one object with a new unique ID_{M_x}. Without knowing its ID, no information about an object can be obtained from the OOWM.

The highest layer is the Service-Layer. At this level ID_P are used and all previously described privacy enforcement techniques are applied.

D. Distribution and Usage of IDs at the Service-Layer

IDs can be maped between the sets ID_P and ID_M in three different ways:

- Bijective mapping between ID_P and ID_M, which is constant within a scenario. The assignment is saved in the PM.
- Unique mapping between ID_P and ID_M, but for every communication a new ID $x \in ID_S$ is created. The assignment is saved in the PM.
- For every communication at the Service-Level a new $ID_{P_x} \in ID_P$ is created, the assignment is not saved. Hence, no mapping between ID_{P_x} and corresponding $ID_{M_y} \in ID_M$ is possible.

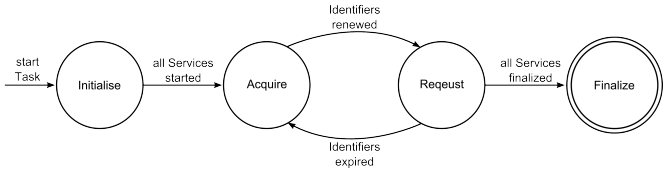


Fig. 5. Four Different States of Surveillance Tasks

Every running task is in one of the four different states, as shown in Figure 5. The first and the last state are important for managing purposes only. The second (Acquire) and the third state (Request) are important for the Privacy Manager and Identity Manager. A more detailed look at the different states is provided below.

In the first state, all needed components are initialized. As soon as all required sensors and services are running, a task changes into the second state.

In the Acquire state, services have to register at the Privacy Manager to obtain information from the OOWM. The authentication is done by established methods and therefore not part of this work. After the authentication, the request is transmitted to the Identity Manager. The Identity Manager checks, which objects from the OOWM contain attributes collected for the requesting service. For every accessible object, the Identity Manager creates a new ID from the set ID_P. Depending on the mapping (see above), IDs in the set ID_M are blocked to ensure that they can only be used in one specific context. When an ID is blocked the assignment is saved.

After that, a list containing all its ID_{P_x} is sent back to the requester. In this state, the Identity Manager assures that all Services use different IDs.

In the Request state, a service can query attributes of a certain object it has access to. This is done by transmitting the ID_{P_x} of the object to the Privacy Manager. The Privacy Manager gives this ID to the Identity Manager which transforms it to the corresponding ID_{M_x} from the OOWM. If an ID_{P_x} can just be used once, a new ID_{P_z} must be requested and the service has to register again. If a critical internal change took place at the IdM, a service as to register again as well. At this point, it is important to stress the fact that without the support from the IdM a ID_{P_x} is useless. The Identity Manager is the only part of the system capable of transforming the ID_P used by the services into ID_M used by the OOWM. Hence, the services have no possibility to circumvent the Privacy Manager and access the OOWM directly.

With the transformed ID_{M_x} the Privacy Manager can access the stored information on the object. It obtains all available information and checks them against its privacy policies. In this step, the Privacy Manager removes all attributes not accessible by the requesting service, either because the attributes were collected by a foreign task or because active policies prohibit their usage. After that, the checked information is delivered to the service.

After finishing its observation, a task changes into the last state. Here all running sensors and services are being terminated. ID_{P_x} are unblocked.

By using these four different states, new privacy enhancing methods can be integrated. Due to the fact that every service has to obtain its information from the Privacy Manager, privacy checks can be enforced at a single point. Finally, it becomes impossible for services to merge their information about objects with the help of IDs. Every service uses a different set of IDs, leaving them no common base to merge information.

IV. Demonstration and Testing Scenario

To evaluate the impact of the IdM on the NEST Architecture, a test scenario is used. In this scenario, the OOWM knows one object, the Person "Max Musterman" with numerous attributes (see Figure 6 (OOWM)). Two services are running to analyze the information. The first one *NEST Emergency Aid* keeps track of all persons in a building. In the case of an emergency, this service can give the last known position of all persons to the rescuers. The second service *NEST Attendance Clock* is the modern form of recording attendance. The service registers when employees are entering and leaving the company buildings so as to monitor their working-time accounts.

A. Scenario Without Privacy Protection

In the first scenario, the OOWM is used without any privacy enforcement. As seen in Figure 6 (a), both services can access all information stored about Max Musterman. Furthermore, both services use the same Identifier "Max Mustermann" to refer to the object. So they could easily share information about the person, when data access is limited. In this scenario, both services have more information than needed to work properly. NEST Emergency Aid only needs the position of every person. Further information, such as name or sex, is not necessary. A similar problem occurs for the second service, as it only needs to know which employees are in the building but not their exact position in it.

Since excess information increases the risk of abuse, this situation is extremely undesirable.

B. Scenario With Privacy Protection

As seen in Figure 6 (b), with privacy control enabled, both services have only restricted access to the information in the OOWM. The NEST Emergency Aid has access to the position attribute. This is everything this service needs to perform its task. The second service, NEST Attendance

Clock, is able to read the name attribute, and thus can map Max Musterman to his personal time account. The attribute position gets abstracted to 'in building'. This is sufficient to perform the task, but much less critical for privacy protection than the real position.

Through the work of the Identity Manager, both services use different IDs to refer to the same observed person, Max Musterman. So the services cannot use their IDs to merge information about him.

With privacy enforcement enabled, both services can perform their tasks, while protecting privacy more than in the scenario without privacy protection.

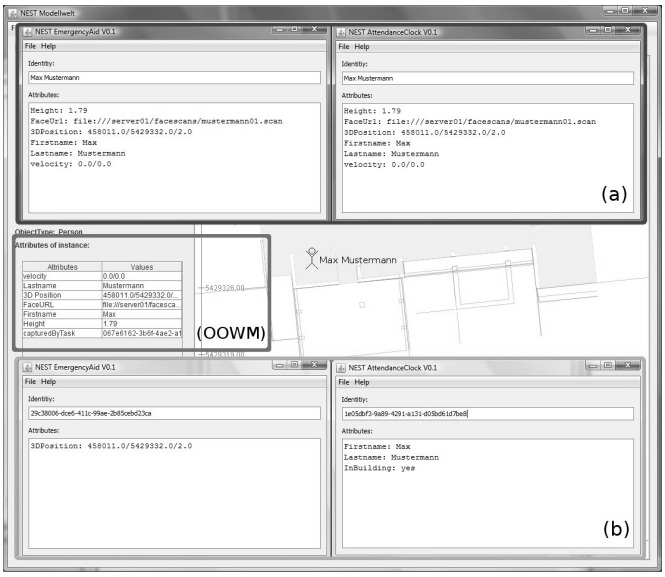


Fig. 6. Scenario Without (a) and With (b) Privacy Protection

V. Conclusion and Outlook

Data protection and privacy enforcement are two of the key challenges in smart surveillance. While systems become more and more powerful, the usage of personal data or even any data must be restricted. The task-based approach of the NEST architecture combined with the work of the Privacy Manager and Identity Manager allows a high grade of protection. As shown in Chapter IV, it is possible to protect the privacy of the surveillance subjects without losing functionality in the system.

Privacy in smart surveillance systems is a recent area of research although a lot of research has already been done. Yet further work, for instance in the area of extended privacy policies needs to be done. With new approaches, it could be possible to create policies which adapt to the current situation and so provide even more privacy protection. Another interesting topic is the integration of mobile devices in the available systems to modify personal privacy policies. This would allow to develop highly dynamic systems.

References

[1] K. Cameron, "The laws of identity," http://www.identityblog.com/?page_id=354, May 2005.

[2] C. Mezler-Andelberg, *Identity management - eine Einführung*, 1st ed. Heidelberg: dpunkt-Verl., 2008.

[3] K. Rannenberg, D. Royer, and A. Deuker, *The Future of Identity in the Information Society: Challenges and Opportunities*. Springer Publishing Company, Incorporated, 2009.

[4] H. Vagts and J. Beyerer, "Security and privacy challenges in modern surveillance systems," in *Future security: 4th Security Research Conference*, 2009.

[5] OECD, Ed., *OECD guidelines on the protection of privacy and transborder flows of personal data*, 268th ed. OECD Publishing, March 2003.

[6] A. Senior, S. Pankanti, A. Hampapur, L. Brown, Y.-L. Tian, A. Ekin, J. Connell, C. F. Shu, and M. Lu, "Enabling video privacy through computer vision," *Security & Privacy, IEEE*, vol. 3, no. 3, pp. 50–57, May-June 2005.

[7] A. Chattopadhyay and T. E. Boult, "Privacycam: a privacy preserving camera using uclinux on the blackfin dsp," in *CVPR*. IEEE Computer Society, 2007.

[8] T. Boult, "Pico: Privacy through invertible cryptographic obscuration," Nov. 2005, pp. 27–38.

[9] F. Dufaux and T. Ebrahimi, "Scrambling for video surveillance with privacy," *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pp. 160–160, June 2006.

[10] J. Schiff, M. Meingast, D. Mulligan, S. Sastry, and K. Goldberg, "Respectful cameras: detecting visual markers in real-time to address privacy concerns," *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pp. 971–978, 29 2007-Nov. 2 2007.

[11] Associated Press, "Talking camera tackles city crime," http://www.cbsnews.com/stories/2005/11/17/tech/main1054526.shtml (last access 10.08.10), November 2005.

[12] S. Fleck and W. Strasser, "Smart camera based monitoring system and its application to assisted living," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1698–1714, Oct. 2008.

[13] D. A. Fidaleo, H.-A. Nguyen, and M. Trivedi, "The networked sensor tapestry (nest): a privacy enhanced software architecture for interactive analysis of data in video-sensor networks," in *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*. New York, NY, USA: ACM, 2004, pp. 46–53.

[14] H. Vagts, A. Bauer, T. Emter, and J. Beyerer, "Privacy enforcement in surveillance systems," in *Future security: 4th Security Research Conference*, 2009.

[15] U. Jendricke, M. Kreutzer, and A. Zugenmaier, "Mobile identity management," Institut für Informatik, Universität Freiburg, Tech. Rep. 178, Oct. 2002, workshop on Security in Ubiquitous Computing, UBICOMP 2002.

[16] J. Moßgraber, F. Reinert, and H. Vagts, "An architecture for a task-oriented surveillance system – a service and event based approach," in *Proc. Fifth International Conference on Systems ICONS*, 11–16 April 2010.

[17] A. Bauer, T. Emter, H. Vagts, and J. Beyerer, "Object oriented world model for surveillance systems," in *Future Security: 4th Security Research Conference*, P. Elsner, Ed. Fraunhofer Verlag, Oct. 2009, pp. 339–345.

[18] Q. Ni, E. Bertino, J. Lobo, and S. B. Calo, "Privacy-aware role-based access control," *IEEE Security & Privacy*, vol. 7, no. 4, pp. 35–43, 2009.

RFID Applications in Hospitals – A Case Study for Emergency Department

Yen-Chieh Huang[1,2], Chih-Ping Chu[1], Yu-Shih Lin[1], Chia-Hao Kuo[1]
[1]*Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan*
[2]*Department of Information Management, Meiho University, Pingtung, Taiwan*
E-mail :p7894121@mail.ncku.edu.tw

Abstract

Radio Frequency Identification (RFID) is a system that uses radio frequency to transmit identification wirelessly. Its functionality is much more powerful than bar code system, with which a contacting reader must be used to read the bar code. RFID, on the other hand, may read the information stored in a tag with a non-contact reader from a distance as far as tens of meters. Health care industry is highly valued throughout the world; applying the cutting edge technology in health care industry to improve patient health care has been the common goal pursued by the hospitals all over the world. The introduction of RFID as medical application is also being enthusiastically studied now. This Study finds the medical application of RFID application for the workflow of the most critical and the busiest unit in a hospital: the emergency department. The results of this Study may significantly increase the operation efficiency by improving the problems commonly experienced at the emergency room, such as congested queue for emergency treatment and sickbed space, understaffed medical personnel, and patient's leave without permission.

Keywords: RFID, Emergency Room, Health Care, Patient Satisfaction

I. INTRODUCTION

The Wal-Mart in the United States demanded the packaging of all merchandises delivered to its distribution center must have re-readable and re-writable passive RFID tags in compliance with the EPC (Electronic Product Code) standard attached before January 2005 from its top 100 suppliers and 2006 from all suppliers. The Food and Drug Administration (FDA) of the United States announced in 2004 to adopt RFID technology for the identification of counterfeit medicines and demanded all pharmaceutical manufacturers must have RFID tags attached to the packaging of their medicines starting from 2007 for the tracking, managing, and recycling of medicines. As a result, all wholesalers and retailers must follow and adopt RFID system [1]. While many researches are undergoing for other industries, the introduction of RFID for health care industry is still in the experimental stage. Plans of RFID application tests in hospitals have been proposed by many suppliers, such as medicine control, patients contact history, patient identification, equipment/apparatus tracking, injection management, physician order monitoring, medical malpractice prevention, blood bag quality control, and operation room workflow; all of which concern primarily the safety of the patients. Medical application of RFID for patient health care enhancement is therefore the focal point of this Study.

The emergency room is the most complicate and busiest place in a hospital. It is the center for treating patients with accident injuries and/or acute sicknesses of different levels of medical needs on a daily basis that the patients must receive the most adequate treatment and care in the shortest possible time from both the physicians and the nurses. Upon the arrival at an emergency room a patient goes through a medical operation procedure, including triage, registration, treatment, cashier, pharmacy, and admission, which require seamless collaboration with the medical treatment team [2].

A patient goes missing and leaves registration record but no clinical record is a frequent scenario in an emergency room. Such patient often leaves without excuse the hospital while waiting in queue, inhibiting hospital management from tracking the patient head count and location and making patient scheduling a difficult task for the hospital personnel [3]. It is therefore a crucial issue for the medical workers being allowed to control the accurate information of the patients' location in real time.

Some of the patients after diagnosis by physicians are concluded to be non-urgent cases and are redirected to the provisional observation room instead of hospitalization. However, chances are some of them have to stay and wait for hours before receiving any further attention from the medical personnel again and being left alone to deal with the anxiety brought by the emergency room surrounding on top of the stress of illness. Therefore, it is paramount to allow the patients to know the examination results and the diagnosis conclusion as soon as possible. The reduction of patients' waiting time is urgently needed [4, 5].

As a result, enhancing medical treatment quality, increasing patients' satisfaction toward hospital, making more time available for emergency treatment,

and providing even better medical care environment in the emergency room through the use of information technology are now the common goals in the health care industry. The scope of this Study focuses on improving medical treatment safety through RFID applications in medical treatment process for the patients in the emergency room, including alert of excessively prolonged queue, patient location tracking, and alert of excessively prolonged stay.

II. LITERATURE REVIEW

A RFID system, in the simplest form, consists of three components: a Tag or Transponder, a Reader and an antenna. A passive tag (or powerless tag) has no battery built in and requires external energy to trigger and initiate signal transmission. An active tag, another common tag, has built-in battery to provide power to actively transmit signal. Both types of tag have a chip inside to carry and process information. Upon entering an electromagnetic zone a passive tag is triggered by the interrogating signal sent by the reader, from which it obtains power through induced current, and begins transmitting the product information stored on its chip. An active tag, on the other hand, as its name suggests, actively and constantly transmits the information on its own. As the reader receives the information, it relays the information to a central information processing center for further application. Other hardware and software are required to support RFID application [6, 7]. A tag may come in different shapes and sizes: business card, coin, button, price tag, wristband, and grain; the list can go on and on. In general, a tag may be passive, semi-passive, or active [8]. A passive tag has no battery built in to supply power and requires external induced current to power and trigger signal transmission. This is why its working range is limited to one meter or shorter. Both semi-passive and active tags have built-in battery, and that is the reason why their working distance ranges from several meters to tens of meters, only that a semi-active tag still needs external induced current from the reader to activate the built-in battery to supply power for signal transmission. An active tag will constantly transmit signal; if it detects the presence of a reader, it begins to transmit information. Because a passive tag is powered externally by a reader, its service length is virtually unlimited (until it is damaged). The service length of an active tag may last up to 8 years theoretically or merely one month in practice as it is limited to the battery life and power consumption.

The Ministry of Economic Affairs (MOEA) has taken initiatives to seek for preventive countermeasures since the SARS epidemic, and RFID application in medical care is one of them. Currently the pilot projects have been rolled out at the Taipei Medical University Hospital, Show Chwan Hospital, Sun Yat-sen Cancer Center, and Kaohsiung Veterans General Hospital. With financial support from the MOEA, the Industrial Technology Research Institute is also cooperating with Hsinchu Dong Yuan Hospital for RFID trial project. Each of these medical institutes concentrates on different subjects of RFID applications: SARS control, area control, restricted area control, bio-medical waste control and monitor, and patient contact tracking.

Some hospitals in the United States are keen to adopt new technology for medical care application [9]. For examples, RFID time sheet for ER personnel [10] and medical record tracking by RFID [11]. These pilot projects have achieved quite satisfying results and these hospitals use 2.45GHz active tags, for which readers have a scope of signal interrogation as long as 30 meters, far more accurate than passive tags that allow less than one meter of scope.

Several human right groups in the United States oppose RFID applications on patients for the reason of privacy violation by unauthorized RFID tracking [12]. Some manufacturers in return respond with erasable tags [13].

RFID applications in Taiwan are still very limited. The policies of the Bureau of National Health Insurance have caused great reduction in hospital profitability and, consequently, displaced the available budget for the induction and application of new technology. Another factor is that it is inherently difficult to change the status quo of the operational workflow already in place and accustomed to hospital employee. The most critical issue, eventually, is the price of RFID system is not yet attainably acceptable to the majority of consumers. There are, practically, difficulties in applying RFID for any medical institute. The pilot projects of RFID, subsidized by the MOEA, are currently implemented at the Taipei Medical University Hospital, Show Chwan Hospital, Sun Yat-sen Cancer Center, Kaohsiung Veterans General Hospital, and Hsinchu Dong Yuan Hospital. Nonetheless, these projects are all not comprehensive and have still a long way to go before the full scale implementation [14,15,16,17].

Shiumn-Jen Liaw et al. (2002) had studied the factors of patient premature departure from emergency room with incomplete diagnosis procedure and categorized such patients into four types: (1) departure without diagnosis, (2) departure due to denied insurance of emergency coverage, (3) departure without excuse, and (4) voluntary departure. The departure without excuse is further sub-categorized as incompletely diagnosed and completely diagnosed. Liao concludes that these patients departed mostly because the excessively long waiting time for diagnosis, time poverty, or unpleasant encounter with hospital personnel. Liao suggests 42% of the patients expect the waiting time not to exceed 15 minutes, and 49% do not accept more than one hour of waiting time. Failure of either one shall result in dissatisfaction with the medical

service provided by the hospital [3].

Su-chou Siao (2002) proposes that the major factors of prolonged patient stay for provisional observation in emergency room are (1) excessively long waiting time for available sickbed, (2) excessively long provisional observation time, (3) understaffed medical personnel, and (4) insufficient space [5]. Ming-ling Liou (2001) has conducted field investigations in emergency room to explore the distribution and variance of waiting times among the stages of medical care in emergency room, and studied their significances concerning patient satisfaction.

The literatures abovementioned reveal that most of the patients depart from hospital due to excessively long waiting time. The long the waiting time at emergency room, the more likeliness of patient dissatisfaction toward hospital will be and eventually cause patients to depart or even permanently abandon the hospital. There is a significant causal relationship between patient waiting time and departure.

III. SYSTEM ENVIRONMENT STRUCTURE AND DESIGN

The goal of this Study is to design a query system using RFID, of which major function shall be actively providing real-time information of the locations of the patients with excessively long waiting time.

A RFID system structure is a combination of tag, reader, and application system. The type of application system varies according to the needs of an industry and the structure may be specialized according to the specification of the tag used. The system structures in the case of passive or semi-passive tag and in the case of active tag, for example, are different. In this Study focuses on applying semi-active tag of UHF frequency on the patients at emergency room to propose a system structure for safe emergency room service. Upon the arrival of a patient at an emergency room for triage, a RFID wristband is printed and mounted to the wrist of the patient. All following procedures must be reconfirmed with a reader, including diagnosis by physician, medication distribution by nurse, injection, examination, urgent operation, and etc.

As shown in Figure 1, the process at emergency department in most hospitals begins with triage at the entrance. A patient of level one or two triage must complete registration first and then queue in the waiting room. A level three or four triage may proceed to emergency care by physicians first and complete the registration process afterward. The emergency departments in some larger hospitals may be specialized into different departments, such as emergency surgery, emergency internal medicine, and emergency pediatrics. Whatever the department it may be, the physicians after the initial diagnosis may request certain examinations, and prescribe medications accordingly. The relationship between physician diagnosis and radiology or laboratory department is expressed with two-way arrows in Figure 1, denoting that such process is not necessarily required depending on the physician's decision. If so required, the process following the examination is still the physician for further diagnosis.

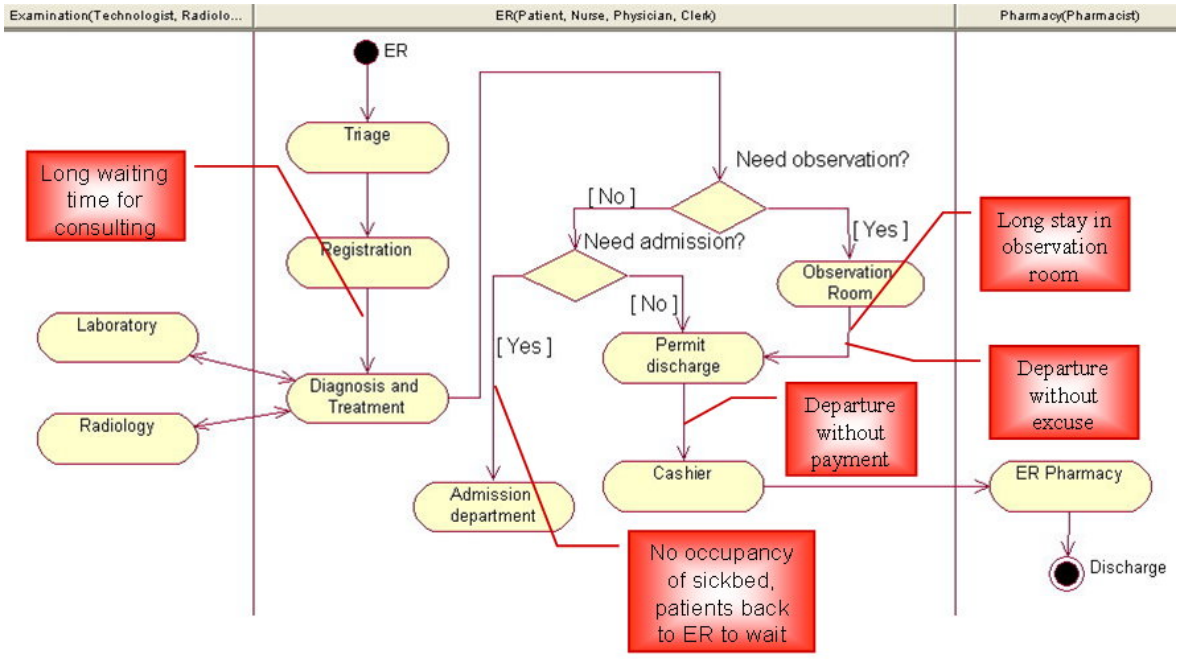


Figure 1. ER operation process and possible procedural problems

Five areas of the process of emergency department are highlighted in Figure 1 as possible procedural problems. The first one is patients' excessively long waiting time for diagnosis by physician after triage and registration. In some cases large amount of patients are rushed to the emergency room, and the patients with critical conditions are prioritized by the physicians, causing longer waiting time for other patients. In some other cases the physician on duty at the emergency department is not specialized for the patient's condition and must wait for the qualified specialist, such as an eye-related problem shall wait for an ophthalmologist, and an ear, throat, or nose (collectively referred as ENT) issue should be determined by an otolaryngologist. If a specialist is not available or not present, the patient will have to wait for even long time. Therefore, the first problem is excessively long waiting time for the patients. The second factor is patient after being determined as hospitalization is not necessary by physician receives no attention during provisional observation due to medical personnel is preoccupied by other tasks, causing the patients to endure excessively long waiting time. Therefore the second problem is excessively long stay in the provisional observation area for the patients, resulting in patients' departure without excuse compelled by their dissatisfaction toward the hospital personnel. This leads to the third problem: departure without excuse. A patient may depart from the hospital for certain reasons before the process of cashier after being acknowledged that leave is authorized. For examples, lack of money to pay the hospital or no medication is prescribed. This is the fourth problem: departure without payment. The fifth issue is no occupancy of sickbed after a patient is instructed after diagnosis to proceed with hospitalization process, causing the patient to wait in the emergency room for sickbed. The availability of sickbed in some popular hospital often runs out, and the waiting time for a sickbed may last for as long as two or three days, causing patients' dissatisfaction. Therefore, the fifth problem is no occupancy that drives patients back to emergency department to wait.

Because an emergency process is prone to negligence, the application of RFID in emergency department, as shown in Figure 2, by immediate printing and mounting a RFID wristband on the wrist of a patient at triage by the medical personnel before the patient is redirected to the waiting area. If a physician fails to diagnose the patient in due time stipulated by the hospital, a short message is immediately sent to the management center where high management may initiate an emergency protocol. For example, dispatching extra physician may avoid excessively long waiting time for the patient. This measure solves the first problem mentioned in above. A patient in the provisional observation area may be waiting for examination report or intravenous drip. However, if such patient is made to wait for excessively long time without further attention form the medical personnel, the RFID will not be read for a predetermined amount of time. The information of the patient will be automatically sent to the management center where the management of the hospital may respond accordingly, such as instructing the medical personnel to give further attention provide more medical service, or request the examination reports to be completed within a certain amount of time. This measure may solve the second problem mentioned in above. In the case that a patient is not yet authorized to leave, or that a patient is notified that leave is authorized before the cashier process, and the patient departs without excuse, the RFID reader at the exit of the hospital will automatically send an instant message to remind the medical personnel that the patient has departed. The intention of this message is helping the medical personnel to realize the absence of the patient, not necessarily policing on the patient's departure without excuse. There are times the patient is simply going out for lunch or having a cigarette outside. This measure takes care of the third and the fourth problems. Before a patient in need of a sickbed is assigned with one, the RFID will not receive the information of the completion of hospitalization process of the patient at admission counter. If the RFID after a predetermined period of time receives no such information, then it means the patient is suffering from excessively long waiting time and the RFID system will automatically send a short message to alert the management center where decision can be made to direct the medical personnel to assist the patient to proceed with transfer, and patient dissatisfaction of waiting time may be avoided. This measure eliminates the fifth problem.

IV. RESULT

As consumerism continues to rise in recent years, more scholars are studying the quality of the medical care provided by medical service providers form the aspect of consumer. The term of medical quality in the eyes of the patients has escalated from the simple treatment result to customer satisfaction of the entire medical service experience. This phenomenon has confirmed that the medical service concept based on patient's right shall be the direction of medical quality development in the future. RFID can be adopted by a hospital for a variety of applications. The Emergency Department of the Kaohsiung Veterans General Hospital has adopted RFID for reconfirmation of medication distribution and injection. Dong Yuan Hospital has adopted RFID to manage SARS patient contact history. The Operation Room of Chang Gung Memorial Hospital has adopted RFID for reconfirmation of the surgical target.

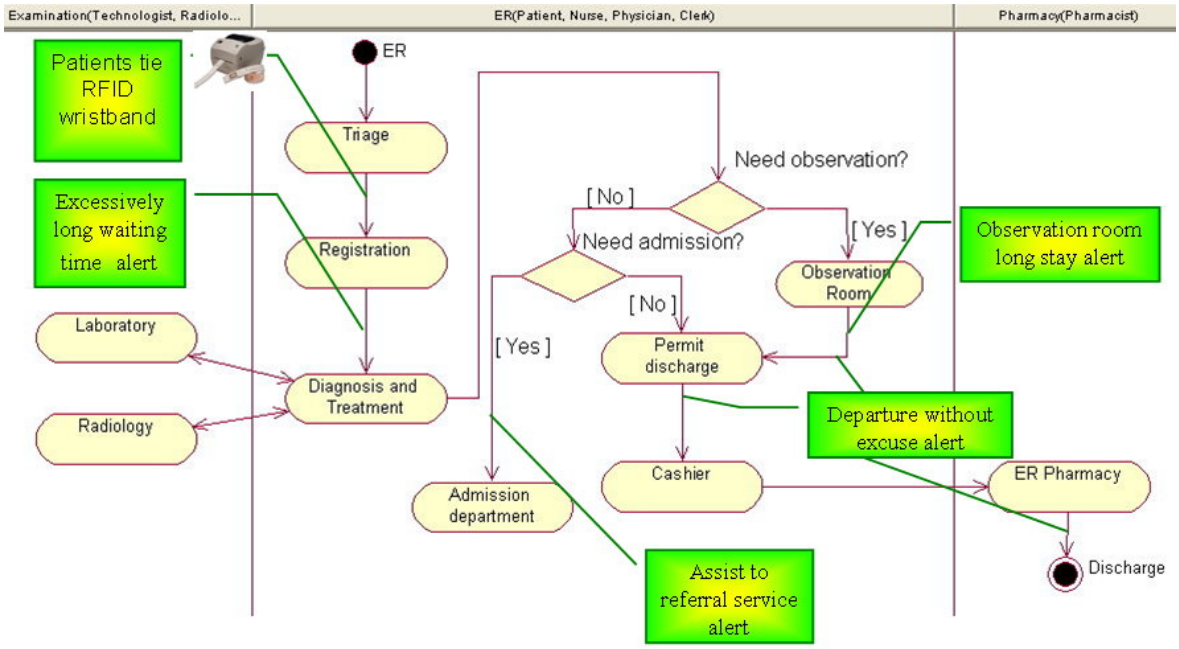


Figure 2. RFID apply to ER process

This Study concerns the experimental introduction of RFID at the emergency department. It is expected that the introduction of RFID not only assures the five-step medical safety checklist (the right patient, the right medication, the right dosage, the right approach, and the right timing), but also solves the problems of excessively long waiting time of diagnosis, excessively long provisional observation period, unattended by medical personnel, excessively long waiting time of sickbed occupancy, and departure without excuse.

This Study has developed, by using Visual Basic, a RFID System for Emergency Department to solve the above-mentioned five problems of the emergency room process. However, the system has only been simulated in laboratory. Despite that the result of this Study is not implemented in any hospital in Taiwan, it indeed can enhance the service and safety for the patients, increase medical quality, and develop trustworthy image for a hospital. In addition, the doubt of privacy violation is also addressed by this Study. Tracking a patient's location in a hospital or preventing a patient from departure without settling medical bill is not the intention of the RFID system proposed by this Study, even though it may be so used. This Study aims to enhance medical quality by assisting medical personnel to seek for the location of an emergency patient and provide adequate assistance for the patients who have waited excessively long time.

V. DISCUSSION AND FUTURE DEVELOPMENT

This Study focuses on the introduction of RFID for emergency department application that not only solves the problem of patients' excessively long waiting time, but also allows the hospital management to monitor the service for the patients provided by the medical personnel at any given time and reduce patient complaint, even prevent potential medical dispute. For the medical personnel, they may stay informed about the location of the patients at any given time, monitor the examination and medication for the patients, reduce data input time in the workflow, and serve more patients with the time saved. For the service provided for the inpatients, the RFID reader installed at the hospital entrance/exit monitors the location of the inpatients to protect them from criminal actions and prevent missing inpatient. For the academic circle, this is an excellent case study for industry-specific application. RFID not only may be applied in other departments of a hospital, but also in other industries, such as manufacturing, service, apparel, and food and beverage. RFID application is important more than ever for the enterprises that emphasize customer relationship management, cost reduction, and profitability.

The triage station personnel, registration/cashier clerks, physicians, nurses, medical technicians/radiologists, pharmacists, patients, and hospital management involved in the introductory stage of RFID application in the emergency department workflow may experience added workload, such as time pressure. However, they shall be enjoying the convenience brought by the new technology after the comprehensive implementation. Every step of the process may comply with the standard operational procedure, and medical safety may be greatly enhanced. If the implementation succeeds, it can serve as a fundamental paradigm for expanding the applications

to other applications in the hospital and encourage the acceptance of RFID among the personnel concerned. The Information Management department is not only a logistic unit of a hospital; it should make the best use of information technology and apply in every process of a hospital to increase operational efficiency and competitiveness, assuring the sustainability of the hospital.

ACKNOWLEDGMENT

The author would like to express appreciation to Dr. Zong-long Yang, Radiology Department of the Kaohsiung Veterans General Hospital, for accepting the study team's interview and offering recommendations for this Study.

REFERENCES

[1] Donna, Y., (2004), "FDA embraces RFID to protect drug supply", Am J Health-System Pharmacy, Vol. 61, No. 24, pp.2612-2615.

[2] Rui-Yu Rao, (2002), "Use PDCA management style to improve emergency patients' healthcare", Annual Conference of Emergency Medicine, pp.133.

[3] Shiumn-Jen Liaw, Pai-min Hu, and How-chin Liao, (2002), "Patients Who Leave Emergency Departments Prematurely", Journal of Taiwan Emergency Medicine, Vol. 4, No. 2, pp.40-49.

[4] Li-Hsiang Wang, (2001), "The difference of Patient Needs Among Patient, Caregiver and Nurse in The Observation Room of Emergency Department", Master Thesis, Chang Gung University, Taiwan.

[5] Su-chou Siao, (2002), "Case of emergency observation to improve patient retention", Annual Conference of Emergency Medicine, pp.136.

[6] Wei Zhang, (2003), "RFID Application Discovery", Journal of the Mechatronic Industry，Vol. 249, pp.88-94.

[7] Yu-Hong Chen, (2004), "RFID System Introduction", Unalis, Taipei Taiwan.

[8] Xiang-Qi Zhou, (2004), "RFID Technology and Application", Flag, Taipei Taiwan.

[9] Glenn, M.H., Robert, D.R., and Mark, E.M., (2004), "Report to the Congress: New Approaches in Medicare," New Jersey, pp.157-181.

[10] Intermec, (2005), Saving Time and Improving Patient Care, http://www.imtermec.com/.

[11] Jonathan, C., (2005), Tracking Medical Emergencies, RFID Journal, http://www.rfidjournal.com/.

[12] David, M., and David, W., (2004), "Privacy: Privacy and security in library RFID: issues, practices, and architectures", Proceedings of the 11th ACM conference on Computer and communications security.

[13] Ari, J., Ronald, L.R., and Michael, S., (2003), "The Blocker Tag: Selective Blocking of RFID Tags for Consumer Privacy", Proceedings of the 10th ACM conference on Computer and communications security, pp.103-111.

[14] Zhi-Qiang Wei, (2004), "Survey Research of RFID application to discuss the RFID promotion in our country." Ministry of Economic Affairs E-navigation, Vol.6, No. 20, pp.1-11.

[15] Rong-Xing Xiao, Wei-Ru Wang and Wei-Ren Su, (2004), "RFID Application Case Study", Ministry of Economic Affairs E-navigation, Vol.6, No. 15, pp.1-16.

[16] Rong-Xing Xiao, Wei-Ren Su, and Yu-Jai Xu, (2004),"The nerve center of RFID technological operation- RFID Middleware", Ministry of Economic Affairs E-navigation, Vol.6, No. 14, pp.1-13.

[17] Rong-Xing Xiao, and Yu-Jai Xu, (2004), "The application of FRID and development trend", Ministry of Economic Affairs E-navigation, Vol.6, No. 13, pp.1-15.

A Technical Model Using Genetic Algorithm For the Cosmetic Tourism Business

[*]Chi-Lu Yang, [*,+]Yen-Chieh Huang, [*]Yeim-Kuan Chang, [*]Chih-Ping Chu
[*]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
[+]Department of Information Management,
Meiho Institute of Technology, Pingtung, Taiwan
{p7896114, p7894121, ykchang, chucp}@mail.ncku.edu.tw

Abstract

The medical skills in some developing countries are almost as advanced as those in developed countries. Medical cost, however, is relatively lower in the developing countries. In recent years, the medical industry in some developing countries has exhibited the trend of providing international medical services. Taiwan's medical quality and instruments are at par with those of most developed countries. The average medical cost is relatively more reasonable. Generally, the infrastructure in Taiwan is at a respectable level to be able to develop international medical treatment and to offer short-term vacations. Furthermore, an oversea tourist will require specific post-cosmetic care after he/she had finished the cosmetic treatments and preferred to have a short-term vacation in that area. Generally, the post-cosmetic care during the vacation is important to effectiveness of the cosmetic treatments. In this paper, we aim to explore technical models for serving the overseas tourist and realizing the collaborative business services of cosmetic treatments and travel tourisms. The proposed Cosmetic Tourist Guiding Model (CTGM), which includes the Genetic Algorithm (GA) scheduling function for planning suitable trips and remote pre-assessable functions, can provide overseas visitors convenient services and allows the cooperation of business owners of medical treatment facilities and travel agencies.

Keywords: genetic algorithm, business model, tour scheduling, cosmetic service, tour guiding

1. Introduction

In recent years, the medical industry has exhibited the trend of providing international medical services in some developing countries such as Korea, Thailand, Singapore, India, and Hungary, among others. These countries earn income by exploring the overseas market and cooperating with travel agencies at the same time. The medical skills of some developing countries are almost at par with those of developed countries. Medical cost is relatively lower in the developing countries. For example, cosmetic surgery and health examinations are relatively economical in the developing countries. Moreover, some transplant operations, such as surgical knee joint, are also cheaper. On the market demand side, people in advanced countries have more expectations in accepting overseas medical treatment when the economic development is gradually becoming outstanding. The medical quality and instruments in Taiwan are almost at par with those of most developed countries. The doctors are experts in their fields. The average cost of medical treatment is also relatively reasonable. Generally, the infrastructure in Taiwan is a convenient reason for the development of international medical treatments and for fostering cooperation with short-term vacation tours.

This study aims to explore technical models that can be used to realize combinative businesses, such as cosmetic treatments and travel tourism. The proposed model, which includes Genetic Algorithm (GA), will be able to provide overseas visitors convenient services and to create opportunities for cooperation among the business owners of medical treatments and travel agencies. The relevant technologies to satisfy these targets are intelligent scheduling, lightweight mobility, cosmetic care, and touring knowledge. We focus on the technical model and scheduling technology. Besides, cosmetic care and touring knowledge are obtained by interviewing relevant domain experts, such as cosmetic doctors and travel agencies in Taiwan.

The rest of the paper is organized as follows. In Section 2, the relevant technologies about scheduling are introduced. The definition of the considerable arguments and representations to construct the proposed Genetic Algorithm (GA) are defined in Section 3. It also explains the proposed fitness function and the GA flow. In Section 4, the technical components and their interconnections are modeled in detail. In Section 5, the experimental results for validating the model are analyzed and discussed. Finally, the conclusion and future work are summarized in Section 6.

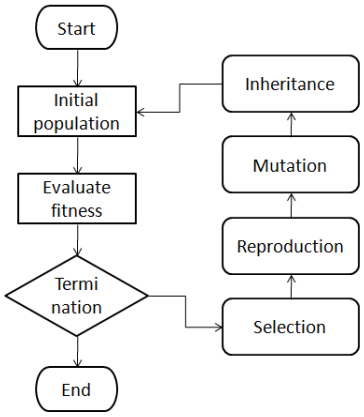


Fig. 1 The general GA process flow

Table 1 Algorithm comparison

	Performance	Stability for precision	Adaption for Problem scale
GA	Fast	Stable	Large
ACO	Normal	High stable	Middle
Tabu	Slow	Normal	Small
SA	Fast	Normal	Large

2. Related Technologies

In this section, the related algorithms that are probably applicable to scheduling tours are carefully introduced. The strengths and weaknesses are compared for the final choice. Those popular algorithms for arranging tours are as follows:

Genetic Algorithm (GA) is a computing technique used in probing the optimization solutions of problem domains. The GA technique is inspired by evolutionary biology fields, such as inheritance, mutation, selection, and crossover. The GA methods are categorized as global search heuristics that belong to one particular class of evolutionary algorithms. When GA technologies are realized in a computer simulation, the solution evolves towards optimization along with a set of abstract chromosomes (i.e., key factors of the problem) in the generations. The evolution usually starts from a random generated solution and occurs within generations. In each generation, the fitness of every solution is evaluated, and multiple solutions are reproduced, which will be selected from the current solutions based on their fitness and modified through recombination or random mutation to form a new solution. The new solution is then used in the next iteration. The next solutions are propagated towards elimination in each generation [1-6]. The behavior is similar to the propagating plant with slow intelligence system [7]. The difference is that the GA is generally terminated when either a maximum number of generations have been reproduced or a satisfactory fitness level has been achieved for the inherited solution. The general GA flow is shown in Figure 1.

The **ant colony optimization algorithm (ACO)** is a probabilistic technique for solving computational problems that can be transformed to geometric graphs used for finding the paths. The ACO was designed to progress similar to the behavior of ants seeking a path between their colony and a food source [8]. The first ACO was proposed by Marco Dorigo in a Ph.D. thesis in 1992 [9-10]. The idea involved some meta-heuristic optimizations and aimed to search for an optimal path in a graph. The original idea has been diversified to solve a wider class of numerical problems. As a result, several problems have emerged, drawing on the various aspects of the behavior of ants.

Tabu search algorithm is a mathematical optimization method that can be used for solving combinatorial optimization problems. Tabu search belongs to the class of local search techniques, such as the traveling salesman problem. Tabu search improves the performance of a local search method using memory structures: once a potential solution has been determined, it is marked "taboo" to prevent the algorithm from visiting that possibility repeatedly. Tabu search was initially proposed by to Fred W. Glover in 1989 [11–12]. It utilizes a neighborhood or local search process to move iteratively from a solution x to a solution x' in the neighborhood of x until some stopping criterion is satisfied [13]. It is also used to explore regions of the search space that will be left unexplored by the local search process.

Simulated annealing (SA) is a generic meta-heuristic method for the global optimization problem of applied mathematics. SA seeks a locating approximation to the global minimum through a given function in a large-scale search space. Besides, it is often used when the search space is discrete (e.g., all tours visiting a given set of cities or spots). For certain problems, SA may be more effective than exhaustive enumeration, provided that the goal is merely to find an acceptably good solution in a fixed amount of time rather than the best possible solution [14-15].

For large-scale problems, GA and SA have better performance than ACO and Tabu. However, ACO is the most stable. GA's stability is also better than that of Tabu and SA [16]. The characteristics of these algorithms are compared in the Table 1. The existing tour scheduling models, like the multi-agents for city travel [19], are not applicable in this cooperative business. For the tour scheduler in our system, we surveyed the related technologies and compared their conveniences. The scheduler in our system requires characteristically a large amount of data to arrange

tours for visitors. Therefore, the proper algorithms should either be GA or SA. Furthermore, the GA solution has better precision than SA. Thus, we finally chose the GA algorithm for the scheduling tours, and implemented it in the kernel of the proposed system.

3. Proposed Algorithm Design

This section describes how the GA is applied to tour scheduling. In the GA data structure and evolution, the definitions of chromosome and genes play the most important roles. Therefore, the chromosome representation and the genes are first illustrated. The effecting factors to the genes are then exposed in detail. The solution evolution is iteratively improved by generating next chromosome. The fitness function is the criterion used to stop the evolution. Thus, fitness function is also defined in this work. Finally, the GA flow is expanded to explain evolution progress step by step.

3.1 Chromosome representations

A chromosome is composed of a number of genes. A gene is a minimal unit in heredity. In our chromosome representations, the genes in the left chromosome represent the sequence of the attractions, which the visitors intend to visit. The final evolution sequence represents the most proper tour schedule for visitors. The genes in the right chromosome represent how long the visitors will stay at the attractions in relation to the left sequence of the chromosome. The gene in the middle chromosome represents the time when the visitors start their daily tour. To work out the arrival and departure time of each attraction, the function *f(p1,p2)* is utilized to calculate the delay time of the whole tour. The delay time can be approximately calculated from Google maps. A chromosome example in our algorithm is shown in Figure 2.

In the example figure, the visitors start their daily tour at 8 o'clock in the morning, as the gene in the middle is equal to eight. Thus, the *Ts* value is equal to 8 o'clock. The function *f (S, B) = 1* represents the traffic time from *S* to *B* is one hour. According to function *f*, the algorithm can derive the arrival time at scenic spot B as 9 o'clock. Furthermore, the stay time at spot B can be defined as three hours in the gene next to the middle gene by a travel agency in advance. Thus, the difference in time between the left and right gene *B* is three hours. In the same way, the time 12:30 is derived based on *f (B, A) = 0.5*. In this example, the sequence plan of this tour can be transformed to a user-friendly interface and be used to guide the visitors from scenic spot B to A, D, and finally to C. The visitors will arrive back to the hotel at 11 o'clock in the evening.

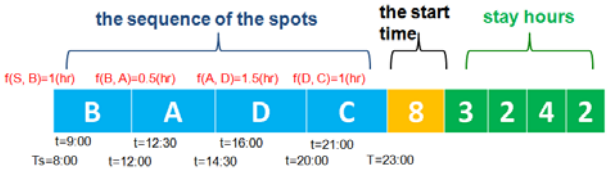


Fig. 2 Chromosome representations

3.2 Fitness function definition

The fitness function is designed to derive the fitness value when a user visits a scenic spot. The factors in the fitness function are as follows: attraction, category, and stay time at a scenic spot. The visitor's hobbies and medical treatments are also included in the factors of the function. The specific definitions are represented as follows:

$\boldsymbol{ATTRACTION}_p(t)$ = *attraction of the spot P at time t*

$\boldsymbol{USER}_{attraction}(P)$ = *attraction of the spot P for the user*

$\boldsymbol{STAY}p(T)$ = *fitness value of staying t hours at spot P*

$\boldsymbol{Environment}_p(t)$ = *environmental factor of the spot P at time t, e.g.,* $\boldsymbol{Light}_p(t)$ *represents the sunshine factor*

$\boldsymbol{USER}_{environment}(x)$ = *fitness of a visitor when he/she stays at environment x, e.g.,* $\boldsymbol{USER}_{light}(x)$

The environmental factors can contain sunshine, temperature, humidity, ultraviolet rays, and wind force, among others. However, based on the interviews with cosmetic doctors, we chose sunshine as our key environmental factor to implement our fitness function. The fitness value in our GA function is defined as follows:

$$f = \sum_{all\ P} \{ \mathbf{USER}_{attraction}(P) \times [\mathbf{STAY}_p(T) + \sum_{t=arrival}^{leave} [\mathbf{USER}_{light}(\mathbf{LIGHT}_p(t)) + \mathbf{ATTRACTION}_p(t)] \}\}$$

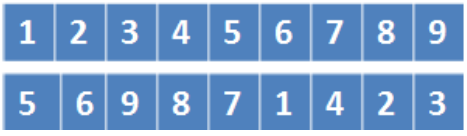
The fitness value of a scenic spot will be diverse at different times because sunshine varies accordingly. This is true in a real situation in the real world. The estimated fitness of a spot at time *t* is used to leverage the fitness value discretely during arrival and departures times. Furthermore, the final fitness of a spot at time *t* is the product of the estimated fitness and a visitor's hobby. Finally, the total fitness value of this tour for the visitors is the sum of the fitness of all the spots.

3.3 The characteristic evolution in GA flow

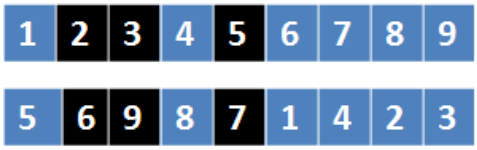
Selection: The selection in each generation is based on the fitness value, which is generated by the fitness function. The higher fitness value has a better chance to be selected for reproduction than the lower one. The higher fitness value can be selected for reproduction on large scale in the next generation. Meanwhile, the lower fitness is eliminated little by little. For each new solution to be reproduced, a pair of previous solutions is selected for breeding from the previous selection pool. By reproducing the next solution using the crossover and mutation methods, a new solution is created.

Crossover: In our algorithm, the position-based crossover (POS) method is chosen to perform crossover in GA flow. The POS is explained as follows:

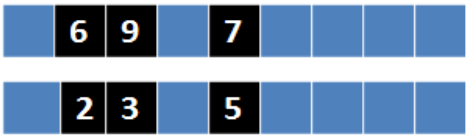
(1) To initially select two chromosomes as parents of next generating chromosomes



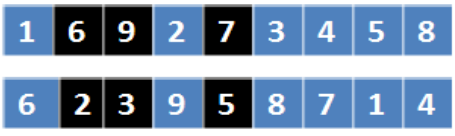
(2) To choose randomly three genes (i.e., 2, 3, 5) in the chromosomes



(3) To exchange the genes of the two chromosomes



(4) To fill in the other genes



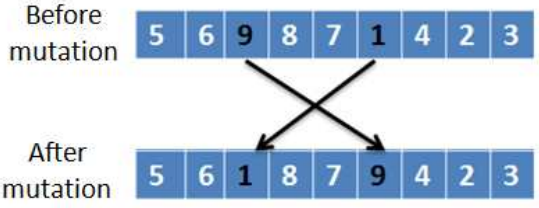


Fig. 3 Chromosome mutation

Mutation: Mutation is the change of genes with probability in a chromosome. In our GA function, we chose exchange mutation by exchanging two or more genes in a chromosome. For example, the mutated chromosome is shown in Figure 3.

4. Cosmetic Tourist Guiding Model (CTGM) Architecture

The CTGM was designed to include three parts: remote server, mobile user, and medical service. The CTGM architecture is shown in Figure 4. The remote server is a back-end server with a scheduling engine and a personal reminder. The engine is constructed by the generic algorithm mentioned in Section 3. The personal reminder is used to generate personal care information according to scenic spots, medical care information, and personal needs. The mobile secretary can query the necessary information from the reminder and tour database and then send the information to the user's mobile device. Furthermore, the secretary could also feedback the user's location to the scheduling engine and personal reminder during the visitor's tour. Therefore, the scheduling engine can reschedule according to the feedback if necessary.

The cosmetic doctor can recommend a user's medical notice through the APITE system, which is a remote consulting system for cosmetic users and doctors. Some facial expression technologies are applicable to the consulting system [17-18]. The user can acquire his/her touring schedule, including medical concerns, through the user browser in mobile devices. The travel agency can maintain the scenic spots in the CTGM and schedule the guest's tour through the user interface provided by the CTGM. From business viewpoint, the whole system is indeed applicable to some product, and the results in some businesses can be beneficial if the medical provider and travel agency agree on a specific service level agreement.

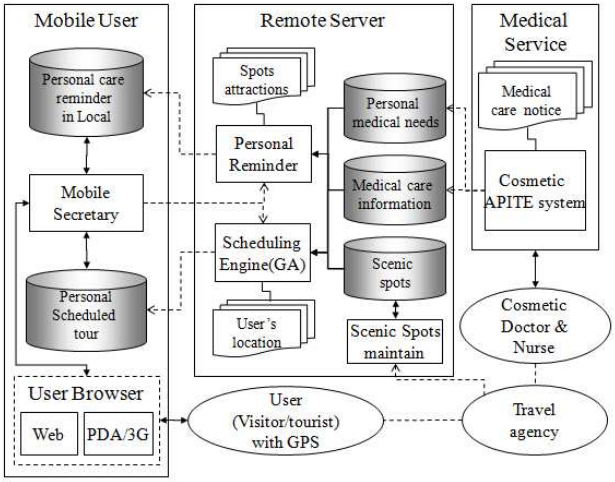


Fig. 4 CTGM architecture

Furthermore, the remote server in the CTGM can provide general business web services that can support GA service to enhance service quality. For example, there are additional components such as user registration and login service, user authorization, global positioning service, real-time weather query service, and suggestion feedback service.

The GA function that includes other factors is a high complexity function. The mobile device is usually equipped with low computing and limiting power. Moreover, the GA function in the server has high maintainability and extensibility. Owing to these viewpoints, the GA is designed to execute from a remote back-end server and dispatch computing results to a mobile user's device.

If a user triggers the scheduling function on the schedule webpage, the scheduling service is remotely called on to invoke a GA object in the server side. Meanwhile, the related factors (i.e., genes) are pre-loaded into the memory through db_gpsdata object to enhance performance. The db_gpsdata is an object that queries the necessary data from the database. This preload activity must be executed before the GA is initialized. In each GA generation loop of the GA evolution, the GA instance executes the following steps: to evaluate fitness, to select two parents, to crossover, and to mutate randomly. The sequence diagram is shown in Figure 5.

5. Evaluation

The components in both the server side and the mobile side were developed using Microsoft Visual C#.NET 2008. They were expressly encapsulated as Web Services for network communication. The server components were installed in an IIS 7.5.7600 server, and the computing data were stored in a mysql database. The designed GA function was installed in the server side and was used to communicate with the components in the mobile side through web service protocols.

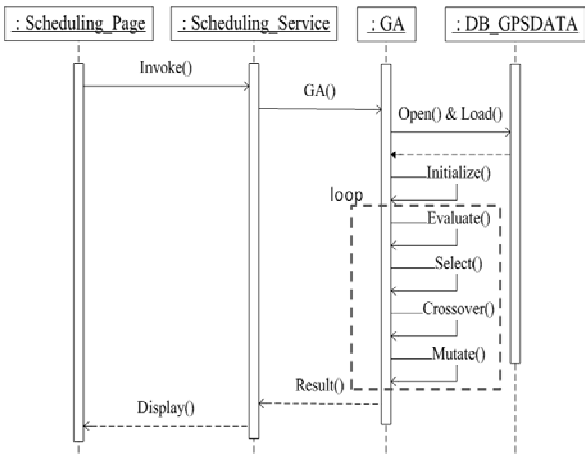


Fig. 5 the GA sequence diagram

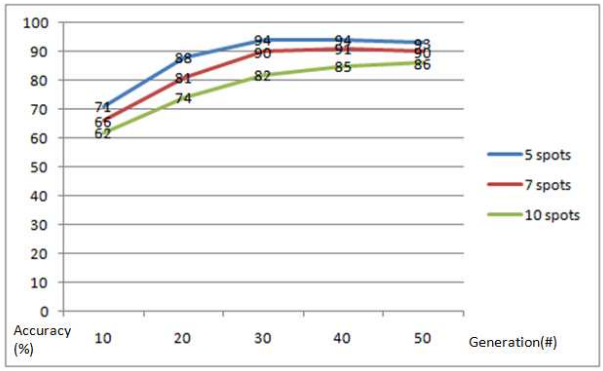


Fig. 6 GA: Accuracy (%) vs. generation (#)

To evaluate the proposed GA function, we scheduled the various spots in a one-day tour in southern Taiwan for visitors. A tour schedule was simulated by combining various numbers of scenic spots, such as 5, 7, and 10 spots, in a one-day tour. For each case, the various results were collected by executing the GA functions 10, 20, 30, 40, and 50 generation times. The GA accuracies of the cases of 5, 7, and 10 spots are shown in Figure 6.

As seen in the curves, accuracy will significantly keep on increasing if the generation times are less than 30. On the other hand, the accurate results will be stable if the generation times of the proposed GA are up to 30. Comparing different curves in the same generation times, the accuracy of 5 spots is higher than that of the 7 spots and 10 spots. The higher complexity resulting from more spots accounts for this. Therefore, the highest accuracy occurs when 5 spots are scheduled in a one-day tour.

6. Conclusion and Future Work

An oversea tourist will require specific post-cosmetic care after he/she had finished the cosmetic treatments and preferred to have a short-term vacation in that area. Generally, the post-cosmetic care during the vacation is important to effectiveness of the cosmetic treatments. In this paper, we aimed to explore technical models including Genetic Algorithm (GA) to realize the combinative business of cosmetic treatment and travel tourism for serving the overseas tourist. The proposed Cosmetic Tourist Guiding Model (CTGM) was connected to the GA and other technologies, and was successfully extended to construct a cosmetic tourist guiding system. The CTGM is composed of the necessary functions and is useful in the business domain of international cosmetic tourism. The relevant technologies in CTGM are GA, lightweight mobility, cosmetic care, and touring knowledge. Cosmetic care and touring knowledge was obtained by interviewing relevant domain experts, such as cosmetic doctors and travel agencies, in Taiwan. In the future, the CTGM model will be extended to provide overseas

visitors with convenient services. The model will also be enhanced to provide an open platform for the business owners to collaborate with each other over the Internet.

Acknowledgments

This research was partially supported by the second phase of the Applied Information Services Development and Integration project of the Institute for Information Industry (III) and sponsored by MOEA, Taiwan R.O.C.

References

[1] K.L. Hsieh, C.C. Shen, C.Y. Chou, C.C. Cheng, "Employing Genetic Algorithm into Constructing the Adaptive Customer's Tour Scheduling Suggestion Model," Tourism Management Research Vol.4 No.1 ,pp. 53-66, June 2005.

[2] A. Maruyama, N. Shibata, Y. Murata, K. Yasumoto and M. Ito, "A Personal Tourism Navigation System to Support Traveling Multiple Destinations with Time Restrictions," in Proc. of the 18th International Conference on Advanced Information Networking and Applications, Volume 2, pp. 18-21, March, 2004.

[3] Kanoh, H. and Nakamura, N., "Route Guidance with Unspecified Staging Posts using Genetic Algorithm for Car Navigation Systems," IEEE Conference on Intelligent Transportation Systems (ITSC 2000), pp. 119–124, Oct. 2000.

[4] Li Yao, Bo Huang and D. H. Lee, "A criteria-based approach for selecting touring paths using GIS & GA," in Proc. of IEEE Intelligent Transportation Systems, Volume 2, pp 958-963, Oct. 2003.

[5] Moon, C. , Kim J., Choi, G. and Seo, Y., "An efficient genetic algorithm for the traveling salesman problem with precedence constraints," European Journal of Operational Research Vol. 140, Issue 3, pp. 606–617, Aug. 2002.

[6] Genetic Algorithm, Wikipedia, http://en.wikipedia.org/wiki/Genetic_algorithm, 2010

[7] Shi-Kuo Chang, "A General Framework for Slow Intelligence Systems", International Journal of Software Engineering and Knowledge Engineering, Volume 20, Number 1, pp. 1-15, Feb. 2010

[8] B. Bontoux, Dominique and Feillet, "Ant colony optimization for the traveling purchaser problem," Computers and Operations Research, Vol. 35, Issue 2, pp. 628-637, Feb. 2008.

[9] Colorni, A., Dorigo, M., and Maniezzo, V., "Distributed Optimization by Ant Colonies," in Proc. of the First European Conference on Artificial Life, Paris, France, F.Varela and P.Bourgine (Eds.), Elsevier Publishing, pp. 134-142, 1992.

[10] Dorigo M., Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italy, in Italian. 1992.

[11] Glover, F. "Tabu Search — Part I", ORSA Journal on Computing, Vol. 1 Issue 3, pp. 190-206, 1989.

[12] Glover, F. "Tabu Search — Part II", ORSA Journal on Computing Vol. 2 Issue 1, pp. 4-32, 1990.

[13] Y. Suna and L. Lee, "Agent-based Personalized Tourist Route Advice System," International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences, Vol. 35, Part 2, pp. 319 -324, 2004.

[14] Das A., Chakrabarti and Bikas K., "Quantum Annealing and Related Optimization Methods," Lecture Note in Physics, Vol. 679, Springer, Heidelberg, 2005.

[15] Yu-Cheng Chen, "A Tour Planning System Based on Solving the Traveling Salesman Problem Using the Taiwan Map," Thesis for Master, Tamkang University, Taiwan, 2007.

[16] Ruo-Ping Huang, "A Comparative Study of Meta-Heuristic Approaches on Data Clustering," Thesis for Master of Business Administration Department of Information Management Tatung University, Dec. 2007.

[17] Chuan-Yu Chang, Yan-Chiang Huang, and Chi-Lu Yang, "Personalized Facial Expression Recognition in Color Image," in Proceedings of the Fourth International Conference on Innovative Computing, Information and Control (ICICIC 2009), pp. 1164-1167, Dec. 2009.

[18] Meng-Hsuan Fu, Yau-Hwang Kuo, Kuan-Rong Lee, and Chi-Lu Yang, "Fusing Remote Control Usages and Facial Expression for Emotion Recognition," in Proceedings of the Fourth International Conference on Innovative Computing, Information and Control (ICICIC 2009), pp.132-135, Dec. 2009.

[19] C. S. Lee, Y. C. Chang, and M. H. Wang, "Ontological recommendation multi-agent for Tainan city travel," Expert Systems with Applications, vol. 36, no. 3, pp. 6740-6753, Apr. 2009.

Remote Query Processing in Wireless Sensor Networks using Coordinated Mobile Objects

Juzheng Li and Sol M. Shatz
Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607
{jli48, shatz}@uic.edu

Abstract. The introduction of mobile devices into wireless sensor network systems has attracted significant attention in recent years. Mobile devices can provide a bridging mechanism to connect geographically separated sensor networks, forming a global coverage sensing network. However, effective techniques for organizing multiple mobile devices and supporting coordinated sensor-data sharing have not yet been established. This paper presents two Coordinated Query Processing mechanisms, CQP-1 and CQP-2. Simulation studies indicate the potential for significant energy savings with the proposed techniques.

Keywords: sensor data, mobile objects, global-coverage sensing networks, query coordination

1 Introduction

A wireless sensor network (WSN) refers to a large collection of spatially distributed sensor nodes with sensing, computation, and communication capabilities. The deployment of sensor networks opens a "window to the world," allowing for sensor-data collection and data fusion for area monitoring, object or event tracking, etc. One unique characteristic of such networks is that sensors pose a severe constraint on design issues such as energy, communication, computation, and data storage [2,3,4]. Generally, sensor nodes are powered by batteries. This constrained energy-supply property results in energy conservation becoming a critical system design consideration. Furthermore, sensor nodes are mostly equipped with relatively limited memory capacity and slower processing units, which restricts data management and processing capabilities. Finally, in terms of communication, data is normally transmitted in a multi-hop fashion. Thus, acquisition of sensor data requires the activation (and hence energy consumption) of target sensors, as well as those sensors along the transmission path. As a consequence, sensor networks are normally designed to serve local sensing queries only. Remote and global querying is generally unsupported.

In contrast to resource-limited sensor networks, wireless mobile networks have significantly relaxed power constraints; and mobile objects (also referred to as mobile devices or mobile nodes) can support large memory and fast processors. Various communication technologies, including Cellular, Wi-Fi, WiMAX, and Bluetooth, have been recently developed for connecting billions of mobile products, such as PDAs, cell phones, laptops, and cars [3,5,9].

Inspired by the above facts, recent research has begun to exploit the interplay between mobile objects and static sensor nodes. Thus, the concept of multi-tier networks, with mobile objects at the "higher" level and sensor nodes at the "lower" level, has been proposed. The mobile devices can be regarded as an overlay network on an existing sensor-level network. This will enable a heterogeneous design space that can offset the stringent resource and power constraints encountered in traditional static sensor networks by taking advantage of the mobility and communication flexibility of the mobile objects. The mobile-object level also provides a bridge to connect individual sensor networks into a global sensing net.

1.1 Remote Query Processing

One class of applications ideally suited for such multi-tier networks is remote sensing. For example, (1) before starting on a particular driving trip, learning beforehand the traffic load can help us wisely choose our best route, and (2) at the start of lunch time, querying for the open-table status at a selection of eating locations can cut our waiting time significantly. An architectural example is illustrated in Figure 1. Note that although mobile objects and sensor nodes are physically located in the same two-dimensional coordinate system, we logically classify mobile objects as being in the "upper" layer, and sensor nodes as being in the "lower" layer.

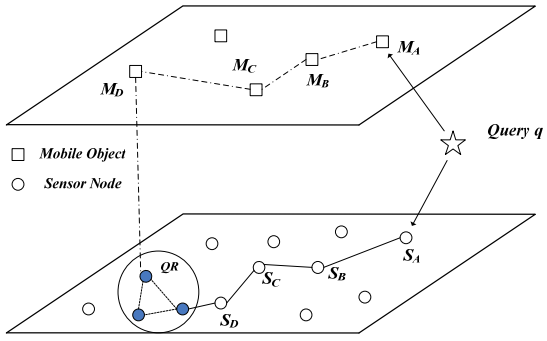


Figure 1. A heterogeneous two-layer design

In the figure, a query q is generated, targeted at some remote region QR. If only the sensor network layer exists, the query would be injected into the sensor layer immediately, to the sensor node S_A in this example. S_A will then route the query through some intermediate sensors to the remote query region. Query results can then be routed back in a similar way, provided that sensor nodes are globally connected [1,16]. Another option, thanks to the employment of the two layer design, is to pass the query to a mobile object, M_A in this example. M_A can then route the query through the upper layer mobile network. At some point, a mobile object, M_D in this

example, receives this query and finds itself close enough to the target query region so that it can inject the query. This option not only achieves a global querying objective without requiring sensors to be globally connected, but it also reduces sensor node energy consumption by using mobile objects to support query (and query result) routing. Intuitively, such savings will be more pronounced when the query generator is farther away from the query region.

Extensive research effort has been engaged with this remote query processing topic. Problems such as how the query should be generated; how to make mobile routing efficient; how to choose an injecting mobile object; and how sensors should reply to the query injections, have been studied [14]. Many questions have been answered regarding this remote query processing topic. However, most previous research focuses only on the processing of one single query, while in practice there are likely to be many concurrent remote queries in the multi-tier networks considered here.

1.2 Multiple Remote Query Processing

A high-level application scenario is illustrated in Figure 2. Mobile object M_1 is handling a query targeted at sensor data within the region labeled RQ_1. Similarly, there can be other mobile objects that are also concurrently interested in specific regions of the sensor field – for example, M_2 and M_3 are handling queries for regions RQ_2 and RQ_3, respectively.

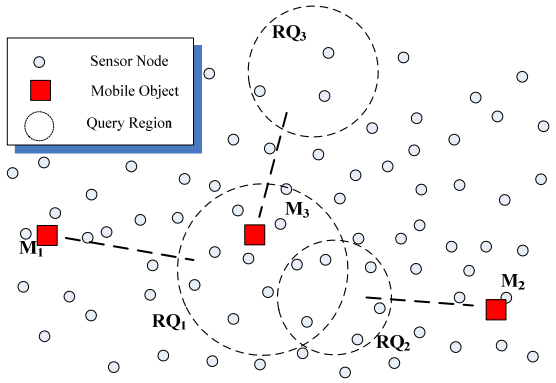


Figure 2. Multiple remote queries

One fundamental problem for multiple remote queries processing that has not yet been well investigated is how to effectively coordinate individual mobile objects with respect to their individually issued remote queries. Note that a key problem is how to coordinate the sensor-data requests (queries actually) in a way that minimizes interactions with sensors and thus conserves sensor-node energy. This can be especially challenging, but of significant value, in the context of applications that impose high-volume data-retrieval requests.

In this paper, we begin to define and study potential coordination mechanisms for the multiple remote queries processing problem. We begin with a baseline naïve approach, Naïve Query Processing (NQP), where no coordination is considered. Later, two more mechanisms are proposed which do allow mobile objects to cooperate with each other to gather and share sensor data. The first one, which we call Level-1 Coordinated Query Processing (CQP-1), resembles the traditional publish-and-subscribe mechanism and explores *query-query (Q-Q) overlapping*. The second one, which we call Level-2 Coordinated Query Processing (CQP-2), takes a totally different approach. It leverages *mobile object-query (M-Q) overlapping*. For example, M_3 and RQ_1 in Figure 2 form a mobile object-query overlapping – M_3 is currently located within the query region associated with RQ_1. Thus, there is an opportunity for Level-2 cooperation between M_3 and M_1, since M_1 is the mobile object with direct interest in RQ_1.

2 Fundamentals and NQP

Regarding query origination, we consider all queries as if they are mobile-object generated. Any non-mobile object generated queries will eventually be handled by a mobile object (Figure 1). Further, we assume that the queries of concern are "one-shot" queries, meaning that a mobile device only requires a one-time query result rather than some on-going, continuous retrieval of information from the query region [6,12,13,14]. Formally, we can represent a query q by a 6–tuple:

q = (Q_ID, M_ID, IT, CL, QR, ERL),

where Q_ID is an identifier for the query; M_ID is the identity of the source mobile device that generates this query; IT and CL represent, respectively, the issue time and current location (x-y coordinate position) of the source mobile device when this query is generated; QR (Query Region) is the region of interest for the mobile device; and ERL (Expected Response Limit) denotes a time limit expressing how long the query issuer is willing to wait to receive the query response.

2.1 Naïve Query Processing

When multiple remote queries are generated in the system, the simplest way is to handle them in a purely distributed fashion, so that each mobile object, and consequently each issued query, is handled independently. For example, mobile object M_1 (from Figure 2) would route its query within the mobile network to the remote query region RQ_1 to retrieve sensor readings of interest. In this sense, mobile objects gather desired data strictly for themselves. Details of how to retrieve sensor data individually have been well studied in previous work [11,14]. Although this approach responds very fast, for query regions such as RQ_1 and RQ_2, which overlap, it is not efficient to interpret and service the queries independently, since this would involve redundant sensor data retrieval and thus increase consumption of sensor node energy. Over time, sensors in areas with frequent sensor reading retrievals ("hot area" nodes) will die first, resulting in the disconnection of the whole network. To avoid this problem, it is desirable for M_1 and M_2 to *share* common-interest sensor data. This requires mobile objects to have some knowledge of each other, and of each other's interests (in the form of queries). Apparently, some form of central authority is needed to provide such knowledge.

3 Q-Q Overlapping and CQP-1

3.1 Query-Query Overlapping

The central authority we employ in this study is a server named the Interest-Management Server (IMS). Note that although we called it a server, it may be actually implemented as a group of brokers [15], as in contemporary distributed architectures. When a mobile object generates a query, the mobile object directs the query to the IMS. Using the terminology from traditional publish and subscribe mechanisms, we can say that the mobile object *subscribes* its interest. The IMS server will then manage the mobile devices' interests (queries) and explore potential cooperation opportunities. The IMS can identify two forms of potential cooperation opportunities: (1) Q-Q overlapping, as will be discussed below, and (2) M-Q overlapping, which will be presented in Section 4.1.

The basic idea of Q-Q overlapping is quite straightforward. RQ_1 and RQ_2 in Figure 2 illustrate an example. This form of overlapping shows that mobile objects M_1 and M_2 share some common interest, and implies opportunities for cooperation. Thus we group those mobile objects in one special purpose group, which we call a Common Interest Group (CIG). To efficiently measure such overlapping, we use the sensor clustering technique [8,10]. In brief, our criterion leverages the geographical partition of the monitored region by sensor clusters. Those mobile devices whose query regions overlap with the coverage region defined by a particular cluster will be grouped together and formed into a common interest group. This paper does not cover the actual clustering process and the algorithms to calculate cluster sizes [8,10], which depends heavily on the particular applications under concern.

Definition: CIG_J is the *common interest group* associated with some cluster J. $CIG_J = \{\text{mobile devices}, m_i \mid \exists q_{ik} \in Q_i \text{ and } QR_{ik} \bigcap CR_J \neq \varnothing\}$

Note that we have relaxed and enlarged the common interest definition. Two queries may not actually overlap physically with each other, but if they both overlap with the same cluster, we still consider this as a case of Q-Q overlapping. Figure 3 illustrates this idea. In the figure, the four big circles represent the four mobile devices' query regions. Based on the CIG definition, mobile devices x, y, and z are elements of CIG_J since their query regions overlap with cluster J's coverage region. Associated with each group there are two important attributes: (1) group address, which is a unique communication identifier reserved for multicasting to the group members; and (2) cluster ID, which is the ID for the cluster that is associated with this group, i.e., group address for CIG_K is GA_K and the cluster ID for CIG_K is C_K;

3.2 Level-1 Coordinated Query Processing

CQP-1 exploits query correlations and the existing and growing communication capability associated with modern mobile devices. Members of the same common interest group communicate for the purpose of sharing sensor-data information. When a mobile device obtains sensor data from sensors of an intended query region, that mobile device will publish this data to peers that belong to the same group, i.e., to other mobile devices with a common interest in this sensor data. This will successfully avoid the expense associated with each mobile device needing to individually acquire data from sensors. Consequently, this sharing will save considerable energy for the sensor nodes.

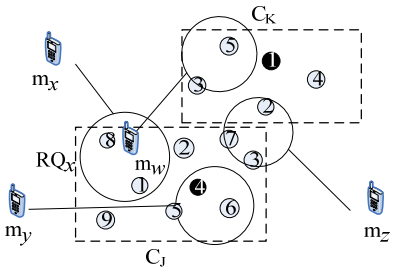


Figure 3. Clusters, query regions and groups

Remember that mobile objects will subscribe their interest by a form of query to the interest-management server, expecting to participate in the coordinated query processing system and receive cooperation help. Those subscription messages can then be utilized by the server to perform the action of joining the mobile device to one or more groups according to the device's specified query region, thus allowing for future coordination. The server checks for any overlap between the query's query region (QR) and the sensor clusters' coverage regions (CRs) to determine if the sensor deployment includes sensors within the query region. If there is no such overlap – due to low sensor-density or very small query regions – the query cannot be served and a failure message will be returned to the mobile device. This is simply due to the fact that the query region is not covered by any sensors. But, if there is such overlap, the server identifies those common interest groups associated with the overlapped clusters and adds the current subscribing mobile device to those common interest groups. Finally, a CQP-1 Instruction Message – containing the group addresses and the corresponding coverage regions for those identified common interest groups – will be sent to the mobile device. A pseudo-code description of the server's responsibility for processing CQP-1 for mobile devices is given in Figure 4. The mobile object's protocol is summarized in Figure 7.

```
1   //server logic for query reception
2
3   IF ((QR_ij ⋂ any cluster CR) ≠ empty)
4
5       FOREACH (Coverage Region, CR_M, in the above found overlapping CRs)
6
7           find the corresponding cluster ID C_M;
8
9               IF ( ∃CIG_M and (Cluster ID for CIG_M == C_M))
10                  find the corresponding group address GA_M;
11                  join md_i to CIG_M through GA_M;
12
13              ELSE IF (there is no CIG associated with C_M)
14                  set up a new CIG, CIG_M and join md_i to CIG_M;
15
16          send an CQP-1 Instruction Message to m_i containing GA_M and CR_M;
17
18  ELSE query region is not covered by sensors, return a failure message to m_i;
```

Figure 4. Server logic in response to query reception

Once a mobile device joins a group, it has the potential to become involved in CQP-1 cooperation among those group members, either by receiving sensor data from peer members or sending such data. Referring to Figure 3 again, we can easily understand that within common interest group CIG_J, if mobile object M_y obtains the information about CIG_J's common interest, it will share its observation with all of the CIG_J's group members by a form of CQP-1 Date-Share Message, which in this case will multicast the information to M_x and M_z. This multicast can be easily and efficiently implemented given the common infrastructure, for example, cellular-based networks supporting the mobile objects.

4 M-Q Overlapping and CQP-2

4.1 Mobile Object-Query Overlapping

The M-Q overlapping is similar with Q-Q overlapping in that we also use the sensor clustering technique to determine actual overlaps. An M-Q overlapping occurs when a mobile object M locates nearby, or even within, a query region Q established by another mobile object at the time that M issues its own query. Strictly speaking, if M and Q overlap with the same cluster, then M and Q form an M-Q overlapping. Mobile object m_w and query RQ_x in Figure 3 illustrate such a case where m_w locates within C_J and RQ_x overlaps with C_J.

4.2 Level-2 Coordinated Query Processing

To further explore the potential cooperation behavior among multiple mobile objects, CQP-2 is designed. The CQP-2 processing is an interesting behavior. The principle behind it is based on the observation that mobile devices can efficiently obtain local knowledge regarding sensor nodes that are in close proximity (without needing to route queries or query results in the network). With the help of the subscribe process, such local knowledge can be of great value to our cooperation mechanism. To actually take advantage of this local knowledge, every time a mobile device m_D subscribes – by directing a query to the interest-management server – we identify those groups that share an interest in sensor data associated with m_D's current location. Then mobile device m_D can publish its locally-obtained sensor data to the mobile devices belonging to those groups. Note that in CQP-2 the mobile object that provides local sensor data does not belong to the same groups as mobile objects that receive this data. For example, mobile object m_w in Figure 3 has the potential of providing CQP2 cooperation – in this case, providing sensor data from region RQ_x to mobile object m_x.

Figure 5 elaborates the steps that define the CQP-2 cooperation processing. Step 1 is the initial sending of a query to the server as a subscription. In Step 2, the server will check if the *Current Location* (contained in the subscription message as attribute CL) associated with this mobile device belongs to any group's coverage region. Current information on existing coverage regions is stored in the server's cluster scheme database. If not, no CQP-2 cooperation can be performed. Otherwise, since the mobile device is within the coverage region of an existing group, for example CIG_P, the server can perform an CQP2 cooperation for CIG_P as follows: First, the server will respond with an CQP2 Instruction Message (Step 3) that contains the group address, GA_P, of CIG_P, thus establishing a temporary member-group relationship between the subscripting mobile device and CIG_P. In response to receiving this message, the mobile device will retrieve sensor data from the cluster head of Cluster C_p (Step 4) and multicast this updated data to members in CIG_P (Step 5). This multicasting, or transmitting, of updated data is achieved by sending an CQP2 Data-Share Message to the group address GA_P, established in Step 3. In response to receiving such a message, each mobile device can then extract from the aggregated data the specific information that is important to their own query region, assuming a lossless compression scheme for sensor data-aggregation technology is employed and the original data can be reconstructed [7].

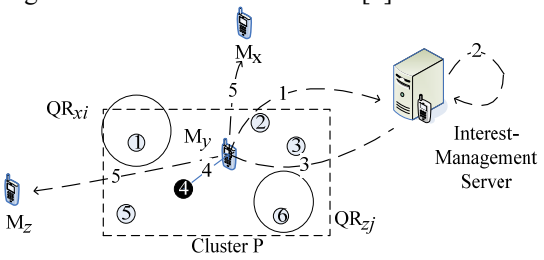


Figure 5. CQP-2 coordination

After the mobile device provides updated information for its temporary group, the temporary member-group relationship is removed. A pseudo-code description of the server's responsibility for processing CQP2 for subscribing mobile devices is given in Figure 6.

```
1   // code for instructing an CQP-2 cooperation
2
3   WHILE (request message queue ≠ empty)
4
5       extract the first request message, qij, from queue;
6
7       IF (CLij ∈ any CR)
8           Find the corresponding cluster ID, CK, of this CR;
9           IF (   CIGK and (Cluster ID for CIGK == CK)
10              find the corresponding group address, GAK, of the CIG;
11              send an CQP-2 Instruction Message to mi containing GAK;
12
13      ELSE mi cannot contribute a CQP-2 cooperation;
```

Figure 6. Server logic for CQP-2 coordination

5 Query Expiration & Mobile Protocol

5.1 Query Expiration

So far we have seen two cooperation approaches for mobile objects to receive desired sensor data: CQP-1 and CQP-2. Remember that the ERL (Expected Response Limit) defines a time limit for the query result retrieval. A natural question is what happens if the coordination system cannot return a query response within that limit - how can we handle such a query expiration event?

Definition: The *query expiration* event for a query q_{ij} occurs at time $IT_{ij}+ERL_{ij}$, in the case that q_{ij} does not receive any query response before $IT_{ij}+ERL_{ij}$.

A simple and naïve way to handle such a situation is for the query issuer to simply accept that this particular query fails and maybe later regenerate a new identical or similar query if the sensor data is still of interest. Alternatively, the mobile object can give up the coordinated query processing system, and instead handle the query independently. That is, the mobile object will route the query and then inject the query. Note that this solution forces us to retreat to our starting point of viewing the whole architecture in a pure distributed way and process each query individually, as discussed in Section 1. We call this back-up measure "Query Injection (QI)". The query injection means actual interactions between a mobile object and sensor nodes. Since we are trying to minimize such interactions, we would like to reduce the probability of using QI.

5.2 Mobile Object Protocol

With the server-side protocols summarized in Figure 5 and Figure 6, we can now show how mobile devices would behave. Figure 7 summarizes the event-driven protocol for mobile objects.

```
Upon receiving remote sensor data retrieval requirement
    Generate a query and subscribe
Upon receiving CQP-1 Instruction Message
    Associate the subscribed query with the received Group Address and Coverage
      Region;
Upon receiving CQP-2 Instruction Message
    Obtain the Group Address GA_i, attached in the message;
    Retrieve sensor data from cluster head ch_j from Cluster C_j;
    Multicast the sensor data using the group address GA_i;
Upon receiving CQP-1 Data-Share Message
    Extract from the received data the specific information that is important to its
      own query;
Upon receiving CQP-2 Data-Share Message
    Extract from the received data the specific information that is important to its
      own query;
Upon Query Expiration
    Conduct a Query Injection;
Upon receiving Query Injection Response
    Extract from the received data D the specific information that is important to its
      own query;
    Multicast the received data D using the group address GA_i;
    // GA_i is the group address received in CQP-1 Instruction Message
```

Figure 7. Mobile device protocol

6 Simulation Analysis

We implemented our group-based coordination protocol under the dotNet Compact Framework. The monitored environment was set to 5000 units width and 5000 units length with 25000 sensor nodes deployed evenly. We divided the sensors into 25 clusters according to their locations. To simplify the simulation, we assume the clusters' coverage regions are same-size squares. 2000 mobile devices were deployed randomly in this environment with speed set to a random number between 1 and 100 units while leaving the motion direction randomly chosen from east, west, south and north every simulated time unit. Since the primary focus of this work is on mobile object coordination, we employed a simplified sensor communication MAC layer protocol, in which sensors have a 5% packet loss rate when communication is within a unit disk communication range. A more complex collision-decay MAC model that takes into consideration MAC layer packet collisions and transmission signal decay was studied in previous work that focused specifically on mobile object and sensor interactions [14].

Queries were generated with the system clock at a specified query generation rate (QGR), which defines how many new queries are generated by the system per simulated time unit. For each issued query, the source mobile device and target region are both randomly selected. To place focus on the coordination mechanism, we defined each query to cover properly one cluster in the environment. Each instance of a simulation run used 480 simulated time units (t = 480), and the results were averaged over 50 simulations and studied below.

As discussed, one primary objective of this architecture is to reduce the mobile/sensor interactions and to shift as much communication burden as possible into the mobile network layer and its associated supporting infrastructure. To measure our design's efficiency, the previously discussed concept named query injection (QI) is used.

In Figure 8, we can see that using the NQP approach, the percentage of QI stays at 100%. This is because NQP considers no cooperation at all, so that to process each query, a QI is needed. However, as cooperation behaviors are introduced, the percentage drops quite a lot. With CQP-1 enabled, less than 20% of queries rely on a QI and when both CQP-1 and CQP-2 are enabled, the same measure is reduced to 5%. It shows clearly the power of employing proper cooperation behaviors. All simulations in Figure 8 were conducted under an ERL of 10 and a QGR of 10.

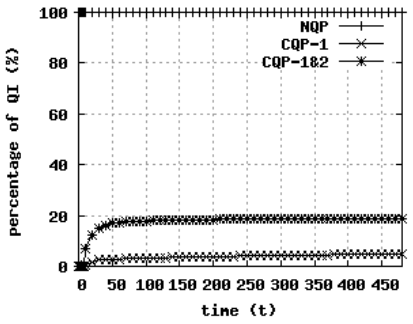


Figure 8. % of Query Injections under 3 approaches

In Figure 9, we fixed the expected response limit, ERL, at 10 to study the impact of query generation rate (QGR) when both CQP-1 and CQP-2 are enabled. Surprisingly, we found that when the workload increases, the percentage of query injection decreases. This is because, as query generation rate increases, the number of members in each group increases. Consequently, the likelihood that members of this group obtain a query-response increases, since there are now more peers looking to achieve this common interest. Once any mobile device receives the common interest data, all peers in the group also receive that data. So, the enhanced

cooperative behavior offsets the increase in query generation rate. Another vital contributing factor to the observed system performance is the CQP-2. As we increase the query generation rate, there are more opportunities for CQP-2 cooperation. The combination of CQP-1 and CQP-2 reduces the impact of an increasing workload.

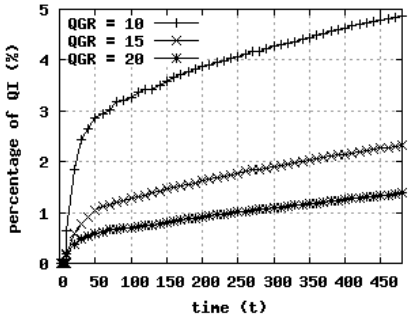


Figure 9. % of QI vs QGR with CQP-2 coordination

The other simulation is based on four different expected response limit (ERL) rates to evaluate the sensor energy consumption of our approach when the system receives queries with different ERL values. Results are plotted in Figure 10 with the QGR set to 10. As expected, a longer ERL can produce a lower query injection percentage, since the increase in ERL gives the queries more time to use our coordination mechanism rather than require query injection. We noticed that the worst percentage is 16.8% when we set ERL to 5. If we can extend the ERL to 10, this statistic can be lowered to 4.87%, meaning that for every 100 queries that are serviced, less than 5 of them require energy associated with sensors for the query and query-result routing.

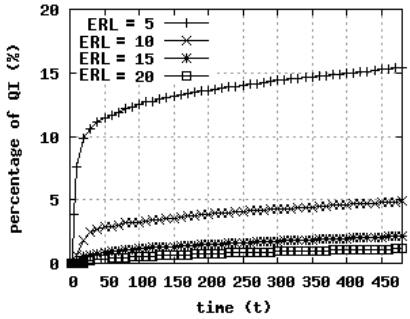


Figure 10. % of QI vs ERL with CQP-2 coordination

The ERL sensitivity result firmly supports our idea of deploying our coordination mechanism in the context of WSN systems, many of which are delay-tolerant to some extent [2].

6 Conclusion

This paper presents coordinated query processing mechanisms for handling mobile generated remote queries in mobile-sensor multi-tier heterogeneous networks. Three approaches were discussed: a naïve query processing approach, NQP, and two coordinated query processing mechanisms, CQP-1 and CQP-2. Simulation analysis indicates the potential for significant energy savings with the proposed techniques.

Two topics for future research are "stream" query processing; and query result caching. To process and provide service for queries that require continuous data stream retrieval, one solution is to cut a "stream" query into several "one-shot" queries according to a specified sampling rate and process each such resulting query. To further optimize the coordination system, a mobile object can cache query results received. Subsequent queries can then use ("recently") cached results. In this sense, the mobile objects form a distributed knowledge storage structure about the sensor field.

References

[1] J. Al-Karaki and A. Kamal, "Routing Techniques in Wireless Sensor Networks: A Survey," *IEEE Wireless Communications*, vol. 11, no. 6, Dec. 2004.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "Wireless Sensor Networks: A Survey," *IEEE Computer*, vol. 38, no. 4, Mar. 2002.

[3] C. Chen, J. Ma and K. Yu, "Designing Energy-Efficient Wireless Sensor Networks with Mobile Sinks," *ACM Conference on Embedded Networks Sensor System*, pp. 195-206, Nov. 2006.

[4] S. Cheng and M. Wu, "Optimization of Multilevel Power Adjustment in Wireless Sensor Networks," 14th Int. Conf. on *Distributed Multimedia Systems*, pp. 109-121, Sept. 2008.

[5] E. Ekici, Y. Gu and D. Bozdag, "Mobility-Based Communication in Wireless Sensor Networks," *IEEE Communications Magazine*, vol. 44, no. 7, Jul. 2006.

[6] B. Gedik and L. Liu. "MobiEyes: A Distributed Location Monitoring Service using Moving Location Queries," *IEEE Transactions on Mobile Computing*. vol. 5, no. 10, 2006.

[7] T. He, B. M. Blum, J. A. Stankovic and T. Abdelzaher, "AIDA: Adaptive Application-Independent Data Aggregation in Wireless Sensor Networks," *ACM Transactions on Embedded Computing Systems*, vol. 3, no. 2, May 2005.

[8] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy–Efficient Communication Protocol for Wireless Sensor Networks," *Hawaii International Conference System Sciences*, Jan. 2000.

[9] S. Jain, R. C. Shah, G. Borriello, W. Brunette and S. Roy, "Exploiting Mobility for Energy Efficient Data Collection in Sensor Networks," *Mobile Networks and Applications*, vol. 11, no. 3, Jun. 2006.

[10] J. Liu and C. R. Lin, "Energy-efficiency Clustering Protocol in Wireless Sensor Networks," *Ad Hoc Networks Journal*, Elsevier Science, vol. 3, no. 3, May 2005.

[11] J. Li and S. M. Shatz, "Toward Using Node Mobility to Enhance Greedy-Forwarding in Geographic Routing for Mobile Ad Hoc Networks," *International Workshop On Mobile Device and Urban Sensing* (MODUS), Apr. 2008.

[12] S. R. Madden, M. J. Franklin, J. M. Hellerstein and W. Hong, "Tinydb: An Acquisitional Query Processing System for Sensor Networks," *ACM Transactions on Database System*, vol. 30, no. 1, 2005.

[13] N. Sadagopan, B. Krishnamachari and A. Helmy, "Active Query Forwarding in Sensor Networks (ACQUIRE)," *Ad Hoc Networks Journal*, Elsevier Science, vol. 3, no.1, Jan. 2005.

[14] S. Tian, S. M. Shatz, Y. Yu and J. Li, Querying Sensor Networks Using Ad-hoc Mobile Devices: A Two Layer Networking Approach, *Ad Hoc Networks Journal,* Elsevier Science, vol. 7, no. 5, 2009.

[15] J. Wang, J. Cao, J. Li and J. Wu, "MHH: A Novel Protocol for Mobility Management in Publish/Subscribe System," *International Conference on Parallel Processing*, pp. 1-8, Sept. 2007.

[16] A. Woo, T. Tong and D. Culler, "Taming the Underlying Challenges of Multi-hop Routing in Sensor Networks," *The 1st ACM Conference on Embedded Networked Sensor Systems,* pp. 14-27, Nov. 2003.

Congestion Optimization in Data Networks

Doraid Dalalah
Assistant Prof., IE Dept., Faculty of Eng., Jordan University of Science and Technology
P.O. Box: 3030, Irbid 22110
doraid@just.edu.jo

***Abstract*—In this paper, we propose a decentralized optimization approach for flow control in data networks. The presented model maximizes the aggregate utilities taking in consideration the links' capacities constraints. A decentralized algorithmic solution is presented through relaxing the constraints and imposing a penalty function on the surplus in the constraints. The algorithm provided in this paper operates in a decentralized environment and falls in the category of Additive Increase, Multiplicative Decrease. The network links and data sources are considered as processors of a distributed computational system having a global objective function. The model works with different shapes of utility curves under the proposition of elastic data traffic. The approach relies on real-time marking of the data packets at the routers. Finally, we present an extensive simulation results to demonstrate the reliability of the algorithm.**

1. Introduction

A great deal of literature exists on flow control including the original TCP self-clocking flow control [1, 2, 3], the economic optimization approaches [4, 5, 6, 7, 8, 9] and control theoretic based solutions [10, 11, 12]. Theoretical research on this topic has dealt mostly with simple scenarios, i.e. single-bottleneck and per-flow queuing [13]. However, in the last few years, significant advances have been taken in bringing analytical models to flow and internet congestion control. Explicit modeling of congestion on the other hand has been widely used [5, 6, 7, 8, 11, 14, 15]. Although the traditional TCP-Reno performed remarkably above expectations and is still widely used, there are some limitations that motivate the advances in flow and congestion control. The work in [16] and [17] focused mainly on demonstrating the rate-to-queue proportionality. Proportional fairness was first presented by [5] particularly on the data transmission rates. The work presented by [18] and [19] shows that it may be applied to the queues as well.

In this work, we assuming the minimum roundtrip delay as an available quantity, the algorithm will work without any access to global information or router support. However, for the sake of improved performance, we use the REM approach presented in [20] and [21] as a tool to help find the roundtrip delay at the transport layer of each source.

2. The Basic Model

We assume an elastic traffic and model the network as a set of resources (links) L with constant transmission capacities $c_l \; \forall \; l \in L$, and a set of data sources R. Each source i traverses a subset $L(i) \subseteq L$ and attains a utility of $U_i(x_i)$ when the source i transmits at a rate of x_i. Following [20], we call the traffic which leads to such utilities as elastic, where $U_i(x_i)$ is an increasing strictly concave function and $lim_{x \to 0} U_i'(x) = \infty$, $lim_{x \to \infty} U_i'(x) = 0$. Each link l will have a subset of data sources $R(\ l\ )$ such that each source traffic will traverse the link l, i.e. $R\ (\ l\ ) = \{ i \in R \mid l \in L(i)\ \}$. Set $A_{ij} = 1$ if $j \in L(i)$, that is, resource j falls on the route of i, and set $A_{ij} = 0$ otherwise. This defines the routing matrix $\boldsymbol{A}$ between the users and the links. Assuming further, that the utilities are additive, the aggregate utility of the rates is given by $\sum_{i \in R} U_i(x_i)$. Under this model, the system optimal rates solve the following problem (P):

$$P: Max \sum_{i \in R} U_i(x_i)$$
$$st.:$$
$$\boldsymbol{A}^T \boldsymbol{x} \leq \boldsymbol{C}$$
$$\boldsymbol{x} \geq 0$$

where $\boldsymbol{x}$ is the transmission rate vector and $\boldsymbol{C}$ is the links capacity vector. The above optimization problem is mathematically fairly tractable, with a strictly concave objective function over a convex feasible region.

3. Constraint Relaxation

In this section, we relax the constraint set by introducing a penalty on the deviation of the constraints. At a certain router we measure the deviation of the constraint by a suggested function shown in Fig.1.

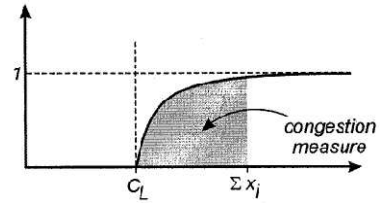


Fig. 1: Congestion pricing.

This function will be used to measure the charges of each transmission session path that will be explicitly fed back to the users using ECN marking. A congestion pricing of the above form for a link *l* can be modeled as:

$$\int_{C_l}^{\sum_{r\in R(l)} x_r} 1 - e^{-\{[y-C_l]^+\}}\, dy \qquad (1)$$

Where $[y - C_l]^+ = \max\{0, y - C_l\}$. This quantity can be calculated at each router and fed back to the end-users as a total charge for using the network resources. Since the objective is a utility maximization, the problem can be remodeled with the congestion prices as a penalty on the relaxation of the set of constraints. This is called a barrier function method in optimization techniques [21]. Hence, problem (P) will look like:

$$max \sum_{i\in R} U_i(x_i) - \beta_i \sum_{l\in L} \int_{C_l}^{\sum_{r\in R(l)} x_r} 1 - e^{-\{[y-C_l]^+\}}\, dy \qquad (2)$$

where β_i is a design parameter. The above optimization is a joint team game that admits a unique equilibrium and has a unique maximum, where the users jointly optimize the objective function. As a result we have,

$$\frac{df}{dx_i} = U'(x_i) - \beta_i \sum_{l\in L(i)} 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+} = 0 \qquad (3)$$

The equilibrium point at which all the users will be satisfied can be found from (3). That is:

$$x_i^* = U_i'^{-1}\left\{\beta_i \sum_{l\in L(i)} 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+}\right\} \qquad (4)$$

Or alternatively: $$x_i^* = U_i'^{-1}\{\beta_i q_i\} \qquad (5)$$

where q_i is the aggregate congestion measure along the path of connection *i*.

Following [21], the optimal solution x_i^* of the above optimization problem in equation (2) approaches the optimal solution of the optimization problem (P) as β_i approaches infinity. This is, high values of β_i mean that the solution in equation (5) gets closer to the optimal point of problem (P).

4. A Decentralized Rate Control Approach

The problem mentioned in (2) is solved using the local information available at the routers and the data source sides. We show that the steepest decent method is a good choice to solve problem (2). That is:

$$x(t+1) = x(t) + \gamma \nabla f \qquad (6)$$

where γ is a small step size. The above equation assumes the knowledge of the gradients of the function *f* at the end data sources, which in reality not true. Therefore, we assume the quantity $\sum_{l\in L(i)} 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+}$ to be available at the end users as a total congestion price for the path of session *i*. Section 6 will show how to estimate this quantity at the end application. Accordingly, the rates control can be given by:

$$x_i(t+1) = x_i(t) + \beta_i \left\{U'(x_i) - \gamma_i \sum_{l\in L(i)} 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+}\right\} \qquad (7)$$

The above control law is a decentralized version of the steepest method where each data source will increase or decrease its rate based on the congestion price. Following [20] we assume elastic traffic with a utility curve given by: $U(x_i) = \alpha_i \log x_i, \forall i \in R$. Substituting this in equation (7) we get the following:

$$x_i(t+1) = x_i(t) + \beta_i \left\{\frac{\alpha_i}{x_i} - \gamma_i \sum_{l\in L(i)} 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+}\right\} \qquad (8)$$

5. The Role of the Router

The congestion charges in this model are conveyed using Explicit Congestion Notification, [22]. The idea behind it is to mark packets which encounter congestion. This study measures the average arrival rate over a certain period of time, preferably, the minimum roundtrip time, then, calculates the marking probability at the receivers. For a router *l*, the marking policy is given as:

$$p_l = 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+} \qquad (9)$$

where the end-to-end marking probability neglecting the delay is given by:

$$q_i = 1 - \prod_{l\in L(i)} (1 - p_l) \approx \sum_{l\in L(i)} p_l \qquad (10)$$

Fig.2 depicts the marking role of the routers.

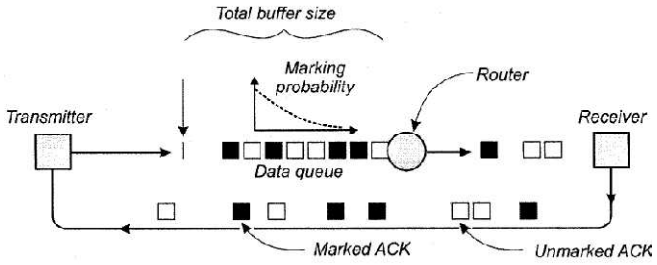


Fig. 2: Marking illustrated.

Although the current working TCP does not completely use ECN, however, the trend in this field is heavily encouraged for its use. In our model, the data sources will find the average of the marked packets using a counting

process. With an accurate mapping, the amount $\sum_{l\in L(i)} p_l$ can be estimated with minimum communication between the network links and the end entities.

6. Penalties on Packet Loss

To make the model packet-drop aware, we add a penalty on the drop fraction that is proportional to each session transmission rate. Let the congestion dropped fraction of x_i be ε_i, the new maximization problem will look like:

$$Max: \sum_{i\in R}(U_i(x_i) - \varepsilon_i x_i) - \beta_i \sum_{l\in L} \int_{C_l}^{\sum_{r\in R(l)} x_r} 1 - e^{y-C_l}\, dy \quad (11)$$

The above optimization problem is strictly concave and admits a unique maximum over the region $x_i > 0$, $\forall i \in R$. The numerical solution of the above problem is:

$$x_i(t+1) = x_i(t) + \beta_i \left\{ \frac{\alpha_i}{x_i} - \gamma_i \sum_{l\in L(i)} 1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+} - \eta_i \varepsilon_i \right\} \quad (12)$$

which is a decentralized solution of the problem:

$$Max: \sum_{i\in R} \alpha_r \log(x_r) - \gamma_i \sum_{l\in L} \int_{C_l}^{\sum_{r\in R(l)} x_r} 1 - e^{y-C_l}\, dy - \eta_r \varepsilon_r x_r \quad (13)$$

where $\beta_i, \alpha_r, \gamma_i, \eta_r$ are locally defined constants.

7. Summary of the Control Method

To implement the decentralized solution, the routers will mark the packets using the given probability distribution:

$$1 - e^{-\{\sum_{r\in R(l)} x_r - C_l\}^+}$$

The data sources will compute the average of the marked packets (q_i), which represents the congestion price and then will monitor and calculate the packet loss fraction ε_i. Next, they will calculate their rates according to:

$$x_i(t+1) = x_i(t) + \beta_i \left\{ \frac{\alpha_i}{x_i} - \gamma_i q_i - \eta_i \varepsilon_i \right\}$$

8. Simulation

This section presents simulation results of several examples for our model. NS-2.26 software package were used for simulation. The data agents in our simulation runs are an extension of the CBR agents where the rate is update every time an ACK is received. We assume full duplex links and constant packet size of 1 kbyte with identical utility curves for the different sessions.

Experiment I: This experiment shows a simple example consisting of single user and a single full duplex link network, Fig. 3.

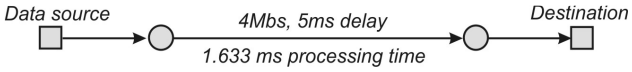


Fig. 3: Network simple example, minimum roundtrip delay = 11.633ms.

The full duplex link capacity is 4Mbit/s. The parameters used for this example are: β_i =10, 100, α_i = 1, η_i = 0.1, γ_i=0.01 and a queue limit of 5 packets. We tested our example for two different values of β_i. Fig. 4 shows the output rate for the case of β =100. Clearly, with higher β_i values, the algorithm approaches the optimal solution faster than lower values. However, larger values make the control law more aggressive and more oscillations may be observed. On the other hand, lower values of β are preferred to reduce oscillations at the expense of longer equilibrium times.

When congestion is detected at a full capacity, the router will start marking the packets and congestion charges are conveyed back to the data sources. The control law at the data source will reduce the rate according to the value of the estimated congestion of the arriving packets..

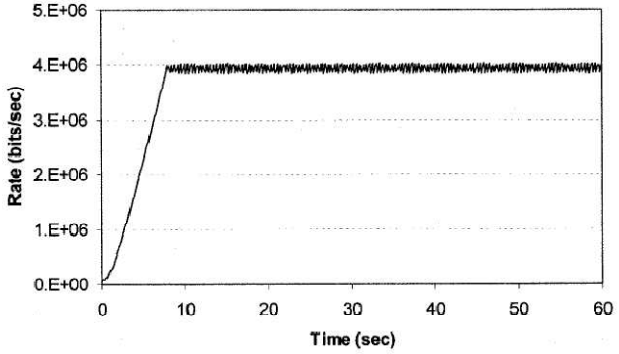


Fig. 4: Rate vs. time, β = 100.

Fig. 5 shows the output of same example for β =10. Clearly, the probing slope is less steep as compared to the case of β = 100.

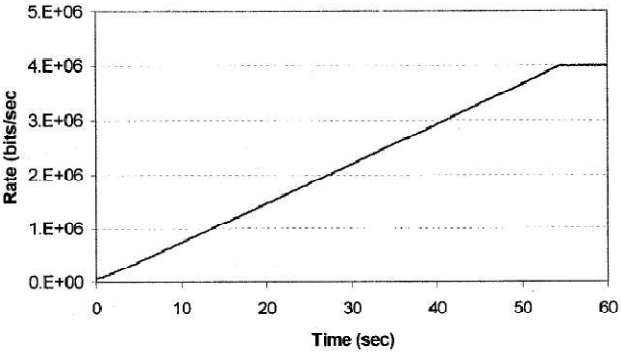


Fig. 5: Rate vs. time, β = 10.

Experiment II: Fig. 6 shows a simple network with 4 users and a single full duplex link of 4Mbit/sec capacity. The simulation is run for 140 seconds and the users start

their transmission sessions in sequence respectively separated by 20 seconds time periods. The users stop their sessions in the reverse order with 20 seconds between each. Therefore, at a certain time, all the users will have active transmission sessions and will share the available bandwidth. The model parameters used for this simulated example are: β_i =100, α_i = 1, η_i = 0.1, γ_i=0.01 and a queue limit of 5 packets.

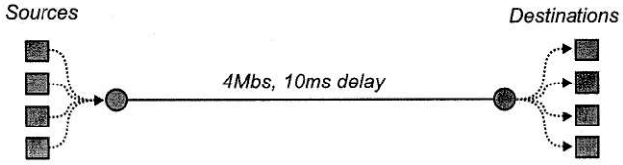


Fig. 6: A single link network example.

Clearly, for the first few seconds when only one traffic type traverses the link, the user will increase its rate proportionally to $(\alpha_i\beta_i/x_i)$ until the link is saturated. Later, it will oscillate around the optimal solution which is the maximum capacity. When the second user joins the network at t = 20 seconds and after one roundtrip delay, both users will start working on a new rate which will divide the link fairly between both. At t = 40, the three existing data sources will agree on new rates. Fig. 7 depicts the calculated rates.

Fig. 8 shows the utilization of the single link. We notice that the link is almost 100% utilized through the transmission periods of the four users. The downward spikes at the moment a user leaves the network are due to immediate bandwidth availability because of session termination.

Experiment III: Next, we present a more heterogeneous network instance in which a link of l0Mbit/s is shared between 20 users as shown in Fig. 9. The parameters used are: β_i =100, η_i = 0.1, γ_i=0.03 and a queue limit of 50 packets.

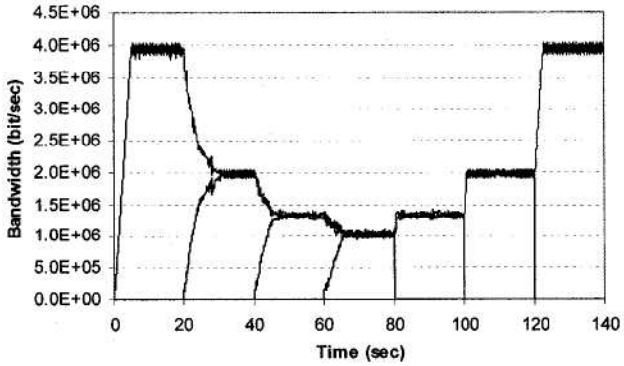


Fig. 7: Data source rates, four data sources and a single link of 4Mbit/s.

The data sources start their transmission sessions 10 seconds apart. The simulation was run for 210 seconds so that at the end, the twenty users all have active sessions. Fig. 10 demonstrates the calculated rates of all the users. Noticeably, at the beginning of the simulation, the first user will have the full link capacity available. When the other users start joining the network, individual rates are reduced. At the end of the simulation, each user enjoys the bandwidth of approximately 0.5Mbit/sec. The aggregate traffic traversing the link is shown in Fig. 11. During the first 100 seconds, the link tends to be underutilized. However, by increasing the number of the users, the link utilization approaches its maximum average.

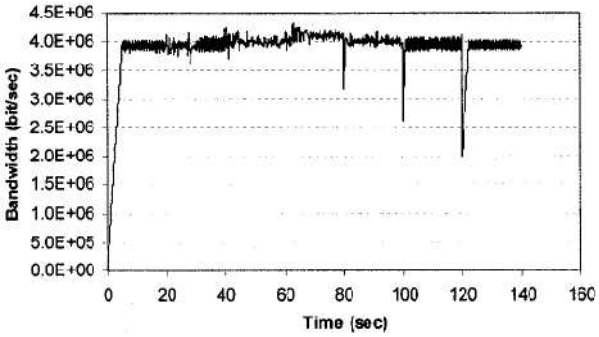


Fig. 8: Aggregate traffic traversing the link.

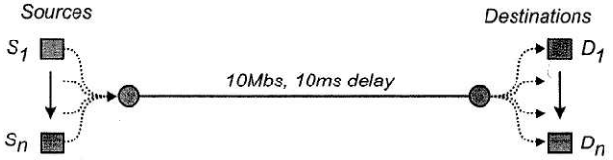


Fig. 9: Heterogeneous network example: a single link and 20 users.

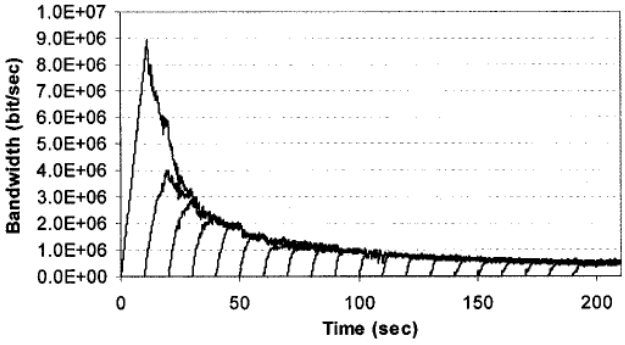


Fig. 10: Calculated rates for a network of 10Mbit/s capacity and 20 data sources.

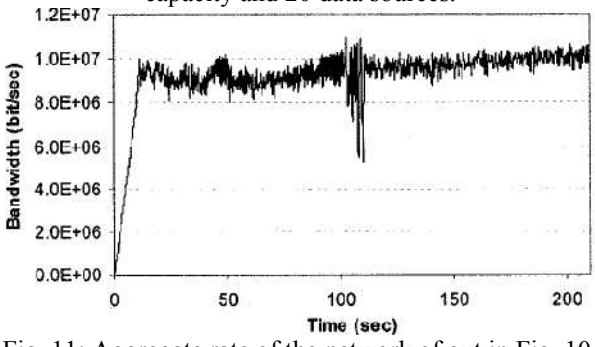


Fig. 11: Aggregate rate of the network of out in Fig. 10.

Experiment IV: Next, we present a reversed scenario of the last example using the same parameters. That is, all the data sources will start their transmission sessions simultaneously at the beginning of the simulation and will successively terminate in a period of 10 seconds, respectively. Fig. 12 depicts the individual calculated rates. The graph shows that all the users approximately agree on the theoretical-optimal point. The last transmission session will approximately approach the full available capacity of the link.

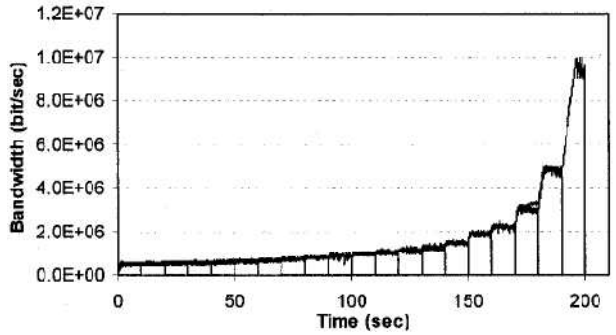


Fig. 12: the transmission rates of 20 traffic flows traversing a 10Mbit/s link.

Fig. 13 depicts the aggregate traffic traversing the link. The figure demonstrates more oscillation when more bandwidth is available due to session termination.

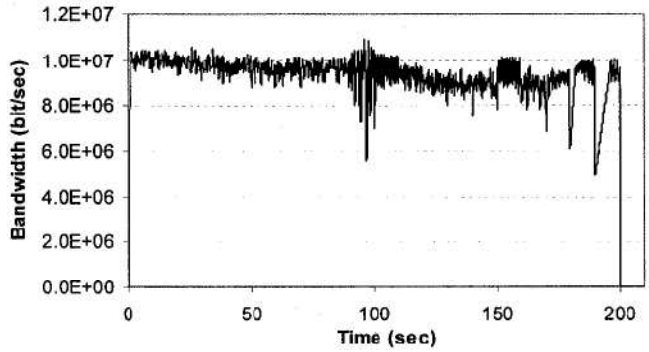


Fig. 13: Total aggregate rate. 20 traffic flows traversing a 10Mbit/s link.

From the above simulated example we notice that the criteria of choosing the parameters, in particular β, depends on the expected number of users and the link capacities. In our simulation, a range of $1 \leq \beta \leq 100$ is acceptable for networks with links of the range 0 to l00Mbit/sec. However, if the network is expected to have many users, which can be determined from the IP class used for the network, a small β is preferred.

The User Behavior and Fairness:

In this section we consider the aspects of the user behavior in particular environments and try to perform a sensitivity analysis of the model parameters.

The end users have a tendency to increase their rates by $\alpha_i \beta_i / x_i$ anytime a positive acknowledgement is received. Clearly, this increase is additive. However, the decrease part is not necessarily multiplicative, rather, it is a function of the rate deviations relative to the capacities. The decrease part will occur in the event of having more arriving packets than the capacity of the outgoing links in a certain router. When a router is overwhelmed with the arriving packets, it will start marking the packets based on the penalty function presented in equation (10).

The receivers, in turn, will evaluate the congestion price and ultimately perform the decrease part of the control method where the estimated congestion depends on the severity of the congestion downstream.

Fairness between users is an important objective and should be always maintained. In this research, we consider the proportional fairness presented and Max-Min fairness between the users in case of similar utilities. If the users have an identical utility curves, all will have the same shares of bandwidth. Conversely, by varying the value α of the utility curves, the users will have different shares depending on how greedy their utility curves.

9. Effect of the parameters β, α, γ , η, and the Number of Users:

Since our algorithm falls in the category of increase/decrease methods, it will have stability design parameters. The presented examples showed the effect of β on the optimal solution. Higher β values guarantee closer rate *averages* to the optimal solution but they make the objective function (2) more steep around the optimal point at which small step size will lead to a significant jumps over the profile of the objective function. Therefore, for a numerical solution, smaller step size is essential to avoid ill-conditioning in the algorithm. However, small values of β will lead to longer time to reach the optimal solution.

The value of α informs the algorithm how much utility each user needs relative to the others. Similar values of a mean all the users compete fairly and will get similar bandwidth shares.

Generally, the step size γ affects the convergence time of the model. Smaller values of γ means underestimating the congestion charges and therefore the end users will deliver packets more than the capacity of the outgoing links of the routers. Higher packet losses are expected in case of small γ values.

With few differences, the parameter η has the same impact on the control law as the parameter γ. Higher

values of η will make the algorithm aggressive and will lead to a tooth-shape rate outputs. Increasing η to excessive value will bring the problem to the same traditional TCP/IP behavior. Usually, higher values of η are not preferred and will cause severe oscillations.

The number of network users affects how the congestion feedback is distributed between the data sessions. In our model and since we use a joint probabilistic marking, less synchronization will be observed. In other words, the end users will not increase/decrease their rates simultaneously at the same time.

When new transmission sessions are admitted into the network, the new data sources will affect the optimal rates that have been agreed on by the already existing users. This instable situation will create more marked packets and all the users will start tuning their rates again including the newly admitted sessions to a point where all will agree on a solution that provides fair shares in case of identical utilities. On the other hand, when some transmission sessions are terminated, after a certain delay the remaining data sources will detect the availability of more bandwidth. That is, when there are no marked packets received at the data sources, it signals for bandwidth availability and therefore the users will increase their rates until they receive newly marked packets where they start tuning their shares again.

10. Conclusions

In this paper, we presented an optimization model to data networks that implements the idea of an economic approach for maximizing the aggregate data sources' utilities, which was solved over the capacities set of constraints. The model showed its ability to handle larger network scales and required less global information.
The presented model can handle different objective functions. It also includes the use of packet marking and the loss rate as criteria for estimating the congestion. The model does not charge the network for single losses as compared to the traditional workhorse of the Internet (TCP), rather, it estimates the rate of the positive ACKs.

In addition, the model achieves a decentralized Max-Min fairness since it solves for fair shares of the bottlenecks in the network in a decentralized manner. Finally, different simulation examples were presented to demonstrate the model robustness in heterogeneous networks. Fairness assurance and decentralization of the control law along with the scalability of the scheme were one of the major conclusions in this study.

References:

[1] Jacobson V. 1988. Congestion Avoidance and Control, Computer Communication Review, Vol. 18, No. 4, pp. 314-329.

[2] Lakshman T.V. and Madhow U. 1997. The performance of TCP/IP for networks with high bandwidth-delay products and random loss, IEEE/ACM Trans. Networking, Vol. 5, pp. 336-350.

[3] Low S. H. 2003. A duality Model of TCP and Queue Management Algorithms, IEEE/ACM Transactions on Networking (TON) ,Vol. 11 , No. 4, pp. 525-536.

[4] Kelly F.P. 1997. Charging and rate control for elastic traffic, Eur. Trans. Telecommun., Vol. 8, pp. 33-37. Available:http://www.statslab.cam.ac.uk/~frank/ elastic.html.

[5] Kelly F.P., A. Maulloo, and D. Tan, 1998. Rate control for communication networks: Shadow prices, proportional fairness and stability, J. Oper. Res. Soc., Vol. 49, No. 3, pp. 237-252.

[6] Srikant R. and Kunniyur S. 2000. End-to-end congestion control: utility functions, random losses and ECN marks, IEEE/ACM Transactions on Networking. An earlier version appeared in INFOCOM'00, Tel Aviv, Israel.

[7] Low S. H. and D.E. Lapsley. 1999. Optimization Flow Control, I: Basic Algorithm and Convergence, IEEE/ACM Trans. Networking, Vol. 7, No. 6, pp. 861-874. Available: http://netlab.caltech.edu.

[8] Low S. H. 2000. A duality model of TCP flow controls, in Proceedings of ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management, September 18-20. http://netlab.caltech.edu.

[9] Low S. H. 2002.[b] Understanding Vegas: A duality Model, Journal of the ACM, Vol. 49, No. 2, pp 207-235. Available: http://netlab.caltech.edu/ pub.html.

[10] Pitsillides A., Sekercioglu A. 1999. Fuzzy logic based Congestion Control, COST 257: Impacts of new services on the architecture and network performance of broadband networks, Larnaca, Cyprus.

[11] Hollot C., Misra V., Towsley D., and Gong W. B. 2001. A control theoretic analysis of RED, in Proc IEEE Infocom, pp. 1510-1519, Anchorage, Alaska, April 22-26. Available at: http://www-net.cs.umass.edu/papers.html.

[12] Lachlan L. H. Andrew, Stephen V. Hanly and Rami G. Mukhtar, 2007. Active Queue Management For Fair Resource Allocation in Wireless Networks, IEEE Trans. Mobile Computing.

[13] Altman E., T. Basar, and R. Srikant, 1999. Congestion control as a stochastic control problem with action delays, Automatica.

[14] Athuraliya S., Low S.H., 2000. Optimization flow control with Newton like algorithm, J. Telcom. Syst, vol 15. no. 3/4. pp. 345-358.

[15] Dalalah, D., Cheng, L., Tonkay, G. 2008. Modeling end-to-end wireless lossy channels: A finite-state markov approach (2008) IEEE Transactions on Wireless Communications, 7 (4), pp. 1236-1243.

[16] Dalalah D., 2008. "Optimization Flow Control: A Queue Proportional Algorithm," in Proc. of IEEE International Symposium on Intelligent Control (ISIC'08), pp. 401-407.

[17] Dalalah D. 2010. " Real time optimization flow control" 2010. Computer Networks, Vol. 54, No. 5, pp. 797-810.

[18] Floyd Sally and Jacobson Van. 1993. Random early detection gateways for congestion avoidance, IEEE/ACM Transactions on Networking, Vol. 1, No. 4, pp. 397-413. Available: http://www.ece.ucsb.edu/Faculty/Madhow/Publications/ton97.ps

[19] Athuraliya S., Victor H. Li, Low Steven H. and Qinghe Yin. 2001. REM: active queue management, IEEE Network, Vol. 15 No. 3, pp. 48-53. Available: http://netlab.caltech.edu.

[20] Shenker S. 1995. Fundamental design issues for the future Internet, IEEE J. selected areas in communicant, Vol. 13, pp. 1176-1188.

[21] Mokhtar S. Bazaraa, Hanif D. Sherali, and Shetty C. M. "Noninear Programming Theory and Algorithms", 2nd Edition, Weiley Text Books, 1993.

[22] Floyd Sally. 1994. TCP and Explicit Congestion Notification, ACM Computer Communication Review, Vol. 24, No. 5, pp. 8-23.

SICSDA: an Adaptive Configurable Distributed Software Architecture Applied to Satellite Control Missions

[1]ADRIANA CURSINO THOMÉ
adriana.thome@dir.inpe.br
[1]MAURICIO G. V. FERREIRA
mauricio@ccs.inpe.br
[2]JOÃO BOSCO S. CUNHA
bosco@unifei.edu.br

[1] National Institute for Space Research (INPE)
Av. dos Astronautas, 1758 - São José dos Campos - SP - CEP 12227-010 – Brazil
[2] School of Engineering of Itajubá (UNIFEI) – Brazil
Caixa Postal 50 – CEP 37500-903 - Itajubá – MG – Brazil

***Abstract*-This paper proposes an adaptive configurable distributed software architecture applied to satellite control missions called SICSDA. The main purpose of this architecture is to control more than one satellite through one set of computers, enabling the choice of each satellite to be monitored in any given period of time. This architecture allows a new mission to be settled without the need for the creation and addition of a specific software component for the satellite being launched, thus minimizing the effort needed to adapt the complete system to the new requirement. It also provides domain specialists and software developers with the capability to configure, if necessary, attributes and business rules to the satellites already launched, adding new elements to business domain without the need of extra codification. The functionalities offered by the application, for example, telemetry visualization and the sending of telecommands, can be distributed into a network pre-defined domain. The system charge distribution service will define the objects location, what means that each machine in the network will be able to have a different view of the metadata stored in the database. A "view", in this context, is the piece of the adaptive object model that will be instantiated in that machine.**

I. Introduction

This work presents an adaptive configurable distributed software architecture applied to satellite control missions called SICSDA. This architecture was proposed as a Doctoral Thesis of Applied Computing Course at Brazilian National Institute for Space Research (INPE).

INPE, as one of the main organizations involved in the accomplishment of Brazilian´s space program, has taken the responsibility for producing and controlling the Brazilian satellites. In its initial stage, the Brazilian space program comprises the launching of four satellites, all of which have already been launched: SCD1 (Data Collector System 1) and SCD2 (Data Collector System 2), both data collector satellites; and CBERS1 (China Brazil Earth Research Satellite 1) and CBERS2 (China Brazil Earth Research Satellite 2), both remote sensing satellites.

INPE has created a robust infrastructure, aiming at the development of its capability for fully producing and controlling satellites, composed by the Satellite Control Center (CCS), by two remote tracking stations (Cuiabá and Alcântara Stations), by a communication network (RECDAS), and by a satellite control software application (SICS), as shown in Fig. 1. The remote stations are located at Cuiabá and Alcântara and they build together with CCS the ground support for the satellites in orbit. CCS is located in São José dos Campos and it is responsible for assuring the perfect operation of the satellite. RECDAS links CCS with Cuiabá and Alcântara stations.

A specific machine or a specific set of machines is used to run an application for collecting the data and for controlling the internal state of each satellite. The software application is uniquely developed to control each satellite from Earth. When a new satellite is launched, a new software application must be developed or adapted to that specific satellite, and a set of machines must be allocated to run this software, causing an extra development cost of hardware and software. This is necessary because each satellite has its own features that normally vary a little bit from one satellite to another. The access to these applications is restricted to satellite controllers physically located at CCS.

The problem stated above caused one to think about the creation of an architecture for the development of a more adaptive software system to control satellites of different kinds making use of the same machine or of the same set of machines. The requirements on the new system should include the capability for expanding it without many difficulties, for example in the case of adding features of a new satellite to be controlled, and the capability for making these additions without jeopardizing the quality of the overall system being modified.

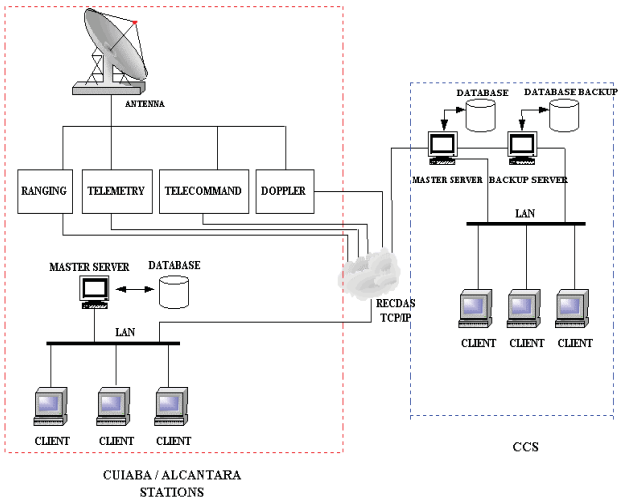


Fig.1- Satellites control system simplified architecture.
SOURCE: adapted from [3].

These requirements led further to the proposal for the creation of an architecture to be used for the development of applications that need to evolve as the problem domain evolves, based on the idea of building more configurable software components, which are flexible and adaptive, thus allowing the system to adapt itself to domain necessities more easily, that is, allowing the system to match requirements evolution while maintaining its quality.

According to the literature on the subject, one way for doing this is to move some aspects of the system, like business rules for example, to databases, making them easier to be modified. The resulting model allows systems to adapt themselves to new domain necessities through the modification of the values stored in the databases, instead of modifications of the code [1].

This encourages the development of software features that allow decision makers and managers to introduce new elements to existing software without the need of additional coding, and that these changes are introduced in the domain models at runtime, reducing significantly the time to incorporate new ideas to the software.

Architectures that can dynamically adapt themselves to new users requirements at runtime are called "reflexive architectures" or "meta-architectures. An "Adaptive Object Model Architecture" is a particular kind of reflexive architecture that holds object oriented systems that can be extended to incorporate new elements [1].

Thus, an adaptive object model is a system that represents classes, attributes, relationships, and behavior as metadata. The system is a model based on instances rather than classes. Users change the metadata (object model) to reflect changes in the domain. These changes can modify the system behavior. In other word, the system stores its object model in a database and interprets it. Consequently, the object model is adaptable; when the descriptive information is modified, the system immediately reflects those changes.

The design of adaptive object models differs from most object-oriented designs. Normally, object-oriented design would have classes for describing the different types of business entities and associates attributes and methods with them. The classes model the business, so a change in the business causes a change to the code, which leads to a new version of the application.

An adaptive object model does not model these business entities as classes. Rather they are modeled by descriptions (metadata) that are interpreted at runtime. Thus, whenever a business change is needed, these descriptions are changed which are then immediately reflected in the running application [2].

The use of an adaptive object model approach (AOMs) in the development of systems can ease some kinds of problems that can be found by software developers, mainly the problems related to system flexibility, evolution and maintenance; reducing drastically the total cost of software development [4].

II. SICSDA Architecture

The main purpose of this architecture is to control more than one satellite through one set of computers, enabling the choice of each satellite to be monitored in any given period of time.

Another important point is to have an architecture that allows a new mission to be settled without the need for the creation and addition of a specific software component for the satellite being launched, thus minimizing the effort needed to adapt the complete system to the new requirement

SICSDA architecture models the satellite control application based on adaptive object models. Therefore, in SICSDA architecture, problem domain objects, for example, telemetry, telecommand, ranging; instead of being located in the code that implements the application, will be implemented in an object oriented database to be interpreted and instantiated at runtime. It means that the system will have a generic code that will be able to handle the different satellite object models too.

SICSDA needs to be fault tolerant, so this architecture will be distributed, and the functionalities offered by the application; for example, telemetry visualization and the sending of telecommands, can be distributed into a network pre-defined domain. It means that application objects can be instantiated in different machines of the network, causing a distribution of the system code. The system charge distribution service will define these objects location, that is, in which machine of the network they will be available. It means that each machine in the network will be able to have a different view of the metadata stored in the database. A "view", in this context, is the piece of the adaptive object model that will be instantiated in that machine. Therefore, a given machine in the network will only have access to a part of the object model, and it will instantiate only the objects pre-defined and allocated to it by the system charge service.

In addition, SICSDA architecture suggests the use of distributed databases through the replication of the database for each node of the network. It is important to remark that it will be necessary to have mechanisms that help the control and the storage of metadata recording, and mechanisms that help the versions control.

SICSDA architecture will be configurable because it will be able to: **(1)** allow domain specialists and developers to re-configure the database to handle new classes, by creating these classes, their attributes and methods at runtime; and **(2)** allow, at runtime, the choice of which satellite metadata to use, causing a new object model instantiation in the generic code each time that this kind of context change is asked by the user, that is, each time the user wants to control a different satellite.

SICSDA architecture will be dynamic or adaptive because it will be able to handle possible changes in the business domain, allowing it to match the evolution of the business requirements and to adapt itself to the users necessities. Fig. 2 shows the structure of SICSDA.

This way, domain specialists (satellite controllers and satellite engineers) and software developers can adapt the system to arrange new classes, creating, at runtime, these classes and there attributes. Thus, we can say that SICSDA architecture is configurable considering business rules, since it is possible new business rules (or new methods) to be associated to a domain class at runtime. Following, the elements that appear in Fig. 2 are detailed:

- ***Satellite Control Application:*** contains the objects of the software that controls the satellites (telemetry, telecommand, ranging, etc).
- ***Persistence Service:*** is responsible for storing and retrieving the metadata from the database. Additionally, this service is responsible for storing and recovering users authentication data and objects configuration data (nodes where the architecture is installed) from a database called "Configuration Base.
- ***Configuration Service:*** belongs to the system presentation layer and is responsible for maintaining the metadata and the objects configuration in the nodes where the architecture is installed. Offers to domain specialists and software developers an appropriate interface to perform there activities.
- ***Adaptation Service:*** is responsible for providing the metamodel that allows the objects recovered from the database can be adequately inicialized and modified at runtime.
- ***Satellite Simulator:*** software that simulates the interaction with satellites.
- ***Charge Service:*** is responsible to give the initial charge in the system, in other words, to run the system charge routine in the nodes where the architecture is installed.
- *User Service:* belongs to the system presentation layer and it is responsible for offering to domain specialists an appropriate interface to visualize telemetry, to emit telecommands and to obtain distance and calibration measures for the desired satellite. Besides, this service is responsible for providing for the users the appropriate interface to login in the system.

Thus, as shown in Figure 2, the Charge Service charges the satellite metadata corresponding to the satellite that were chosen to be monitored in a given period of time. To do that the Charge Service activates the Persistence Service, that is responsible for recovering the metadata from the database. The Charge Service instantiates the objects recovered from the database, creates the objects of the chosen satellite application, and constructs, in this way, the real object model of the application.

The user can interact with the objects through the User Service. If the user wants to initiate the monitoring of a satellite, the User Service activates the Charge Service to charge the appropriate objects from the database. The data from satellites can be obtained through the Satellite Simulator.

The application objects will be distributed and can be accessed through middleware. The connection service will be responsible to locate an object in the system.

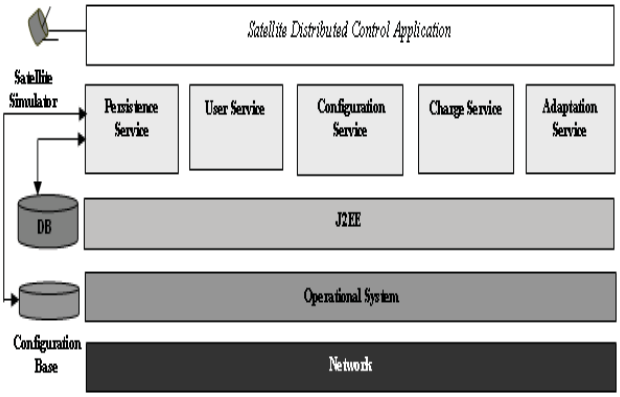


Fig.2- SICSDA architecture.

If the satellites metadata need to be changed to reflect changes in the domain, the user responsible for the configuration of the system can do these changes using an appropriated interface. When this kind of change occurs the Adaptation Service is activated. The Adaptaion Service, as a consequence, activates the Persistence Service, that will be responsible for reflecting in the database these changes.

III. Final Results

Relying in the class diagram that represents the business domain classes for SCD1, SCD2 e CBERS2 satellites, it was possible to obtain a generic class diagram that represents the Satellite Constrol System business domain metamodel. This metamodel was achieved after applying the following design patterns: pattern TypeObject, pattern Property, pattern Accountability e pattern Strategy. More details about these patterns can be obtained in [2] e [1].

After the generic class diagram achievement, the metamodel was oriented to fit the choosen implementation platform. As was established J2EE as the implementation platform, each class that appears in the generic class diagram was represented as an Entity Enterprise Java Bean.

The implementation of SICSDA architecture was composed of: Java as the programmation language, Caché as the database management system, JBOSS as the application server and Jbuilder as the development environment.

The metamodel was represented in database and the metadata to each satellite could be stored through a graphic interface offered by the Configuration Service. Through this interface, it is possible to change at runtime the classes and their attributes, and also to associate new methods to the classes.

The database was populated with satellites metadata through the Configuration Service interface. The metadata stored were based, as much as possible, in values approximated to values used in real control systems.

Users interact with the system to perform an operation, for example, “View Telemetry”, “Send Telecommand” or “Obtain Measures”, trhough the interface provided by the User Service. Thus, it is possible, depending on the satellite that is desired to control, ask for an operation to the system and recover the values associated to this operation. This operation is, actually, dynamically invoked, since its name and the information about its parameters and returned values are only known at runtime. In this manner, it is possible to associate methods already created to new kind of messages, what means that it is possible, in a point of view, to change the system behavior at runtime.

IV. Conclusion

This work proposes an adaptive configurable distributed architecture for the development of satellite control systems that are based on adaptive object models called SISDA.

One of the main advantages of using an architecture based on adaptive object models is, on the one hand, the ease to make changes in the system, since it allows developers and domain specialists to extend the system dynamically, that is, they can “code without coding”.

On the other hand, the possibility of creating new business rules without changing system code brings up some difficulties. One reason for this is that, to make it possible, it is necessary to build a specific domain language, what makes the developers inherit all the problems associated to the development of any language, like the construction of debuggers, and the provision of features to make system documentation and the control of versions.

Besides, it is necessary to build an interactive interface so that developers and domain specialists can create new rules at runtime, what requires tools and support for the use of graphical interfaces as well.

Furthermore, architectures based on adaptive object models require the construction of generic code. This generic code is usually difficult to be built since it is necessary to have a high level of abstraction.

Another important issue is the performance of the system. Since the adaptive object model needs to be interpreted to build the real object model and in order for it to reflect the changes made to the system, the system execution is affected accordingly, requiring adequate software and hardware support.

In addition, developers that are not familiarized with this kind of architecture find them more difficult to understand and to maintain than the traditional architectures, since two systems co-exist: the generic code written in a object oriented codification language and the AOM that is interpreted. On the other hand, developers familiarized with this kind of architecture believe that it makes the system easier to maintain in comparison with traditional architectures, since the systems have less code lines and a small change in it can typically cause a big change in the system being run.

Despite of the fact that many aspects point out that, at least in the beginning, architectures based on AOMs require more effort to be built than the traditional architectures, this work intends to take a significant step towards systems reusability, since adaptive systems have the property of matching business requirements evolution, and at the same time, they have a general-purpose architecture, what make them highly configurable and adaptable.

Another important step can be taken in the direction of systems maintainability. This is possible because the effort to make changes in the system can be minimized, since changes in the code can be minimized substantially. In addition, with this kind of architecture it is possible to allow domain specialists to make some changes in the system by themselves, increasing the maintainability and minimizing developer intervention in the system evolution.

Besides, one also hopes to take an important step towards economy, since future missions will be able to use all the hardware and software investment that has been made for previous missions.

The work proposed here has a multidisciplinary nature, since it is a binding of some features that have been explored independently in different research areas until now. Besides, this work is based upon relevant efforts made in the past for the creation of an architecture to model satellites control application based individually on the SOFTBOARD and on the SICSD architectures presented in [5] e [6], respectively.

The studies developed originated the thesis proposed in [7] and had been presented and/or published in [8], [9], [10], [11], [7], [12]; and originated the studies developed in [13] and [14].

This way, one hopes to be collaborating with to the advance of computational research in Brazil, and to be contributing to the success of the Brazilian space mission, by proposing a new alternative architecture for the development of satellites control software, and mainly, by proposing a new approach for the development of adaptive systems.

References

[1] YODER, J. W.; ET AL. Architecture and design of adaptive object-models. **ACM Sigplan Notices.** Vol. 36, Fasc. 12, pg. 50-60, December of 2001.

[2] JOHNSON, R.; YODER, J. W. The adaptive object-model architectural style. **IEEE/IFIP Conference on Software Architecture 2002 (WICSA3´02).** Canada. August of 2002.

[3] ROZENFELD, P.; ORLANDO, V.; FERREIRA, M.G.V. Applying the 21th century technology to the 20th century mission control. **SPACE OPS 2002.** USA. October of 2002.

[4] LEDECZI, A.; ET AL. Synthesis of self-adaptive software. **IEEE Areospace Conference Proceedings.** USA. V 4, pg. 501-507. 2000.

[5] CUNHA, J. B. S. **An Approch to Development of Complex Software Systems Using Softboard Architecture .** São José dos Campos, 1997. Applied Computing Phd Thesis. National Institute for Space Research (INPE).

[6] FERREIRA, M. G. V. **A dynamic and flexible architecture for distributed objects applied to satellites control software.** São José dos Campos, 2001. Applied Computing Phd Thesis. National Institute for Space Research (INPE).

[7] THOMÉ, A. C.; ET AL. SICSDA – **An adaptive configurable distributed software architecture applied to satellite control missions.** São José dos Campos, 2004. Applied Computing Phd Thesis. National Institute for Space Research (INPE).

[8] THOMÉ, A. C.; ET AL. SICSDA – An adaptive configurable distributed software architecture applied to satellite control missions. **XVII SBES – Brazilian Software Engineering Symposium - III Workshop of Software Engineering Thesis.** Manaus, Brazil. October of 2003.

[9] THOMÉ, A. C.; ET AL. An adaptive configurable distributed software architecture applied to satellite control missions (SICSDA).**III Workshop of Applied Computing Courses of INPE**. São José dos Campos, SP. November of 2003.

[10] THOMÉ, A. C.; ET AL. Establishing an adaptive configurable distributed software architecture applied to satellite control missions. **Eighth International Conference on Space Operations – SpaceOps 2004**. Montreal, Canadá. May of 2004.

[11] THOMÉ, A. C.; ET AL. SICSDA: an Adaptive Configurable Distributed Software Architecture Applied to Satellite Control Missions. **18th European Conference on Object-Oriented Programming - ECOOP 2004.** Oslo, Norway. June of 2004.

[12] THOMÉ, A. C.; ET AL (f). An adaptive configurable distributed software architecture applied to satellite control missions. **DaVinci Magazine.** Volume II, Number 1, pg. 117-132**.** ISSN 1807-6432. 2005.

[13] ALMEIDA, W. R. **An approach to persistence of models of adaptive and configurable systems.** São José dos Campos, 2004. Applied Computing Master Thesis. National Institute for Space Research (INPE).

[14] CARDOSO, P. E. **A new architecture for representing business rules i dynamic object models.** São José dos Campos, 2004. Applied Computing Master Thesis. National Institute for Space Research (INPE).

A Media Sharing Platform Built With Open Source Software

Karel Braeckman, Robbie De Sutter, Mike Matton, Tine Blomme
VRT-Medialab
B-1043, Brussel,
Belgium
E-mail: {karel.braeckman|robbie.desutter|mike.matton|tine.blomme}@vrt.be

***Abstract*— Sharing and handling media files in a professional context often requires expensive software packages. Within the EBU P/SCAIE project, a platform was required that could handle an abundance of professional file formats, arbitrary large file sizes and which did not pose restrictions on the metadata format used. As there was no such software available, we decided to build a custom web-based platform, based on loosely coupled open source components.**

This paper explains the architecture of the resulting platform. With a minimum of custom code, we have created a powerful platform that meets our requirements. This integration, described in the paper, is of use to organizations wishing to build their own media platform using open source components.

***Index Terms*—Media Repository, Open Source, Online Exchange Platform**

I. INTRODUCTION

Nowadays more and more automated information extraction tools from audiovisual material, such as face recognition, scene segmentation, and alike, are being developed by (academic) research groups. These tools are of interest to broadcasters, but unfortunately broadcasters are not always aware of their existence and possibilities. Vice versa, researchers do not know the particular functionality broadcasters desire and need from such tools. The main goal of the European Broadcasting Union (EBU) workgroup P/SCAIE [4] is to create a platform for broadcasters and researchers that allows them to exchange ideas and results with regard to automated information extraction tools. Whereas the TRECVID [10] initiative of NIST is more oriented towards military applications, the P/SCAIE initiative focuses on applications for broadcasters.

An objective of the P/SCAIE initiative was to construct and disclose an online media library which contains relevant and genuine broadcast material. As such, an online platform was created, allowing broadcasters to share and researchers to retrieve and download media material. Researchers can freely use the material (for research purposes) to test and evaluate their information extraction tools.

This paper describes the architecture, requirements and set-up of the online media exchange platform. A first version of the platform is available at http://media.ibbt.be. It was built by VRT-medialab, the research division of the Flemish public broadcaster, using open source components and some custom development.

This paper is organized as follows. First we will list the requirements of the media platform. In the following section, we discuss the different components used and their configuration. The next section elaborates upon the integration of these components in a consistent platform. Finally, we explain the ingest workflow and we end with conclusions.

II. REQUIREMENTS

The platform should meet the following requirements:

1. **Open Source components**. The components used are preferably available under an open source license. This ensures continued development, easy customization and absence of license fees.
2. **Loose coupling**. The different components of the architecture must not be tightly integrated. The final architecture must be extensible and configurable.
3. **No metadata restrictions**. The components used should not restrict the choice of a metadata model.
4. **No file size restrictions**. As audio and especially video files may have very large file sizes, the Upload Component should not pose any file size restrictions.
5. **In-browser experience**. The entire application should run inside the browser; the system and firewall of the user should not need any modification; the user interface can be simple, but should at least be able to show a preview video along with the metadata.

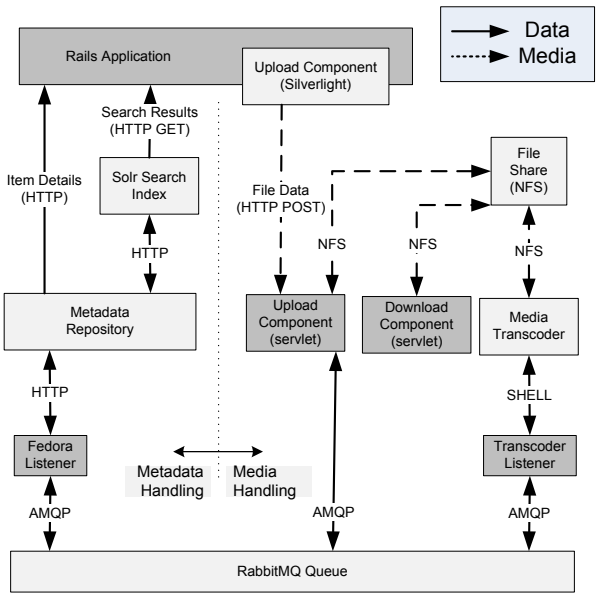


Fig. 1. Architecture Overview. The Blocks in White are the Existing Open Source Components.

TABLE I
OPEN SOURCE COMPONENTS USED IN THE PLATFORM

Component	Project	License
Upload (client side)	Silverlight Uploader *http://silverlightuploader.codeplex.com/*	Ms-PL
Metadata repository	Fedora Repository *http://www.fedora-commons.org*	Apache
Search Index	Apache Solr *http://lucene.apache.org/solr/*	Apache
Transcoder	FFmpeg *http://ffmpeg.org/*	GNU GPL
Message Queue	RabbitMQ *http://www.rabbitmq.com/*	Mozilla PL
Ruby-Fedora Integration	RubyFedora *http://mediashelf.us/rubyfedora*	GNU GPL
Solr – Rails Integration	solr-ruby *http://wiki.apache.org/solr/solr-ruby*	Apache

III. COMPONENTS

An overview of the platform architecture is shown in Fig. 1. The blocks in white represent existing open source components, while the blocks in gray are components that were custom-developed. Each of these components will be discussed in this section, while the integration between the components will be discussed in the next section. An overview of the open source components used, their project page and license can be found in Table I.

A. Components for Media Handling

The right hand side of Fig. 1 contains the components that handle the media files. The Upload Component allows the user to upload a media file from their computer to the online platform, while the Download Component is used to download media files. The Transcoder creates thumbnails and converts the uploaded media files to a format that is playable in the browser.

Upload Component

Due to requirements (4) and (5), the upload component proved to be quite a challenge to implement. Two protocols are often used to transfer files over the Internet: FTP and HTTP. The FTP protocol allows arbitrarily large file sizes, as required by (4). However, FTP requires opening certain firewall ports, which violates requirement (5). In contrast, HTTP traffic is allowed in almost any firewall configuration; otherwise the user would not be able to browse the web. This means our upload component should use the HTTP protocol to satisfy (5).

Furthermore, current browsers do not support uploading files with a file size larger than 2GB [6]. Therefore, we decided to split the media file in chunks smaller than 2GB, upload these chunks one by one and re-assemble them on the server (Fig. 2).

The technology choice for doing this kind of file manipulation in the browser is rather limited because of security restrictions. A regular JavaScript is unable to manipulate the local filesystem of the client. An Adobe Flash component on the other hand has access to the local file system (Flash Player 10 and above), but lacks the ability to read a portion of the file. This means the entire file is loaded in memory which is a problem for large files.

We chose to use an open source Microsoft Silverlight component. Similar to Flash, Silverlight allows access to the local file system, plus it is able to access a file byte per byte. This enabled us to read and upload the media file chunk per chunk. As shown in Fig. 2, each chunk is sent using a regular HTTP POST request. Besides the file data, this request

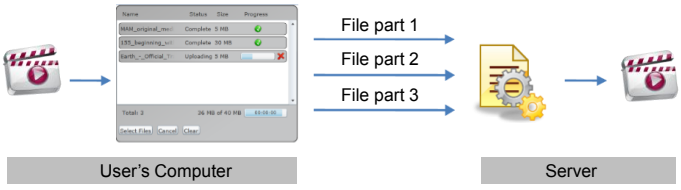


Fig. 2. The Upload Component splits the file up in several HTTP POST Messages, which are Reassembled on the Server by a Java servlet.

To obtain minimal overhead, the Java servlet will immediately stream the file to its final location, instead of first downloading it to a temporary file and copying it afterwards.

Media Transcoder

Once an item is uploaded to the media platform, a thumbnail and an in-browser playable proxy are created. The FFmpeg project is used to transcode the original material, because it supports a wide variety of media file formats (including professional formats). A keyframe, which will be shown in the web application, is extracted from the video. In addition, the video is converted to a low-res (qvga, 320x240) video in the Windows Media Video 8 format. The application uses JW Silverlight Player by LongTail Video to play this video in the browser. The Silverlight player was preferred over JW Flash Player, because the Upload Component also uses the Silverlight plug-in.

Download Component

A simple Java servlet is used to download files. This component allows users to start downloading a file while it is still being uploaded by another user. In future versions we plan to start the transcoding as soon as the upload starts. In combination with the download component, this will allow a user to start downloading the transcoded file while another user is still uploading the file, thus dramatically decreasing the processing time.

B. *Components for Metadata Handling*

The left hand side of Fig. 1 contains the components used to handle the metadata related to the media files. The Metadata Repository stores the metadata. The Search Index contains part of the metadata, structured in a manner optimized to support search.

Metadata Repository

We used the Fedora Repository Project (FRP) as the repository for storing the metadata of the media files. FRP does not impose a metadata structure – any XML is allowed, satisfying requirement (3) – which allows the user to choose a metadata standard that fits the needs of the specific media collection. FRP also supports advanced features such as semantic relations between items using RDF, and metadata versioning, which will be exploited in further iterations of the platform.

A Fedora Object is composed of several so-called Datastreams. We use two types of Datastreams in our platform: Media and Metadata Datastreams. A Media Datastream contains a URL that refers to a media file. A Metadata Datastream contains XML metadata about the media item. A Fedora digital object can have one or more Datastreams of each type. At a minimum, a Metadata Datastream with Dublin Core metadata is required by Fedora. Fedora stores its digital objects in a simple XML format called FOXML. An in-depth explanation of the Fedora Object model is available in [3].

Search Index

FRP has built-in search functionality, but it only indexes the Dublin Core Metadata Datastream. As we want to search through any metadata related to an item, we decided to set up Apache Solr [8] as an external search index.

Solr can index documents with several fields. A Solr document is represented by a simple XML file (SOLR XML format) which defines the values of these fields. HTTP is used to build and query the index. A Solr document is added to Solr using an HTTP POST request. The index can be queried with an HTTP GET request, with the query parameter in the URL. More information on Solr and building a custom video search application on top of it can be found in [2].

C. *User Interface*

The user interface was built with the Ruby on Rails (RoR) framework[1]. Fig. 3 shows the search results page, which contains a thumbnail and a short description per result. The top right corner contains a search box that allows searching for particular media files. The left side of the page contains filters which can be used to narrow down the search result set.

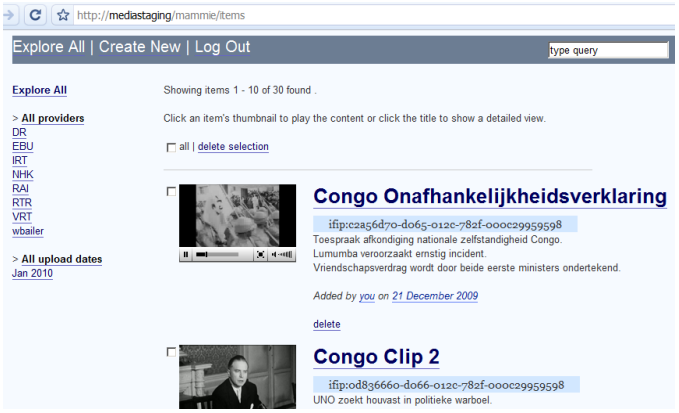


Fig. 3. Overview Page of the Interface.



Fig. 4. Item Details with a Preview Video, Metadata and Media Datastreams.

When the user selects an item, a detailed view is opened (Fig. 4) in which the low-res video or audio can be played and

[1] http://rubyonrails.org/

details about the Datastreams are shown. If the user owns the item, he can also add new metadata or media datastreams to the item. The Dublin Core metadata can be also be edited on this page.

IV. INTEGRATIONS

This framework contains a lot of components which have to cooperate. The integrations between the diverse components are discussed in this section.

A. Solr - Fedora

To transfer the metadata from the Fedora Repository into the Solr index, we used a component called the Fedora Generic Search Service (GSearch). This component, which is available via the Fedora Repository, allows connecting Fedora with various external search indexes, including Apache Solr.
The configuration of GSearch is performed using an XSL transformation that transforms a Fedora digital object (FOXML format) to a Solr document. This XSL transformation should be adapted to the specific metadata format used in the application, in order to ensure the correct indexing of all metadata. When Fedora updates an object, it calls GSearch, which in turn transforms the object to a Solr record and adds it to the index.

B. Rails - Solr

We applied the existing Ruby library solr-ruby to connect the Rails application to the Solr index. This library transforms the Solr search results into Ruby objects which causes the integration to be quite painless.
Solr stores the description and title of every item. This information was used to construct the overview / search results page (Fig. 3). Since no Fedora calls are necessary to show this page, the load on Fedora is reduced and the speed of the user interface is increased.

C. Rails - Fedora

Rails is organized around the Model, View, Controller architecture (MVC). In Rails, a model usually corresponds to a table in the database. Our application is, however, not based on a traditional relational database. Instead, the Fedora Repository was used to store the data related to our media items. For this reason, the mapping from a Fedora item to a Rails model had to be programmed, instead of relying on the existing Rails mappings between models and relational databases.
The communication between Rails and Fedora uses REST [5] and was implemented using an existing Ruby library called ruby-fedora. The interface of our final model to a large extent matches the conventional Rails Resource Model, which allowed us to employ the default configurations of Rails.

D. Rails – Upload Component

Integrating the Upload Component in the Rails interface was done using a simple HTML <embed> tag. A security issue complicated the integration though. Only a user that is logged in, is allowed to upload files to (or download files from) the repository. When a user logs in, the Rails application generates a session id, which is stored on the client in a Cookie. The browser attaches the session id with every future request it makes. The Rails application can use this parameter to verify that the request comes from a valid user.
As the Silverlight component posts its data directly to the Java servlet instead of to the Rails application, the Java servlet should verify that the user is logged in. Otherwise a malicious user could post potentially dangerous files directly to the Java servlet without being logged in.

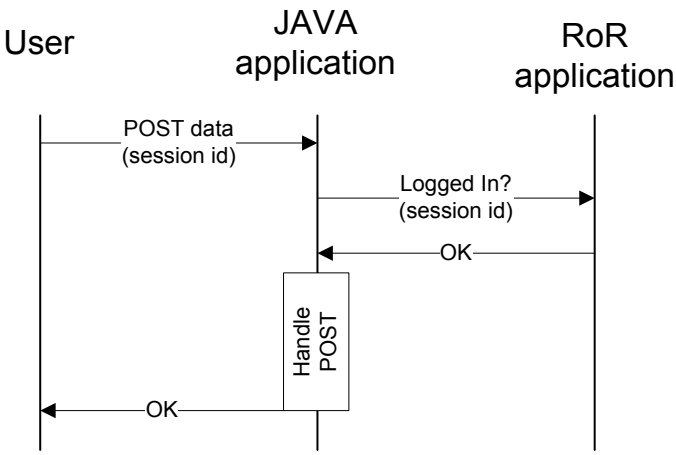


Fig. 5. The JAVA Application Authenticates the User with an Extra Request to the RoR Application.

Fig. 5 explains how the verification is accomplished. Assume the user is logged in and tries to upload a file using the Silverlight component. The Silverlight component will perform an HTTP POST to the Java servlet. Because the user is logged in, the browser will attach the session id parameter to the request. In the Java application, a new request to the RoR application, containing the session id parameter, is created. This request is sent to a controller that replies with a simple YES or NO depending on whether a session with this session id exists or not. If the Java application receives a YES it continues handling the request, otherwise it will fail.
Note that checking the session is transparent for the user (no code on the client needs to change). To separate the session logic from the file handling logic in the servlet, we have put the session logic in a servlet filter [9]. This makes it possible to attach the session logic to any servlet by simply changing the web.xml configuration file of the Java web application.

E. Fedora – FFmpeg – Upload Servlet

We have chosen a loosely coupled architecture using a message queue for the integration between Fedora, FFmpeg and the Upload Component. As will be shown in the following

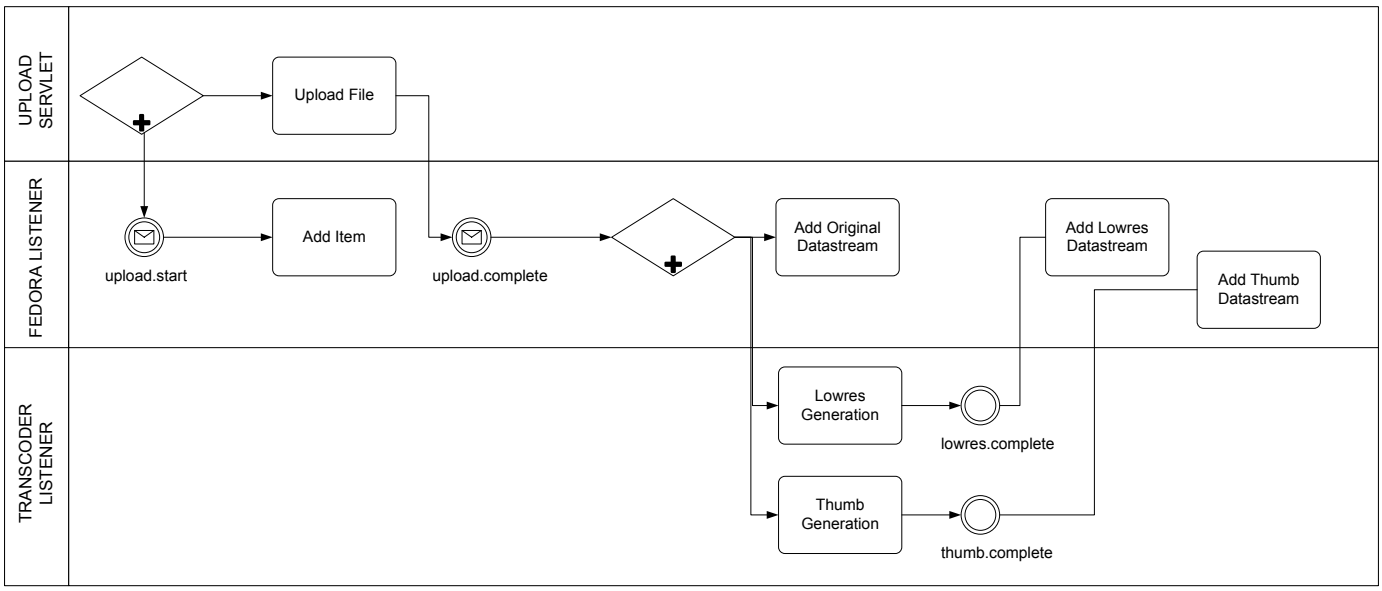


Fig. 6. The Upload Workflow is defined by the message configuration.

paragraphs, this renders our architecture highly configurable and extensible, satisfying requirement (2).

The Advanced Message Queuing Protocol [1], [11] was chosen as communication standard between the different components. In contrast to other messaging middleware such as JMS, which merely defines an API, AMQP describes the messages sent across the network as a stream of octets. Consequently, any tool that can create and interpret these messages can interoperate with any AMQP message broker, irrespective of implementation language. We employ RabbitMQ, an open source AMQP message broker, in the platform.

In our platform we use the 'topic exchange' concept defined in AMQP. Every message has a routing key (e.g. 'upload.start') and a queue can subscribe to all the messages of which the routing key matches a specific pattern (e.g. 'upload.*').

We wrote a small framework in Ruby which facilitates connecting a component to the queue. Every component will have an associated 'Listener', which is implemented as a regular Ruby class, that bridges the gap between the component and the queue. The methods of the Listener class are triggered by certain messages on the queue.

A configuration file is used to define the desired effect of messages. Upon initialization, our framework will load this configuration file and configure RabbitMQ accordingly. An example of the configuration file:

```
Listener1:
  ruby_class: FedoraListener
  mapping:
    uploadStarted: upload.start
    addDatastream: *.complete
        ...
```

This configuration file defines only one Listener. The configuration file will bind the uploadStarted method of the Listener FedoraListener to messages with a routing key 'upload.start'. The body of the message – a hash of parameters containing, for instance, the filename – is passed as a parameter to this method. The implementation of the uploadStarted method will create a new object in Fedora Repository. The next line of the configuration file uses a routing key with a wildcard. Any message with a routing key ending with '.complete' – be it 'transcode.complete' or 'upload.complete' – will trigger the addDatastream method.

As the configuration of the queue, based on the configuration file, is done outside of the Listener class, there is a clean separation between Listener logic and queue logic. Table II briefly describes the Listeners we have developed.

It is clear that the system is highly extensible, as was required by (2). Adding new functionality only requires a regular Ruby class that implements the desired functionality, and a change of the configuration file which couples the new functionality to the queue.

TABLE II
BRIEF DESCRIPTION OF IMPLEMENTED LISTENERS

Listener	Description
FedoraListener	Class that contains methods for creating Items and Datastreams in Fedora. It uses the ruby-fedora library to communicate with Fedora using its REST interface.
TranscoderListener	Class that contains methods for starting the generation of low-res versions and thumbnails. It calls FFMpeg via the command line with the correct encoding parameters.
ErrorListener	Class that contains methods for creating error reports. These reports can be mailed to the platform administrator.

V. INGEST WORKFLOW

In this section we explain the ingest workflow that is currently implemented in the platform. The configuration of the messages sent by a component and their effect influences the tasks that are executed. We will explain this with the upload workflow which is shown in Fig. 6.

When the user clicks the upload button of the Silverlight component, this component starts uploading the file chunk per chunk to a Java servlet, as explained earlier. Upon receiving the first chunk, the Java servlet sends a message with a routing key 'upload.start' to the queue. This message informs other interested components that a new upload has started. The body of the message contains other information such as the user id, the filename, etc.

Because we configured FedoraListener to listen to messages with routing key 'upload.start', it will execute the Ruby method that creates a new item in Fedora. This item does not have any media attached to it yet, because the upload has not finished.

After receiving the last chunk of the upload, the Java servlet sends a message with 'upload.complete' to the queue. Both FedoraListener and TranscoderListener are triggered by this message. FedoraListener adds a Datastream with a URL linking to the uploaded file to the previously created item, while TranscoderListener starts two transcoding tasks that will create the low-res version and a thumbnail of the video.

Once the low-res version (or the thumbnail) is generated, a message is sent to the queue that triggers FedoraListener to add a Datastream linking to the low-res version (or to the thumbnail).

Other workflows can be implemented in the same manner, making it easy to extend the functionality of the system. A drawback of this approach is that it quickly becomes difficult to gain a clear view of the effect a message has. This could be solved by developing a tool in which the user defines the user interface in a diagram (such as Fig. 6) that is automatically converted to the queue configuration file we saw earlier.

VI. CONCLUSIONS AND FUTURE WORK

Within this paper we presented a platform, largely based on open source components, that allows the handling of professional media files with arbitrary large file sizes. The system is agnostic regarding media and metadata formats.

Thanks to the usage of a message queue, our architecture is highly extensible and loosely coupled. We have demonstrated a simple way to implement reconfigurable workflows.

Future work will probably include collapsing the upload-transcoding pipeline, as described briefly in part A. A streaming server for playing out the low-res media is also desirable, because this will improve the user experience and the network throughput. As the work done by the EBU P/SCAIE workgroup evolves, the user interface of the platform will also have to incorporate new features. For instance, the interface will have to support multiple collections of media objects, which will allow to make sets of media objects that are useful for evaluating a certain tool (e.g., news broadcasts contain a large amount of spoken speech and are thus useful for evaluating speech recognition tools).

A first version of the platform is available at http://media.ibbt.be. It is being used by the EBU P/SCAIE workgroup to enable the exchange of professional mediamaterial between broadcasters and research groups, such that the latter can use genuine material to test their automated information extraction tools. The platform is being further developed in the context of the CELTIC project MediaMap (http://www.mediamapproject.org/).

VII. REFERENCES

[1] AMQP Working Group. (2009, Dec.). *AMQP Protocol Specification.* Available: http://www.amqp.org/confluence/display/AMQP/AMQP+Specification

[2] K. Braeckman, R. De Sutter and M. Verwaest, "Building Custom Video Search upon an Existing MAM System" in *Proc. of the Fifth IASTED European Conf. on Internet and Multimedia Systems and Applications*, Cambridge U.K., 2009, pp. 671-677.

[3] C. Lagoze, S. Payette, E. Shin, C. Wilper, "Fedora: An Architecture for Complex Objects and their Relationships," *Int. Journal on Digital Libraries*, vol.6, no. 2, pp. 124-138, Feb. 2006.

[4] EBU Technical. (2007). *P/Scaie Content Analysis-Based Automatic Information Extraction*. Available: http://tech.ebu.ch/groups/pscaie

[5] R.T. Fielding, and R.N. Taylor, "Principled Design of the Modern Web Architecture", *ACM Transactions on Internet Technology (TOIT)* , vol. 2, no. 2, pp. 115-150, May 2002.

[6] *2GB upload limit for http uploads*. Available: http://www.motobit.com/help/scptutl/pa98.htm

[7] L. Richardson and S. Ruby, *RESTful Web Services.* Sebastopol: O'Reilly, 2007.

[8] D. Smiley and E. Pugh, *Solr 1.4 Enterprise Search Server.* Birmingham: Packt Publishing, 2009.

[9] Sun Microsystems. (n.d.). *The Essentials of Filters*. Available: http://java.sun.com/products/servlet/Filters.html

[10] TRECVID. *TREC Video Retrieval Evaluation*. Available: http://www-nlpir.nist.gov/projects/trecvid

[11] S. Vinoski, "Advanced Message Queuing Protocol", *IEEE Internet Computing*, vol. 10, no. 6, pp. 87-89, Nov. 2006.

[12] MediaMap project home page (2010). Available: http://www.mediamapproject.org/

Electronic Payment and Online Gaming

Gerald Draxler, Johannes Sametinger, Andreas Wiesauer

Johannes Kepler University Linz, A-4040 Linz, Austria
draxler@incert.at, johannes.sametinger@jku.at, andreas.wiesauer@jku.at

Abstract. We will evaluate electronic payment (ePayment) systems by employing a use-value analysis. The key success factors of ePayment systems are security and flexibility. Not surprisingly, it turns out, that there is neither a "best" nor a "most secure" ePayment system. The adequacy of these systems depends on the application context. A use-value analysis is an appropriate and easy to use evaluation method, because it allows the consideration of different application perspectives. In fact, many ePayment systems are available today, but there are still contexts that require a tailored solution. Online gaming will be given as an example. For this purpose, we will introduce *BetMPay*, an ePayment system that offers a high level of anonymity, payment guarantee for providers, as well as consumer protection. This system also suffers from drawbacks in comparison to other existing system. This again will be outlined by a use-value analysis.

1 Introduction

Today, transactions on goods, e.g., books or electronic devices are increasingly carried out over the Internet. Customers and merchants regularly face the problem of handling payments. Various payment systems are available that rely on different paradigms, e.g., credit cards, debit notes or payment via cell phones. Online store operators often have difficulties in selecting the appropriate systems for their needs. Criteria include usability, user acceptance or common usage in existing web shops. Many electronic payment systems are available and their use is constantly rising. But there are scenarios, where special requirements exist that are not sufficiently fulfilled by available systems. For example, online gaming clients often prefer to indulge their passion anonymously.

In this paper we present an overview on different electronic payment systems and evaluate their benefits from various perspectives. Additionally, we present a solution that had been implemented with the focus at customer anonymity. This system had been developed in the context of online betting applications and additionally aims at providing effective cost control for customers as well as paying guarantees.

In Section 2 we shortly introduce online gaming. Section 3 describes aspects of electronic payment and introduces a selection of ePayment systems. In Section 4 we focus on requirements that are important for any electronic payment system. We also depict criteria that can be used for evaluation. Section 5 provides an evaluation of ePayment systems by employing a use-value analysis. In Section 6 we introduce an ePayment system that is focused on customer anonymity, and we also compare it to the other systems. Related work follows in Section 7, and a conclusion is given in Section 8.

2 Online Gaming

In this context, "gaming" means the playing of games for something of material value like money. Games focus on an event with an uncertain outcome and the intent of winning additional material value [17]. The outcome of the game is typically evident within a short period of time, e.g., the final score of sports events. Online games are played for electronic money and require electronic payment systems.

An example provider of online gaming is *bwin* who is offering up to 30,000 bets daily with betting action in more than 90 sports [www.bwin.com]. Bwin offers a wide range of payment methods, including credit cards and online banking deposits. Reliable transfer of money is crucial. Above all, gaming providers need to be sure that money of their clients can actually be collected.

3 ePayment

Electronic payment or ePayment is the transfer of electronic means of payment from the payer to the payee through the use of an electronic payment instrument [11].

ePayment systems are used to transfer money from one account to another at the same or another financial institution. ePayment is an important part of eCommerce, as goods and services offered through the Internet are most conveniently paid in electronic form.

Several factors influence payment over the Internet [8]. For example, the Internet does not have an established security architecture. Both seller and buyer are not physically present in an online transaction. Goods are available only as virtual representations. And there is no synchronization between payment and delivery of goods.

3.1 Classification

Numerous ePayment systems are on the market. They can be classified based on several categories, see [1] and [11]. ePayment systems can be divided into electronic cash mechanisms and credit-debit systems. Electronic cash resembles conventional cash and is based on tokens. Electronic tokens represent value and are exchanged for payment.

Credit-debit systems are based on bank accounts. In credit-debit systems, money is represented by records in bank accounts. Payment information is sent over computer networks, e.g., the Internet. Electronic cash has several advantages like the potential for anonymity and the lack of the need to contact a central system. In *pre-paid systems*, the payer's account is debited before the payment. Pay-now systems debit the payer's account at the time of payment. In *post-pay systems*, the account of the payee is credited before the account of the payer is debited.

Another distinctive feature is *micro payment ability*. Micro payments typically amount to a value of less than 1 Euro or 1 Dollar. In contrast, macro payments start at an amount of 10 Euros/Dollars and small payments are in between. There can also be a limit on the amount of money that can be paid, e.g., the amount that had been prepaid or the credit card limit.

3.2 Available Systems

Too many ePayment systems are available to provide an exhaustive list. We have chosen a few that are widely available. Not all systems are offered world-wide; we have also included systems available in Austria, the home country of the authors. Additional ePayment providers and services can be found, for example, in [4].

Credit cards entitle their holders to buy goods and services based on the holders' promise to pay for these [5]. Credit card providers offer various levels of security in order to prevent fraud, e.g., the card security code. *Verified by Visa* and *MasterCard SecureCode* depict additional security measures for online transactions. When credit card information is entered online, the user has to authenticate herself, i.e., to confirm her identity with an additional password. *Debit cards* are used like credit cards for telephone and Internet purchases or like ATM cards for money withdrawal. In both situations, funds are immediately transferred from the holders' bank account. This is in contrast to credit cards, where users have to pay back on a later date. Maestro is an example for a widely known debit card service [www.maestrocard.com].

PayPal is an e-commerce business that allows worldwide payments and money transfers over the Internet [www.paypal.com]. *Paybox* offers payments via users' mobile phones in Austria. A user simply provides her phone number or, for anonymity, an alias number, gets an SMS (simple text message) from PayBox and confirms the payment [www.paybox.at]. *PaySafeCard* is a pre-paid system primarily for online shopping. It allows online cash payments without a bank account or credit card [www.paysafecard.com].

4 ePayment Requirements

Concerns on data security and on the misuse of private data are important factors for electronic payments. These concerns alone can hinder the development of e-commerce [15]. Security requirements will depend on the amount of money being transferred, i.e., macro payments require higher security than micro payments. Smaller risks and cost considerations lead to the acceptance of less security. However, concerns about security are key factors in discouraging consumers from online payments [9].

Security comprises integrity, availability, and confidentiality [8]. For online gaming, anonymity is an additional important security aspect. ePayment providers have to make sure that all these facets are considered sufficiently, including e.g., software security, network security and organizational security. The payment card industry data security standard helps organizations of payment cards to prevent fraud [6]. This can be achieved through increased data controls and decreased exposure of data to potential compromise. Evaluating the security of systems is a difficult and time-consuming task. In order to make this task manageable for our purpose, we will use the following properties that can be assessed with reasonable effort.

Passwords. How many characters must users enter in order to enter the system? We use a qualitative measure – high if a minimum length of eight characters and compliance to defined rules, e.g., use of special characters, are required; medium if passwords are used but do not comply with rules and low if no passwords/PINs are employed.

Password renewal. Password renewal defines whether users are asked to change their passwords periodically.

Login brute-force. Login brute-force prevention addresses effective measures for preventing brute-force attacks, see [12].

Certificates and SSL. Certificates and the use of SSL/TLS address the security of the communication channel, e.g., whether communication is encrypted and authenticity of the counterpart can be determined.

Authentication. We again use a qualitative measure – two-factor authentication, one factor authentication and authentication that is solely based on information that cannot be considered as sufficiently secret, e.g., account information or credit card numbers. Two factor authentication means that authentication is based on knowledge, e.g., passwords, and the possession of artifacts like cell phones. As access or transactions codes are sent to these cell phones, their possession is crucial for payments. One factor authentication is based on knowledge or possession.

Lock-out. Systems may lock out a certain user upon request. A lock-out can be initiated by users whenever they realize that authentication information had been lost or stolen in order to prevent malicious transactions.

Anonymity. We differentiate between information that is easily linked to identities like credit card numbers and information that is more difficult to use to reveal identities, e.g., phone numbers. Thus, we use yes/no values.

Additional security features address the fact whether critical transactions are secured by certain measures, e.g., use of transaction authentication numbers (TAN). Additional non-security requirements include micropayments, guaranteed payments, cross-border payments, offline payments and market penetration.

5 Analysis

As ePayment systems rely on different paradigms, they are difficult to compare. Our analysis does not aim at determining the best or most secure system. Instead, we want to provide a comprehensive overview on functionalities and security aspects of different systems.

5.1 Methodology

An evaluation determines the value or usefulness of a solution with respects to given objectives. Use-value analysis does not only cover quantitative but also qualitative criteria [7]. It involves the following steps [2]:

- Identification of evaluation criteria
- Assessing values of criteria for each system (quantitative or qualitative)
- Quantification of qualitative values (scaling)
- Weighting of criteria depending on importance

For the analysis, we use criteria outlined in section 4 together with requirements concerning system flexibility, e.g. micro payment or cross border payment ability.

5.2 Results

Table 1 shows the parameter values of the different systems. The values were retrieved by literature analysis, statistics, interviews of ePayment system providers and self-experimentation. We have to quantify our qualitative measures. We rank "high" by 3, "medium" by 2 and "low" by 1 and further "yes" by 3 and "no" by 1. Therefore, systems with higher use-values will be preferable. We can compare the systems based on these values. The adequacy of the values depends on the point of view, e.g., seller or buyer. We can reflect this by assigning different weightings to the chosen criteria. These weightings are calculated as follows: We use a base of 100. Since we have 14 criteria, we get an average weighting of 7 (100/14 rounded). We almost triple the average weighting and use 20 as a maximum weighting factor for eminently significant criteria. Since less important criteria shall not be completely disregarded, we use 3 as minimum weighting factor. Different points of views can be reflected by assigning different weightings in the range of 3 to 20.

A user study has shown that it is important from the perspective of customers to guarantee anonymity, ease of use and security. Furthermore, customers prefer systems that they already use or that they can use in other contexts too, i.e., flexible systems with high market penetration [3]. Given these considerations, Table 2 shows the results of our use-value analysis from a customer perspective. *Click & Buy* emerges as the preferable system. In contrast, providers demand payment guarantees as well as high market penetration in order to prevent customer distraction [10]. *Click & Buy* is also preferable from a provider perspective. This table is not shown due to space limitations.

An advantage of such use-value analyses is that they are easy to adjust according to requirements in a given context. For example, if it is central to provide high flexibility, we can adjust the weightings of micropayment ability, cross-border payment ability and offline payment ability and conclude that Paybox is our first choice.

6 BetMPay

Although there are many ePayment systems available, there can be application contexts in which requirements cannot be fulfilled by existing systems. In this section, we will describe *BetMPay* – a system that has been developed to support special needs of online gaming companies. These special needs are anonymity, prevention of customer addiction and a high level of payment guarantees.

Table 1: Values of Payment System Criteria

Criteria/ System	Credit Card using SSL	Verified by VISA	Paypal	Paysafecard	Paybox	Click and Buy	Debit Card
Password Strength	Low	Low	Med	Med	Med	High	Low
Password Renewal	No	No	No	No	No	Yes	No
Login Brute Force	No	No	Yes	No	Yes	Yes	No
Certificates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SSL/TLS	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Authentica-tion Method	Low	Med	Med	High	High	Med	Med
User Lock-out	Yes	Yes	No	No	Yes	Yes	No
Additional Security	No	No	No	Yes	Yes	No	Yes
Market Penetration	High	Low	High	Med	Med	High	Low
Anonymity	No	No	No	Yes	Yes	Yes	Yes
Micro-payment	No	No	Yes	Yes	Yes	Yes	Yes
Payment Guarantee	Low	High	High	Low	Med	High	Low
Cross border Payment	Yes	Yes	Yes	Yes	Yes	Yes	No
Offline Payments	Yes	Yes	No	No	Yes	No	Yes

Table 2: Use-values from a Customer Perspective

Criteria/ System	Credit Card using SSL	Verified by VISA	Paypal	Paysafecard	Paybox	Click and Buy	Debit Card	Weighting
Password Strength	3	3	6	6	6	9	3	3
Password Renewal	3	3	3	3	3	9	3	3
Login Brute Force	6	6	18	6	18	18	6	6
Certificates	15	15	15	15	15	15	15	5
SSL/TLS	15	15	15	15	15	15	15	5
Authentication Method	10	20	20	30	30	20	20	10
User Lock-out	15	15	5	5	15	15	5	5
Additional Security	5	5	5	15	15	5	15	5
Market Penetration	30	10	30	20	20	30	30	10
Anonymity	15	15	15	45	45	45	45	15
Micro-payment	10	10	30	30	30	30	30	10
Payment Guarantee	3	9	9	9	3	6	9	3
Cross border Payment	30	30	30	30	30	10	10	10
Offline Payments	30	30	10	10	30	30	10	10
Use-value	**190**	**186**	**211**	**239**	**275**	**257**	**216**	

6.1 Context

Online gaming is a sensitive area, since it may impose dangers on customers, e.g., getting addicted. We can therefore identify three main requirements: At first, customers need anonymity. They may not want that others know about their activities. Furthermore, we need a mechanism that helps to prevent addiction. Finally, providers demand payment guarantees. They may regularly get in contact with (addicted) indebted or insolvent customers. *BetMPay* supports all three major requirements.

6.2 Scenario

Anonymity is one critical success factor of ePayment systems. *BetMPay* operates according to the following paradigm: Customers can buy vouchers for bets (offline) in gaming offices, kiosks or petrol stations. These vouchers represent a fixed value which can be used for gaming activities on online platforms. For this purpose, the customer has to register at a web application when she uses the payment system for the first time. The only information needed for registration is a valid cell phone number.

The system does not store any names, addresses or bank account data. If buyers do not disclose their cell phone number, e.g., by enlisting in phone directories, their identity cannot be determined using common methods.

After registration, the customer can enter a 16-character code printed on the voucher. The amount of the voucher will be credited to her account and can be used for gaming activities. Whenever customers win a game or decide to withdraw money credited to their account, they can create a payment ticket. The ticket consists of a security code which is sent to the customer via SMS. Afterwards she may go to gaming offices or kiosks, where the payment ticket is validated by cell phone number and the security code. The amount will get cashed out.

Together with a high level of anonymity, this paradigm further supports the other major requirements. On the one

hand, it provides better cost control for customers. As many customers of online gaming companies are compulsive gamblers, the system requires them to buy additional vouchers offline. This is not as comfortable as transferring money from bank accounts by means of credit cards, but it gives gamblers a chance to rethink their decision ("cool down phase") and can prevent them from making overhasty investments. On the other hand, the system provides guarantees for gaming companies: Customers can only bet with payments already made (pre-payment). As compulsive gamblers are often heavily indebted or even insolvent, companies do not have to fear losses or costly legal processes in order to recover debts.

6.3 System Architecture

The architecture of *BetMPay* is conceived in a classical three layer approach**.** The client is designed as a thin client using HTML in order to omit installation of software on the client side. Business logic is realized by ASP.NET components, which are operated by an IIS server. Data logic is realized using SQL databases. Since we do not use stored procedures, views or triggers, we support almost any relational database product.

6.4 Risks and Countermeasures

We will outline *BetMPay*'s major risks and countermeasures from the perspective of design and implementation.

System Design. The system faces two fundamental security risks. At first, malicious users may fake vouchers by auto-generating the 16-character code and therefore be able to transfer credits, for which they never paid. To mitigate this risk, special measures for creating the code are used. The code is separated into an application number, a ticket number and one part of a ciphered message. The other part of the ciphered message is stored in a database when the code is generated. Whenever a user redeems a voucher, the two parts of the message get assembled, deciphered and compared to the template. Thus, the validity of the voucher can be checked.

The second fundamental risk is the fact that cashing out requires the possession of a specific cell phone. If the cell phone gets lost or stolen, other persons may retrieve money from the account. The risk gets reduced primarily by the circumstance that users have to create payment tickets. Users can create such tickets only with the password of the account. We have a two factor authorization in this case, i.e., possession of the cell phone and knowledge of the password. In addition, payment tickets are sent to the cell phone via SMS. Customers can ask their cell phone provider to lock the phone as soon as they recognize their loss. The cell phone will not be able to receive SMSs anymore. This provides further protection even in the case when a malicious person has the phone and knows the account password.

Generated payment tickets could be used when a cell phone gets stolen or lost. In case the malicious person only has the cell phone but does not know the account password, the customer may lock the account and therefore prohibit any cashing outs. If the malicious person also knows the account password, he can change the password and prevent logins by the regular customer. This risk can only be overcome by instructing users to use their payment tickets carefully. For example, they should be generated only shortly before a cashing out is requested.

System Implementation. *BetMPay* uses several strategies for mitigating common security risks, like brute forcing, SQL injection, or man-in-the-middle attacks. In order to prevent login brute force attacks, the number of login attempts within a certain time frame is determined. Whenever a certain threshold is exceeded, the user gets locked. SQL injection attempts can be mitigated by input validation. We use a double strategy and validate input on the client as well as on the server side. On the client side, a configurable AJAX component is used whereas the server side validation is done by .NET methods. For preventing man-in-the-middle attacks, communication between client and server is encrypted by means of HTTPS and the use of a certificate signed by a trusted certificate authority.

Password policy. A configurable AJAX component on the client side informs users whether their password complies with defined rules (minimum length, special characters). Additionally, passwords are checked on the server, e.g., if passwords have been used before. Users have to update password periodically.

6.5 Evaluation

Table 3 shows the results of the use-value analysis of *BetMPay* that have been gathered from a usability study. *BetMPay* is ranked third from a customer perspective and fourth from a provider perspective. This comes from restrictions regarding cross-border payment ability and the market penetration of this system. The aspect of preventing overhasty decisions has not been considered in our analysis. We can therefore conclude that if market penetration can be increased and cross-border support will be added, *BetMPay* will be a good choice for online gaming.

7 Related Work

Benefits and costs of ePayments as well as a tool for system comparison are given in [16]. There are also evaluations and comparisons of electronic payment systems.

Table 3: Use-values of BetMPay

Criteria	Weighting Customer Perspective	Value Customer Perspective	Weighting Provider Perspective	Value Provider Perspective
Password Strength	3	9	3	9
Password Renewal	3	9	3	9
Login Brute Force	6	18	3	9
Certificates	5	15	6	18
SSL/TLS	5	15	7	21
Authentication Method	10	30	5	15
User Lock-out	5	15	5	15
Additional Security	5	15	3	9
Market Penetration	10	10	15	15
Anonymity	15	45	5	15
Micro-payment	10	30	10	30
Payment Guarantee	3	9	20	60
Cross border Payment	10	10	10	10
Offline Payments	10	10	5	5
Use-value		**240**		**240**

Wright evaluates privacy, traceability, transaction cost, and the ability to build up the customers purchasing pattern of credit card payment systems, an electronic check system and digital cash systems. He evaluates advantages and disadvantages to customers, merchants, service providers as well as financial institutions [13]. Wright suggests a new system for increased user acceptance, which allows payments over telephone networks for purchases made over the Internet. Yu et al. explore advantages and limitations of online credit card payment, electronic cash, electronic checks, and small payments. Systematic and detailed comparisons of alternative systems are provided [14]. Their analysis was targeted at companies planning to adopt or to improve an ePayment system.

8 Conclusion

We have shown an evaluation of various ePayment systems by employing a use-value analysis. Key success factors of ePayment systems are security and flexibility. There is no "best" or "most secure" ePayment system. The adequacy of these systems depends on the application context. Use-value analysis is an appropriate and easy to use evaluation method, since it allows the consideration of different application perspectives. Despite the fact that many ePayment systems are available, there are contexts that require tailored solutions, as is the case with online gaming. We have introduced *BetMPay*, an ePayment system that offers a high level of anonymity, payment guarantees for providers, as well as consumer protection. One major drawback of *BetMPay* is that it requires an established selling infrastructure, i.e., concluding contracts with gaming offices or kiosks.

References

1. Abrazhevich D.: Electronic Payment Systems: a User-Centered Perspective and Interaction Design, 2004. http://www.idemployee.id.tue.nl/g.w.m.rauterberg/publications/PhD-Thesis%20%28Abrazhevich%202004%29.pdf
2. Heinrich L.J.: Informationsmanagement, 7th Edition, Oldenbourg 2002.
3. Leibold K., Stölzle R., Stroborn K.: ePayments und die Meinung der Konsumenten – Online-Survey; in Handbuch ePayment. Zahlungsverkehr im Internet: Systeme, Trends, Perspektiven (Ketterer K.H., Stroborn K.). Fachverlag Deutscher Wirtschaftsdienst, Köln 2002, p. 109-118.
4. NACHA – Electronic Payments Association. Electronic Payments Review & Buyer's Guide 2009. www.nacha.org
5. O'Sullivan, A., Sheffrin S.M: Economics - Principles in Action, Prentice Hall, 2003.
6. Payment Card Industry (PCI). Data Security Standard Requirements and Security Assessment Procedures, Version 1.2.1, July 2009. www.pcisecuritystandards.org/security_standards/pci_dss.shtml
7. Pahl G., Beitz W., Wallace K., Feldhusen J., Blessing L.: Engineering Design: A Systematic Approach, 3rd Edition, Springer 2007.
8. Slay, J., Koronios, A.: Information Technology, Security and Risk Management John Wiley & Sons, 2005.
9. Sumanjeet S.: Emergence of Payment Systems in the Age of Electronic Commerce: The State of the Art, Global Journal of International Business Research Vol. 2. No. 2. 2009.
10. Tan, M.: E-Payment – The Digital Exchange, Singapore University Press, 2004.
11. Tsiakis T., Stephanides G., Pekos G.: Trust and Security in Electronic Payments: What We Have and Need to Know? www.waset.org/journals/waset/v5/v5-36.pdf, 2005.
12. Viega J., McGraw G.: Building Secure Software, Addison-Wesley Professional Computing Series, 2002.
13. Wright, D., Comparative Evaluation of Electronic Payment Systems, Infor -Ottawa-, Vol. 40, Part 1, pages 71-85, 2002.
14. Yu H.-C., Hsi K.-H., Kuo P.-J., Electronic payment systems: an analysis and comparison of types, Technology in Society 24, pages 331-347, 2002.
15. Electronic Payment Systems Observatory (ePSO), Building Security and Consumer Trust in Internet Payments, – The potential of "soft" measures, Background Paper No. 7, 2002
16. Electronic Payments, www.electronic-payments.co.uk.
17. Wikipedia,Gambling, http://en.wikipedia.org/wiki/Gambling

An Engine for Local Communities Sharing Services and Content

Pierfrancesco Bellini, Daniele Cenni, Vladimiro Corsi, Paolo Nesi
DISIT-DSI, Distributed Systems and Internet Technology Lab
Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze, Firenze, Italy
http://www.disit.dsi.unifi.it/, nesi@dsi.unifi.it, paolo.nesi@unifi.it

Abstract

The trend of social networking has put in evidence the needs of users in collaborating and sharing information. Users prefer to keep under control their content and to have at home personalized views of their preferred services. Thus local community services and content management systems may be of interest to be developed in contrast to service provider that try to monopolize all services. In this paper, we present LOCUSGENS architecture and solution and in particular the features and the capabilities of the intelligent local Autonomous Community Manager which includes a rule based systems to cope with the harmonization of third party services.

1. Introduction

The state of the art of Social Networks, SN, is grounded on centralized solutions for collection and sharing of user generated content and information. Many different business models have been exploited in social networks, including: subscription, advertising, communities sharing similar interests, access to premium content or services in Pay Per Play/View. Communities whose members' involvement is characterized by occasional attendance, and short passive participation are generally suitable for business models based on advertising [1]. This model is typically viable in the presence of huge number of users, while the premium VOD (video on demand)/ PPV model is mainly associated with solutions for protecting content (e.g., CAS, DRM).

The main problem of SN is their centralized approach. In distributed solutions costs may be shared among middle level servers and/or with final users' devices (such as decoders, computers, mobiles, game stations, etc.) or servers: such as media centers, routers, etc. The concept of media centre is less widespread as a commercial device, while it is more common in the open source community where a set of Media Centre software solutions are available. They have been mainly derived from public open source media centers such as XBMC, LinuxMCE, Boxee, MythTV, etc. They mainly integrate capabilities of playing video streams coming from DVB boards and from IPTV streams, receiving EPG, program recording, connecting to some social networks, etc. One example about the standardization of media centre can be the MPEG M3W [2], which also integrates the MPEG-21 DRM support and concepts [3], [4]. In most cases, the media services are statically connected on receiving devices and offered to users. These solutions integrate poor features of content management, advertising management, gateway to manage the owned content in other social networks, automated content publication, and never rules for automating those activities. Furthermore, many service providers are interested in providing social features to the community that visit their websites, allowing them to have a full control of their social information, activities, and relationships [5]. On the other hand, content services for the community are today divided into: data provider, such as Web Services, RSS and ATOM; (ii) data Aggregator, such as Widgets and Mashups.

The aim of this paper is to present the main engine for setting up scalable local services for community enabling local management of complex services for user and content. The presented solution is part of a larger project called LOCUSGENS: Local Communities Sharing Generated Content, Services and experiences. The components presented in details in this paper refer to the engine that one could host on a media center located into the community premises to provide services to the local community: (1) remapping and reshaping a set of external services and information, (2) providing a set of additional services for integration of the former services and for the fact that the local information and events are locally exploited and managed. The added value services are performed by means of a rule based engine to automatically perform a set of activities for the community such as: download, local recommendation based advertising, privacy control, automated meeting/connection establishment, content filtering, automated content sharing on social networks, dynamic/active query and download, monitoring, notification of events in the community, etc. The definition of a set of rules for the local management of the community may simplify the life to the community users and protect the young elements.

2. LOCUSGENS General Architecture and Scenarios

According to LOCUSGENS, a community is a group of users sharing a set of needs, contents, services and experiences, and may be a small set of locations. LOCUSGENS community can be a family, a group of friends, a public laundry, a restaurant, a golf club, a social group attached to one or several social networks, etc. In Figure 1, the main elements of the LOCUSGENS architecture are reported. The architecture includes a set of Logically Accessible Directories (LAD services) that are providing services to the Autonomous Community Manager (**ACOM**). LAD services can be Social Network, content providers, web TV, weather forecast, EPG for TV or other, travel agency, kids schools, golf club, other ACOMs as well, etc.

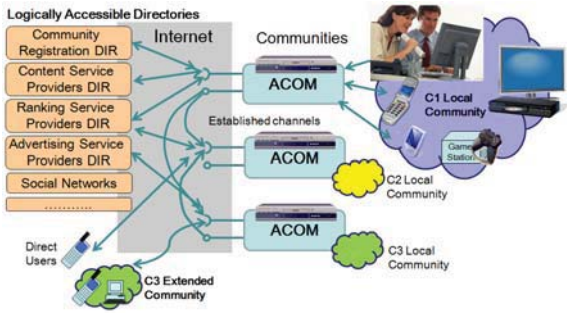


Figure 1 - LOCUSGENS general architecture

The ACOM can be hosted in a dedicated device or in a media center, decoder, set top box, etc. The ACOM can profile local users and according to rules/policies it can manage the community by providing content services to the community devices. In these cases, the content is considered as a set of complex data items where several different kinds of streams of information need to be provided, where multiple paths and possibilities as occurs including classical audio/video (DVB-T, DVB-S..) for IPTV/WebTV, but also educational content, cultural heritage content, games, collections, guides and manuals, cross media reports, play lists, etc. Examples are MPEG-21, AXMEDIS, SCORM, MPEG formats, HTML, SMIL, EPUB, etc.

The advantages of the LOCUSGENS architecture are: (i) scalability in terms of socially produced and shared content; (ii) decoupling of content services and advertising; (iii) flexibility in performing for the community repetitive and boring activities from/to other communities and social networks, etc.

Examples of ACOM policies can be:

- content services identified or matching a certain profile can be accessed only at a given time window during the day or from a specific device of the Community.
- locally produced video has also to be accessible on the community mobiles and on a set of other identified communities.
- At 20:30 the ACOM has to request a comment on Political news from Philip's friend Carl belonging to the Community "Smith family".
- When Tom connects show him last Amazon books on fishing.
- Philip's video of last week-end can be accessed/played from external users of the ACOM belonging to the following list of communities: Smith Family, Pub Happy Boat.
- Access to weather information and make them accessible at 19:00, together with news of Channel RAI1 and BBC.
- At 20:00 the ACOM WEBCAM is activated to establish a contact with the grandmother's.
- Allow reception of Advertising in Push if it matches with certain descriptors provided for or it is associated to specific content services.
- At the 15:00 please activate a query on YouTube to get the first 15 videos on "ski Rocky" at high resolution.

3. ACOM Architecture

The ACOM architecture is presented in Figure 2 and includes:

- **GUI support**: the front end interface for the devices and applications that are used in the local community and virtually from the Internet. This interface allows to mount the selected content services from the LOCUSGENS network of ACOMs communities and to make them available for the community devices and users. As widget manager the Liferay solution has been adopted [6].
- **Service Registry**: It is the entity to which each new service has to be registered to provide connection with the Process Engine to put the connected services at the disposal of the ACOM.
- **Social Media Services**: These services have been developed as native tools to provide a minimum number of basic capabilities in the ACOM middleware such as: managing users, content management system, crawling in the local area network for collecting resources and media, direct connection and chat, events and asynchronous messages. All these services are at disposal of the Process Engine and Policy Engine.
- **Process Engine and Editor:** the Process Engine is an intelligent engine of the ACOM devoted to the user administration with user access control, and manages the actions in response of the widgets mounted on the user profile. The Process Engine is also responsible of considering semantic aspects of

content services: (i) user actions, preference, favorites, etc.; (ii) content descriptors; (iii) local and external events; (iv) actions performed on provided and exported services; etc., to provide local personal recommendations and advertising support (static and dynamic user profile).

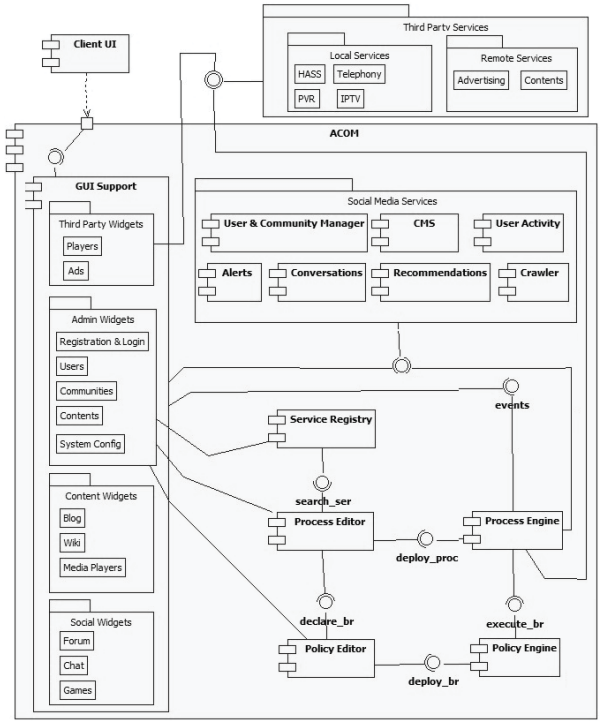


Figure 2 – ACOM architecture

- **Policy Engine and Editor**: The Policy Engine is the "intelligent" engine of the ACOM to process and execute ACOM policies as defined and downloaded by the ACOM and associated with Content Services and other services. The ACOM Policy Engine can take decisions to serve the local community. The capabilities of the engine are enforced by using semantic processing functions. Policies are formalized according to the following EBNF format (in which not all constructs are reported).

<Rule> = WHEN <PredicativePart> DO <ActivePart>
<PredicativePart> = <FiringCondition>,{<MatchExpression>}
<FiringCondition> = <InstantCondition> | <EventCondition>
<EventCondition> = ON EVENT WITH TYPE <EventLabel>
FROM <DomainCondition> [FOR
<DomainCondition>]
<DomainCondition> = <ObjectType> <Identifier>
([<Constraint> {AND <Constraint>}])
<Constraint> = [NOT] <Atom>
<Atom> = <Attribute> <OP> <Parameter> |
<Predicate>({<Parameter>})
<MatchExpression> = MATCH <DomainCondition>
<ActivePart> = <SyncStatement>, {<AsyncStatement>}
<SyncStatement> = Block action | Allow action
<AsyncStatement> = <StatementName> ({<Parameter>})

Some of the constructs of the Policy Rules directly refer to services provided in the Social Media Services (such as: sending events, getting status, etc.), while other are generically referred to dynamically mounted services coming from the Widgets connected by the users for creating their working user interface. The user interface has been realized in Drools Guvnor.

The LOCUGENS widgets are those accessible as Liferay on the web plus those specifically developed to be controlled by the Process and Policies Engines. All the widgets are hosted on Liferay support, only the LOCUGENS widgets can be managed by the Process and Policy Engines.

In order to **validate** ACOM solution a number of Widgets have been implemented according to the ACOM communication model, among them:

- Amazon Ads Widget realized to interact with ACOM engines to provide queries and get the published books matching with the a user profile,
- Alerts Widget has been realized to interact with ACOM engines to receive information and events, for example the reaching of a message, the occurrence of a certain condition, the login, the creation of a connection, etc.,
- Content Widget can be instantiated in three different manners as: content player, content list viewer, and as interface to set preferences. It has been realized to be installed as a Liferay portlet and to interact with ACOM engines. In Figure 3, the class diagram showing Content Widget and related classes is presented.

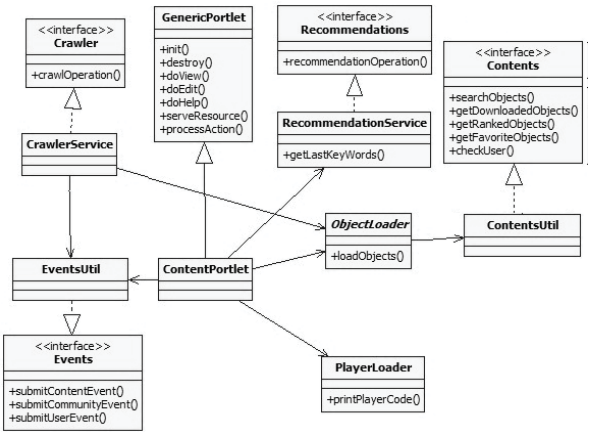


Figure 3 – Class diagram of Content Widget

- Mobile Medicine (http://mobmed.axmedis.org) is a best practice network for medicine. In this case, a widget has been created to provide into the ACOM special services from Mobile Medicine: getting content and providing content from/to the

social network, etc. The Mobile Medicine social network presents a web service to provide those services towards its widget counterpart for the ACOM.

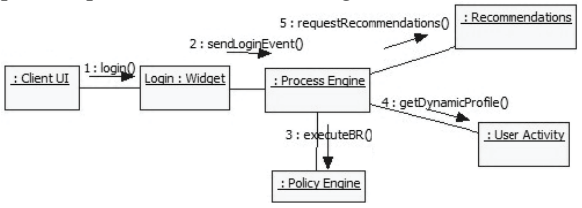
Other services are available as well such as Portal Event Listener, Conversation Service e Download Service, User Activity Service (UAS), etc.

4. Some Examples

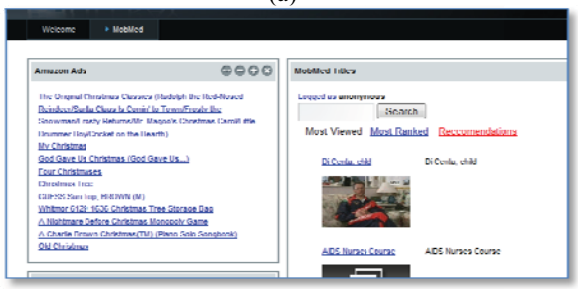
The solution proposed has been tested by setting up a number of ACOM services into communities. These where comprised of few people, typically families and their friends. On these bases, a number of rules have been created by using a graphic user interface. The creation of a new rule is guided by a sort of wizard and should be improved to make it realistically understandable by any final user. An example of rule to prevent access to video content for Junior users belonging to the "Family" community is reported:

```
WHEN
   ON EVENT WITH TYPE "Play"
   FROM User u(age < 18,
         belongsToCommunity("Family"))
   FOR Content c(format == "Video")
DO Block action
```

Figure 4a represents the collaboration diagram describing the relationships among the entities involved into the execution of the previous rule. While Figure 4b shows a segment of the screen in which the advertising produced by querying on Amazon portal on the basis of user profile (static plus dynamic) is presented. So that exploiting the preferences manifested during the accesses on Mobile Medicine to provide personal local advertising of Amazon content.



(a)



(b)

Figure 4 – Personal local advertising produced dynamically on the basis of Amazon content and local user profiling.

5. Conclusions

In this paper, LOCUSGENS architecture and solution and in particular the features and the capabilities of the intelligent local Autonomous Community Manager which includes a rule based system to cope with the harmonization of a number of third party services have been presented. The solution proposed goes in the direction of locally managing services for communities. Among them content services but also news, connection, advertising, control and supervising, automating some boring activities of content collection and sharing among friends and social networks. The components presented in details in this paper refer to the engine that one could host on a media center located into the community premises to provide services to the local community. The added value of the solution resides on the possibility of setting up a set of "smart" rules to automatically perform synchronous and asynchronous activities for the community such as: download, local recommendation based advertising, privacy control, automated meeting/connection establishment, content filtering, automated content sharing on social networks, dynamic/active query and download, monitoring, notification of events in the community, etc. From the experiments we have verified that they can simplify the life of the community reducing boring activities and creating more stable connection among other communities, the larger families, the houses of the close friends, etc.

References

[1] Bourdoncle, F. (2008). Future Internet: Multi-Media Search Engines and Scalable Infrastructures. Conference Session Report on The Future of the Internet, ICT Event.

[2] A. Colzi, T. Martini, P. Nesi, D. Rogai, ``A Component-Based Solution and Architecture for Dynamic Service-Based Applications", The Nineteenth International Conference on Software Engineering and Knowledge Engineering, SEKE'2007, Hyatt Harborside at Logan Int'l Airport, Boston, USA, July 9-11, 2007, Organized by Knowledge Systems Institute.

[3] P. Bellini, P. Nesi, D. Rogai, "Exploiting MPEG-21 File Format for cross media content", Proc. Of the International Conference on Distributed Multimedia Systems, DMS 2007, September 6-8, 2007, San Francisco Bay, USA, Organised by Knowledge Systems Institute.

[4] MPEG Group, "Introducing MPEG-21 DID", http://www.chiariglione.org/mpeg/technologies/mp21-id/index.html

[5] R. Geambasu, C. Cheung, A. Moshchuk, S.D. Gribble, H. M. Levy. Organizing and Sharing Distributed Personal Web-Service Data.

[6] Yuan, J. X. (2009). Liferay Portal 5.2 Systems Development. Packt Publishing.

THE EXPLAINER
Using the Computer to Help Collaborate in Connecting the Dots to Resolve Complex Problems

Donald V Steward – Emeritus Professor – Sacramento State University
Josh Knight – PhD Student – University of Michigan

Abstract

The Explainer is a means whereby people can use a computer to help them collaborate to bring together their various points of view in order to solve very complex problems. The collaborators contribute elementary cause-and-effect statements that describe a situation so that the computer can use propositional logic to combine these statements to explain behaviors that arise within the scope of that situation.

Propositional logic is used to represent the cause-and-effect statements. Abduction is used to find the multiple explanations for the behavior. Deduction is used on each proposed explanation to show what additional behaviors it would also explain. Proposed explanations that explain false behaviors are rejected.

People tend to solve problems by recognizing a pattern that may be occurring now and correlate it with a pattern they have seen before. They assume that the cause and the outcome will be the same. But different people recognize different patterns and argue over them rather than solve the problem.

A better way to solve problems is to connect the dots to build a scenario that shows the connections linking the cause to the symptoms. This scenario is transparent to all and can be the basis for discussion and criticism.

People are naturally much better at pattern recognition and correlation than in connecting the dots. The Explainer provides the ability that allows them to connect the dots.

1.0 Connecting the Dots: A Better Way to Solve Problems

People are naturally very good at pattern recognition and correlating patterns they observe now with other patterns from the past. They assume that the causes and outcomes of these patterns will be the same. Unfortunately, different people recognize different patterns, come to different conclusions as to what should be done, and argue without having solved the problem.

A better way is to connect the dots to provide a scenario that shows the logic that connects the causes to their consequences. These logical scenarios can then be exposed for criticism and changed as needed. This process is transparent.

Unfortunately, this is not as natural for people as correlating patterns. People are very poor at connecting the dots. The Explainer is a means of using the computer to help us do what is not so natural to us. It opens up the capacity to use a new way of thinking. This capability is a paradigm change made possible by programming a computer to handle propositional logic.

An explanation must conform to two conditions. One, it must explain what is desired to be explained. And two, it must not explain behaviors that do not exist.

Abduction is used to find the explanations for what is to be explained that can occur within the scope of the cause-and-effect knowledge. Of course, the validity of the explanation is limited by the validity and scope of the cause-and-effect statements. This is the creative phase.

Deduction is used to determine what other behaviors accompany each proposed explanation beyond the behavior that is intended to be explained, i.e. side effects. If these other behaviors do not occur, then that explanation must be rejected. This is the critical phase.

The abduction process is very straightforward provided there are no circuits in the cause-and-effects needed to explain the behavior. All that is required is to eliminate the variables between the assumptions and the behavior they are to explain. An assumption is a cause that itself has no cause and thus can only be assumed. This elimination is done by substitution much as is done in the solution of simultaneous linear algebraic equations. But this process breaks down when there are circuits.

The rules of logic are based on the assumption of an open system. If A implies B and if A is true, then B is true. But we cannot say in logic that if A is false then B is also false because the truth of B may be implied by something else that is not considered. But if we assume we have a closed system where nothing inside the system can be implied by something outside the system, we can account for everything that might imply that B is true. Then we can say that if A is false and there is no other implication within the system that would make B true, we can then say that if A causes B that if A is false then B must also be

false. This assumption of closure is important to have a practical process of explanation.

1.1 The Explainer Also Solves Other Types of Problems

When we can solve the explainer problem, there are many other types of problems that we can also solve. For example, we can design something from components whose behaviors have been described by cause-and-effect statements by asking that they explain the behavior that would represent the requirements for the design. We can develop a plan by describing the steps we can take as cause-and-effect statements and request an explanation for the goal of the plan.

2.0 Dealing with Circuits

When there are circuits, it is assumed that functions can be found such that the variables within the circuit become functions of variables outside the circuit but not of the other variables within the same circuit. When these functions are used, the circuits are replaced by equivalent cause-and-effect statements that contain no circuits. Then these variables can also be eliminated by substitution, as we had done when there were no circuits.

Unfortunately, finding these functions is a combinatorial problem. This raises the suspicion that the combinatorics might require more computing capability and time than is available in practice.

Fortunately, there are characteristics of this problem that make the computing practical for all reasonable problems. In each combinatorial trial, it is possible to assign multiple variables at the same time. This greatly reduces the number of trials. If for example r variables can be assigned in each trial, the number of trials would be $2^{b/r}$ where b is the number of variables in the block, and r is the number of variables assigned in each trial. A block is the largest subset of variables such that there is a path from each variable in the block to every other variable in the block and back to itself.

When the block contains only a simple circuit, r = b, thus requiring only two trials no matter how large b is.. We can assume that unless the circuits in a block are severely entangled, the number of trials would be expected to be very few.

Also, problems usually consist of many small blocks that can each be considered independently. This greatly reduces the number of trials required.

3.0 Example: Understanding the Economic Crisis Problem

We have used the Explainer to explore how it might be used to shed light on several complex problems. We do not contend that we have solved these problems, but only to suggest how the Explainer might be used in the hands of more knowledgeable people to produce better understandings of these problems.

The Explainer has been used in trials to illustrate how it might be applied to understanding a number of different types of problems such as understanding how terrorists might design a plan to do us harm and how those plots might be thwarted, how we might consider the various factors involved in plans to leave Iraq, how it might be used to provide a medical diagnostic service to physicians who do not have the time to read all the latest literature, and so on.

But none of these examples has done any more than illustrate how these problems might be formulated. Better understanding of these problems would require that the Explainer be put in the hands of more knowledgeable people to develop better cause-and-effect statements.

3.1 The Economic Model

Cause-and-effects can be input to the Explainer as follows:

EFFECT 1
CAUSED BY A
NOT B
CAUSED BY C

Meaning that Effect 1 is caused either by (A and not B) or by C.

Then the input for this partial economic crisis analysis might be as follows:

EFFECT 1. Home prices depressed
CAUSED BY 17. More houses available on market
EFFECT 2. Credit crunch
CAUSED BY NOT 3. Liquidity
EFFECT 3. Liquidity
CAUSED BY NOT 4. Mistrust
EFFECT 4. Mistrust
CAUSED BY 16. High foreclosure rate
29. Pressure to make loans with inadequate documentation in order to collect fees
EFFECT 5. Initial lender possesses home partially paid for by borrower`s payments
CAUSED BY 14. Borrower defaults
EFFECT 6. Initial lender wants to sell ARM mortgage
CAUSED BY 5. Initial lender possesses home partially

paid for by borrower`s payments
10. Rate will rise to make loan profitable to lender
11. Initial lender has low risk in reselling loan
21. Initial lender gets origination fees
EFFECT 7. Borrower wants to buy ARM loan
CAUSED BY 9. Borrower has low income and low equity
12. Initial rate low
EFFECT 8. Many ARM loans made
CAUSED BY 6. Initial lender wants to sell ARM mortgage
7. Borrower wants to buy ARM loan
ASSUMABLE 9. Borrower has low income and low equity
EFFECT 10. Rate will rise to make loan profitable to lender
CAUSED BY 24. Initial lender cannot profit from initial rates
EFFECT 11. Initial lender has low risk in reselling loan
CAUSED BY 19. Initial lender bundles loans
30. Ability of lenders to recoup bad loans
EFFECT 12. Initial rate low
CAUSED BY 23. Nature of ARMs
EFFECT 13. Borrower cannot pay reset rate
CAUSED BY 9. Borrower has low income and low equity
10. Rate will rise to make loan profitable to lender
EFFECT 14. Borrower defaults
CAUSED BY 13. Borrower cannot pay reset rate
EFFECT 15. High delinquency rate
CAUSED BY 8. Many ARM loans made
14. Borrower defaults
EFFECT 16. High foreclosure rate
CAUSED BY 15. High delinquency rate
EFFECT 17. More houses available on market
CAUSED BY 14. Borrower defaults
32. Buyers walk away from homes
EFFECT 19. Initial lender bundles loans
CAUSED BY 22. Bundled loans have low risk rating
EFFECT 20. Artificially low risk assigned to bundles
CAUSED BY 26. Asymmetry of information about risks of bundles
EFFECT 21. Initial lender gets origination fees
CAUSED BY 23. Nature of ARMs
EFFECT 22. Bundled loans have low risk rating
CAUSED BY 20. Artificially low risk assigned to bundles
ASSUMABLE 23. Nature of ARMs
EFFECT 24. Initial lender cannot profit from initial rates
CAUSED BY 9. Borrower has low income and low equity
23. Nature of ARMs
EFFECT 25. Market collapse
CAUSED BY NOT 3. Liquidity
EFFECT 26. Asymmetry of information about risks of bundles
CAUSED BY NOT 34. Regulators guarantee symmetry of information about loan risks
EFFECT 27. Investors unaware of bundle risks
CAUSED BY NOT 28. Open books
29. Pressure to make loans with inadequate documentation in order to collect fees
EFFECT 28. Open books
CAUSED BY 34. Regulators guarantee symmetry of information about loan risks
ASSUMABLE 29. Pressure to make loans with inadequate documentation in order to collect fees
EFFECT 30. Ability of lenders to recoup bad loans
CAUSED BY
EFFECT 31. Negative equity
CAUSED BY 1. Home prices depressed
EFFECT 32. Buyers walk away from homes
CAUSED BY 31. Negative equity
EFFECT 33. Bundled loans bring good prices
CAUSED BY 26. Asymmetry of information about risks of bundles
ASSUMABLE 34. Regulators guarantee symmetry of information about loan risks

Figure 3.1: Initial model for economic crisis situation including circuit

Those with better knowledge of the situation are invited to use the Explainer to produce a more comprehensive analysis.

Input to block Begin
EFFECT 14. Borrower defaults
CAUSED BY 13. Borrower cannot pay reset rate
Input End

Variables Within the Block

Block begin
EFFECT 17. More houses available on market
CAUSED BY 14. Borrower defaults
32. Buyers walk away from homes
EFFECT 1. Home prices depressed
CAUSED BY 17. More houses available on market
EFFECT 31. Negative equity
CAUSED BY 1. Home prices depressed
EFFECT 32. Buyers walk away from homes
CAUSED BY 31. Negative equity
Block End

Figure 3.2 Input to Block Before Substitution

Figure 3.2 identifies a circuit within the cause-and-effect statements that must be resolved. Figure 3.3 shows the cause-and-effect statements after the substitution has

been made to produce an equivalent set of cause-and-effects that no longer contain circuits:

Input Variables to Block
Begin
EFFECT 14. Borrower defaults
CAUSED BY 13. Borrower cannot pay reset rate
End
Variables within Block
Block Begin
EFFECT 17. More houses available on market
CAUSED BY 14. Borrower defaults
32. Buyers walk away from homes
EFFECT 1. Home prices depressed
CAUSED BY 14. Borrower defaults
EFFECT 31. Negative equity
CAUSED BY 14. Borrower defaults
EFFECT 32. Buyers walk away from homes
CAUSED BY 14. Borrower defaults
Block End

Figure 3.3: Input to Block After Substitution So It No Longer Contains a Circuit

Effect 18 has been deleted because during the development it was found to be equivalent to Effect 1. This reveals why it is important to make sure that the variables are not duplicated with different numbers.

After making this substitution, we can run the existing program as though there were no circuits. Then we get the following explanation for the Market Crisis:

EFFECT 25. Market Crisis
CAUSED BY
29. Pressure to make loans with inadequate documentation in order to collect fees
NOT 34. Regulators guarantee symmetry of information about loan risks
23. Nature of ARMs
9. Borrower has low income and low equity

Figure 3.4: Explanation found for '25. Market Crisis'

This tells us that 'EFFECT 25. Market Crisis' can occur only if all these causes occur together and thus can be prevented if any one of them can be prevented.

We cannot change '23 Nature of ARMs' i.e. Adjustable Rate Mortgages, or '9. Borrower has low income and low equity', but we can change 'NOT 34. Regulators guarantee symmetry of information about loan risks'. This suggests that the 'Market Crisis' could have been prevented and now can be resolve by having regulations to guarantee symmetry of information in all financial transactions. As of this writing, the regulations requiring this are still a matter of contention and thus are not yet in place.

We can do this if there are no ORs. If there were ORs, we have to operate on multiple subsets of expressions where there is only one term chosen from each expression.

3.2 Scenario to Explain the Reasoning for the Explanation and Show Side Effects

The scenario shows the reasoning that leads to this explanation and provides the consequences if the explanation were valid. A term is a set of assumptions or actions joined by ANDs.

The scenario uses the following rules:

- A term is True if all its variables are True
- A term is False if any of its variables are False
- A term is Unknown if at least one of its variables is unknown and none of its variables are False
- An explanation is False if any term in its explanation is False and no term is True or Unknown
- An explanation is Unknown if no term is either True or False
- An explanation is True if any term is True

The program assigns values of IS TRUE, IS NOT TRUE or equivalently IS FALSE, or UNKNOWN to the effects, causes and terms. The unknowns have been eliminated in the following for clarity and brevity. Then the scenario is as follows:

DETERMINE THE CONSEQUENCES IF
THESE ASSUMPTIONS ARE ALL TRUE
29. Pressure to make loans with inadequate documentation in order to collect fees IS TRUE
34. Regulators guarantee symmetry of information about loan risks IS NOT TRUE
23. Nature of ARMs IS TRUE
9. Borrower has low income and low equity IS TRUE
THEN THE CONSEQUENCES ARE:
9. Borrower has low income and low equity IS TRUE
23. Nature of ARMs IS TRUE
29. Pressure to make loans with inadequate documentation in order to collect fees IS TRUE
34. Regulators guarantee symmetry of information about loan risks IS FALSE
12. Initial rate low IS TRUE
Term 1 IS TRUE
23. Nature of ARMs IS TRUE
21. Initial lender gets origination fees IS TRUE
Term 1 IS TRUE
23. Nature of ARMs IS TRUE

24. Initial lender cannot profit from initial rates IS TRUE
Term 1 IS TRUE
9. Borrower has low income and low equity IS TRUE
23. Nature of ARMs IS TRUE
26. Asymmetry of information about risks of bundles IS TRUE
Term 1 IS TRUE
NOT 34. Regulators guarantee symmetry of information about loan risks IS TRUE
28. Open books IS FALSE
Term 1 IS FALSE
34. Regulators guarantee symmetry of information about loan risks IS FALSE
33. Bundled loans bring good prices IS TRUE
Term 1 IS TRUE
26. Asymmetry of information about risks of bundles IS TRUE
7. Borrower wants to buy ARM loan IS TRUE
Term 1 IS TRUE
9. Borrower has low income and low equity IS TRUE
12. Initial rate low IS TRUE
10. Rate will rise to make loan profitable to lender IS TRUE
Term 1 IS TRUE
24. Initial lender cannot profit from initial rates IS TRUE
13. Borrower cannot pay reset rate IS TRUE
Term 1 IS TRUE
9. Borrower has low income and low equity IS TRUE
10. Rate will rise to make loan profitable to lender IS TRUE
14. Borrower defaults IS TRUE
Term 1 IS TRUE
13. Borrower cannot pay reset rate IS TRUE
17. More houses available on market IS TRUE
Term 1 IS TRUE
14. Borrower defaults IS TRUE
32. Buyers walk away from homes IS TRUE
1. Home prices depressed IS TRUE
Term 1 IS TRUE
14. Borrower defaults IS TRUE
31. Negative equity IS TRUE
Term 1 IS TRUE
14. Borrower defaults IS TRUE
32. Buyers walk away from homes IS TRUE
Term 1 IS TRUE
14. Borrower defaults IS TRUE
20. Artificially low risk assigned to bundles IS TRUE
Term 1 IS TRUE
26. Asymmetry of information about risks of bundles IS TRUE
22. Bundled loans have low risk rating IS TRUE
Term 1 IS TRUE
20. Artificially low risk assigned to bundles IS TRUE
27. Investors unaware of bundle risks IS TRUE
Term 1 IS TRUE
NOT 28. Open books IS TRUE
29. Pressure to make loans with inadequate documentation in order to collect fees IS TRUE
30. Ability of lenders to recoup bad loans IS TRUE
Term 1 IS TRUE
33. Bundled loans bring good prices IS TRUE
5. Initial lender possesses home partially paid for by borrower's payments IS TRUE
Term 1 IS TRUE
14. Borrower defaults IS TRUE
19. Initial lender bundles loans IS TRUE
Term 1 IS TRUE
22. Bundled loans have low risk rating IS TRUE
11. Initial lender has low risk in reselling loan IS TRUE
Term 1 IS TRUE
19. Initial lender bundles loans IS TRUE
30. Ability of lenders to recoup bad loans IS TRUE
6. Initial lender wants to sell ARM mortgage IS TRUE
Term 1 IS TRUE
5. Initial lender possesses home partially paid for by borrower`s payments IS TRUE
10. Rate will rise to make loan profitable to lender IS TRUE
11. Initial lender has low risk in reselling loan IS TRUE
21. Initial lender gets origination fees IS TRUE
8. Many ARM loans made IS TRUE
Term 1 IS TRUE
6. Initial lender wants to sell ARM mortgage IS TRUE
7. Borrower wants to buy ARM loan IS TRUE
15. High delinquency rate IS TRUE
Term 1 IS TRUE
8. Many ARM loans made IS TRUE
14. Borrower defaults IS TRUE
16. High foreclosure rate IS TRUE
Term 1 IS TRUE
15. High delinquency rate IS TRUE
4. Mistrust IS TRUE
Term 1 IS TRUE
16. High foreclosure rate IS TRUE
29. Pressure to make loans with inadequate documentation in order to collect fees IS TRUE
3. Liquidity IS FALSE
Term 1 IS FALSE

NOT 4. Mistrust IS FALSE
2. Credit crunch IS TRUE
Term 1 IS TRUE
NOT 3. Liquidity IS TRUE
25. Market Crisis IS TRUE
Term 1 IS TRUE
NOT 3. Liquidity IS TRUE

Figure 3.4: Scenario for Cause of Economic Crisis

A scenario can also be generated for the case that variable 34. Regulators guarantee symmetry of information about loan risks is true.

5.0 Plans for Further Work

A more imaginative use of the Explainer would be to design a device that can use sensors and the Explainer method to explain its environment and make plans for how to use actuators to interact with its environment.

Such a device might have two methods for understanding its environment. One, it could use its sensors to correlate pairs of events that always occur in the same order, occur frequently, and preferably with a short duration between the events. Thus, it would find candidates for cause-and-effect statements to be added to its repertoire. It could also use its cause-and-effect statements to plan experiments on its environment that would yield further cause-and-effect statements and perhaps also produced actions that would achieve its needs and desires.

This has not been implemented yet, but is an interesting possibility to be considered. It might provide a hypothesis for how babies learn about the world and how to deal with it. It might be used by robots to learn about their environments and how to deal with them.

The Explainer is also being considered as a tool for negotiations and emergency management.

6.0 Conclusion

The Explainer shows a method that might be used to help people collaborate to understand complex problems and see how they might be resolved. Unlike expert systems, this method can handle circuits, which are not uncommon, and there is no ambiguity in the sequencing of the application of rules.

If equivalent methods are available to study complex problems such those that the Explainer appears capable of understanding, there does not appear to be evident that such methods are being use to deal with the many significant problems that we are facing today.

7.0 References

7.1 Written Materials

[1] Brachman, Ronald & Hector Levesque (2004) Knowledge Representation and Reasoning (Morgan Kaufmann)

[2] Garey, Michael R. & David S. Johnson, (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness (W. H. Freeman and Company)

[3] Giarratano, Joseph C. & Gary D. Riley, (2007) Expert Systems: Principles and Programming 4th ed. (Thomson Course Technology)

[4] Josephson, John R. & Susan G. Josephson (1996) Abductive Inference: Computation, Philosophy, Technology (Cambridge University Press)

[5] Pearl, Judea, (1988) Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference Revised 2nd Printing (Morgan Kaufmann, Publishers, Inc.), particularly pg. 195 footnote

[6]. Pearl, Judea, (2000) Causality: Models, Reasoning, and Inference (Cambridge University Press)

[7] Charles Sanders Peirce, ed. Kenneth Laine Ketner (1992) *Reasoning and the Logic of Things* (Harvard University Press)

[8] Peng, Yun & James A. Reggia (1990) Abductive Inference Models for Diagnostic Problem-Solving (Springer; 1st edition)

[9] Steward, Donald V, (1981) *Systems Analysis and Management: Structure, Strategy and Design*, (Petrocelli Books, Inc., now McGraw-Hill)

[10] Walton, Douglas (2004) *Abductive Reasoning* (The University of Alabama Press)

7.2 Materials on the Web

Go to http://problematics.com/readings.asp
then:

[11] Click on
THE EXPLAINER METHOD - P.doc
for a more in depth discussion of this paper.

[12] Click on APPLICATIONS-E.doc to see demonstrations of how the Explainer might be used to deal with various types of problems.

Note: ***This work was done in conjunction with David Singer of the University of Michigan and Janel Nixon of Integrative Engineering under a grant from the Office of Naval Research.***

Verification of Security Policy of Service Oriented Systems

Antonio Maña, Gimena Pujol
Computer Science Department, University of Málaga
{amg,gimena}@lcc.uma.es

Abstract

In this paper we present a mechanism for verification of security policies that is part of a policy-based security engineering process for service-oriented applications. The mechanism, developed in the SERENITY and MISTICO projects, provides a framework for the automated treatment of security and dependability (S&D) issues. Concretely, we focus on the verification of the compliance with security policies, based on the formal specification of Security and Dependability (S&D) Properties. Precisely, the main advantages of the approach presented in this paper are that it allows us to define high-level policies and to verify that an application developed in these security frameworks complies with such policy.

1. Introduction

The trend in recent years toward open and distributed computing environments, and in particular towards service-oriented computing, increases the emphasis on relationships, negotiations, and agreements between the different components of the systems. This brings particular challenges for the area of security and dependability (S&D), which are traditionally very difficult to manage and measure. It also introduces the issues of accountability and liability. These are often also referred to as responsibility, answerability, blameworthiness and other terms associated with the expectation of account-giving. These topics have been heavily debated because introduce a new concept in security engineering, related with the responsibility to ensure that a system complies with a regulation or a security policy. We can find in the recent past a practical example, this is the case of corporate scandals and breakdowns like the recent loss of 45 Million credit and debit card numbers and other personal data by the TJX Corporation, which have flourished in the last years, highlighing the need for stronger compliance regulations for publicly listed companies.

Work partially supported by DESEOS project (TIC-4257) funded by the government of Andalucía and by the E.U. through SecFutur project (FP7-ICT-2009-5).

This is a direct consequence of the fact that the security has been traditionally overlooked and considered supplementary instead of a core element in the design and development of such systems, which has produced very negative consequences and has undermined the users' confidence in computer systems.

One of the most significant regulations in the financial context is the Sarbanes-Oxley Act developed in 2002, which defines significant tighter personal responsibility of corporate top management for the accuracy of reported financial statements. As in this case, many of the best known initiatives for enhancing the security of computer systems have been based on guidelines, recommendations, best practices, certification, compliance and similar approaches lacking the necessary rigour and precision that one would expect when dealing with "security". Other relevant examples are: Common criteria, FIPS, traditional security patterns, GLBA, FISMA, and HIPA. In some cases other compliance frameworks (such as COBIT) or standards (NIST) inform on how to comply with the regulations.

The fact that all these regulations and policies are expressed informally make it practically impossible to rigorously verify that an application complies with a specific regulation. Precisely, one of the main advantages of our proposal is the formalization of the aforementioned regulations in high-level "Security policies" and to allow verifying that an application developed in the security framework complies with such policy.

In this paper we present a model of secure and dependable systems and its application to the Service-Oriented Computing, focused on the verification of the compliance with security policies. The model has been developed in the SERENITY and MISTICO projects. SERENITY is a research project founded by the European Commission, in its Sixth Framework Programme [1]. MISTICO is a project founded by the Consejería de Ciencia y Tecnología (Government Castilla-La Mancha).

2. Motivating Example – the scenario

Peter is a 76 years old widowed man. Peter has been discharged from hospital after a cardiac arrest. Peter feels quite good at the moment and he is not required to spend his entire day at home, but of course his health status needs to be monitored continuously 24 hours a day.

At this purpose Peter carries a monitoring device that regularly measures his heart rate and some other data such as his blood pressure, body temperature, etc. All these devices are integrated to an e-health mobile terminal that, besides all these things, reminds Peter to take daily medicines and allows him to promptly communicate medical data to his doctor. Peter also carries a motion sensor providing an alert as soon as Peter becomes passive for two minutes. In case of emergency the personal health monitoring system and the motion sensor send a confidential and authentic alert, which is handled by the SERENITY framework running at the emergency response centre.

We going to focus in a concrete use case in order to complete this example. Peter enjoys an amazing day on the seaside. He is walking when suddenly he feels dizzy and shortly after loses consciousness. His personal health monitoring system and his motion sensor send a confidential and authentic alert, which is handled by the SERENITY framework running at the emergency response centre. Any attempt to reach Peter on his mobile phone fails and therefore Peter's current location is determined and an emergency response team is activated and sent to Peter. On the way to reach Peter, the medical team uses a PDA device that, through the SERENITY framework, remotely authenticates with Peter's e-health monitoring terminal and gains access to the confidential information about Peter's health status and his recent medications.

Due to the regulation matters, all the applications in this scenario must comply with the Health Information Protection Act (HIPA). This regulation is designed to improve the privacy of people's health information while ensuring adequate sharing of information is possible to provide health services.

The first challenge for the developer (especially if he is not a security expert) is the integration of the security in his application. The second is ensuring that the application complies with the regulation. SERENITY framework gives a security environment to develop application and to check that an application complies with a regulation as is explained in the sections 4 and 5. After introduce the reader in the SERENITY ambient, we complete this example in the section 6 with a technical point of view of a specific scene.

3. Background

Companies have long sought to integrate existing systems in order to implement information technology (IT) support for business processes that cover all present and prospective systems requirements needed to run the business end-to-end. SOA (Service-Oriented Architecture) offers one such prospective architecture. It unifies business processes by structuring large applications as an ad hoc collection of smaller modules called "services". Different groups of people both inside and outside an organization can use these applications, and new applications built from a mix of services from the global pool exhibit greater flexibility and uniformity. Nevertheless, the lack of security is the main impediment to the adoption of service-oriented systems in open scenarios. Concretely, for Web Service approach, existing security solutions but these are still based on technology that has been used for several years to secure e-commerce sites. In fact, until recently, the most common approach was to use elements like standard firewalls, Secure Sockets Layer encryption, and ad-hoc gateways for authentication and authorization. The WS-Security specification has represented an important advance in the standardization of mechanisms supporting security for WSs. This initiative describes enhancements to the Simple Object Access Protocol (SOAP) to provide protection of messaging. WS-Security also provides a general-purpose mechanism for associating security tokens with messages and describes how to encode binary security tokens such as X.509 certificates. Although WS-Security does not address security issues such as authorization or access control, it represents a useful initiative to support other security services.

Regarding Security and Dependability (S&D) policies, much work has been done for specific purposes, such as access control or security configuration management. In general, a computer security policy defines the security-related goals and assets of an organization's information systems. Policy definitions range from informal to highly formal. In the latter case we can find many examples of formal description of policies, especially related to confidentiality and access control. Examples of these are the Bell-La Padula model [2], which supports multilevel confidentiality, the Chinese Wall [3], or role-based access control (RBAC)[4]. Integrity is also the focus of other policy models such as [5] among others. Verifying informal policies such as regulations, certifications criteria, etc. represent a bigger challenge due to their broad scope and informal nature. It is important to highlight that many of the best known initiatives for enhancing the security of computer systems have been based on informal specifications (e.g. Common criteria, FIPS, Sarbanes-Oxley, etc).

Currently, there are some initiatives that are focused on the verification of the compliance to a policy. This is the case of the MASTER project, which by means of observing the application of several security controls tries to ensure that an application complies with a given policy.

4. The SERENITY Model of Secure and Dependable Systems

The SERENITY objective [6] is to provide a framework for the automated treatment of S&D issues in Ambient Intelligence (AmI) scenarios. To achieve this goal, three key aspects have been identified: Capturing the specific expertise of the security engineers in order to make it available for automated processing (S&D Artefacts), giving to the developer tools to add to the application code calls to the SERENITY framework to request the security solutions (SERENITY-aware applications); and providing means to perform run-time monitoring of the functioning of the security and dependability mechanisms (S&D mechanisms).

4.1. Artefacts for modelling S&D Solutions

One of main SERENITY results is the set of modelling artefacts used in order to capture security expertise; these are semantic descriptions at different levels of abstraction. The main artefacts are:

- *S&D Solutions* is a general term that we use to refer to well-defined mechanisms that provide one or more *S&D Properties*.
- An *S&D Property* is an interpretation of a security or dependability quality [7] such as confidentiality, integrity, availability, etc. A reference to the S&D Properties provided must be included in the specification of each S&D Solution. This topic is analyzed in the next section.
- *Security policies* are used to capture and enforce enterprise-wide policies, and also security regulations. These are detailed descriptions of the S&D Properties that must compliance an application to fulfill a regulation or policy. This artefact is the cornerstone of the paper and this is detailed exposed in the section 6.
- *S&D Patterns* are detailed descriptions of abstract but specific *S&D Solutions*. These descriptions contain all the information necessary for the selection, instantiation, adaptation, and dynamic application of the security solution represented in the S&D Pattern.
- *S&D Classes* represent abstractions of a security service that can be provided by different S&D Solutions. All solutions belonging to an S&D Class are characterized by providing the same S&D Properties and complying with a common interface.
- S&D Implementations represent the components that realize the S&D Solutions and therefore contain the specific features of that component (e.g. additional restrictions). These components are made accessible to applications throughout the Serenity Runtime Framework (SRF).
- Finally an ExecutableComponent (EC) is the actual implementation of an S&D Pattern. ECs are represented by S&D Implementations. These elements are not used during the development phase. Instead, they are the functional realization of S&D Solutions that is used at run-time. An ExecutableComponent works as a stand alone executable S&D Solution ready to provide security and dependability services to applications.

Due to the inherently unpredictable and dynamic nature of AmI environments SERENITY provides two frameworks to support the security and dependability of applications operating in AmI scenarios: (i) the SERENITY Development Framework (SDF), which includes tools for developing S&D Solutions and supports the development of SERENITY aware applications; (ii) the SERENITY Run-Time Framework (SRF) is responsible for dealing with S&D Solutions referenced by the SERENITY-aware applications at run-time (by means of what we call S&D Requests). The SRF uses a set of mechanisms for context evaluation. In this way, the SRF knows the current and the past device context conditions. The provision of security solutions by the SRF is explained in more detail in [8].

The process for developing SERENITY-aware applications deals with S&D issues from the very early engineering phases. SERENITY-aware applications contain calls to the SRF requesting it to provide references to ECs. These calls are called S&D Requests. Then, at runtime the SRF processes these requests, selects the best available EC to fulfill the request, activates it and returns a handler to the application. From that moment on the application can access the EC directly without the intermediation of the SRF. ECs are deactivated when context changes make them no longer suitable or by explicit request from the application.

4.2. The SERENITY Semantic Approach for S&D Properties

Our work in the definition of security policies and the verification that applications comply with them is based on the formal framework for the definition of properties developed in SERENITY [9][10]. This framework is based on the hypothesis that there will be different interpretations of the S&D properties; which in turn requires the ability to allow the definition of

different interpretations that express subtle differences of security properties. There are many reasons supporting this assumption; for instance, legislation, cultural etc. Under this assumption, and in order to guarantee the interoperability of systems based on different interpretations of the properties and on policies fulfilling those properties, we need to be able to exploit the relations between those properties.

4.2.1. Goals and Architecture of S&D Properties Language

Our goal for the property specification language was to support (i) the definition of the semantics of different S&D properties; (ii) the analysis of these definitions in order to find relations between S&D properties; and (iii) the use of the relations found in order to flexibly select S&D Solutions suitable for fulfilling the requirements of the system developers. To achieve those goals we decided to split the specification in two different layers, each one with its own language.

On the one hand, in order to be able to create formal and precise definitions for S&D properties, to compare them and reason about them, S&D experts need a language that has no room for ambiguity and that is precise and flexible. For these reasons we have defined the Formal S&D Properties Language (FPL). This language is based on High Order Logic (HOL).

On the other hand, in order to select S&D solutions to fulfill their requirements, system developers do not need to know the formal definition of the properties. Once having identified the required property all they need to know are the relations between properties. Relations found are then expressed in the Operational S&D Properties Language. The OPL describes S&D Properties using an XML specific format.

4.3. Security Policy description

One of the important advantages of the approach presented in this paper is precisely that we can verify that an application complies with a Security policy, with the only requirement that it has been developed following the SERENITY approach. This can be useful even if it does not use SERENITY to provide runtime support it only used to develop the application and verify the policy compliances. As it is explained in previous sections, two differentiated parts compose the SERENITY framework. The SDF (Serenity Developed-Time Framework) and the SRF (Serenity Run-Time Framework).

Applications developed following the SERENITY approach are designed to make use of the security and dependability mechanisms provided by SERENITY. In order to do that, the developer, includes references to S&D solutions in the applications' source code. As has been introduced in previous sections, security solutions are well-defined mechanisms that provide one or more S&D Properties such as confidentiality, integrity, availability, etc. Here, an important point is the aforementioned semantically point of view of the S&D properties, which allows several interpretation of the same property to coexist and introduces the possibility of relating the properties.

The S&D properties and the Security policies are closely related because the latter are expressed in terms of the former. That is, Security policies contain information about the S&D Properties that must be ensured for each type of information asset. These rules are composed by one or several types of objects and the S&D properties that must fulfil be applied to that object.

Now, focusing on their syntax, Security policies are digitally signed XML documents. At this level, an Security Policy is composed by: (i) *SandDPolicy*: This field includes the policy name, domain, version; (ii) *InformationalPart* section contains the creator field, a label with the main objective of the policy, and a field describing rules. These rules relate properties to types of computational assets; and (iii) The *operationalPart* section is composed by two elements: a list of trust mechanisms, and the artefact validity period of time.

5. Assurance and compliance of Service-oriented systems to of Security policies

When we focus on Service-based Systems assurance and compliance are not so popular terms. In fact, assurance and certification of service-based systems is a difficult task, especially in the future open scenarios. Currently, S&D assurance is strictly intertwined with development process. Assurance is mainly carried out during analysis and coding phases, putting heavy constraints on the needed design documentation and models. In this way, assurance results are only valid for the system and its environment planned at design time, making this approach unsuitable for systems that are meant to continuously evolve (for instance AmI environments).

5.1. Ensuring compliance of service-oriented systems with dynamic selection of services

S&D solutions are described by means of a set of S&D Artefacts (S&D Classes, S&D Patterns and S&D Implementations). In fact, the specific WSs themselves are considered SERENITY's ExecutableComponents and are therefore described as S&D Implementations,

while the other artefacts provide support for applications in order to use these components.

Figure 1 shows the SERENITY model of running systems. In this model there is a security management component aforementioned SERENITY Runtime Framework (SRF). In our case, this component can be realized by a component on the same security domain as the application or even by a WS. In any case we assume that the communications between the application and the SRF are appropriately secured using the mechanisms described in [11]. One interesting aspect to highlight and one of the main elements of SERENITY is the monitoring of the operation of the system. In this way, SERENITY follows the emerging trend towards supporting transparency in distributed components, which is the key to establish the necessary trust in these components and in the systems that use them.

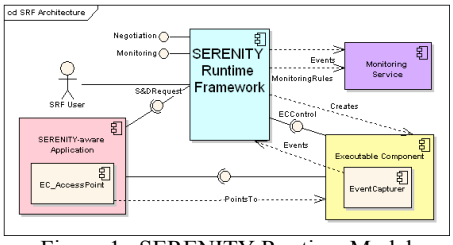


Figure 1. SERENITY Runtime Model

In the current situation, WS-based applications that allow some degree of dynamic selection and replacement of the WSs rely on a discovery or directory service to locate the services they need (using UDDI or some other mechanism). However, the selection and access to the WS ignores security aspects. In our architecture, the SRF provides a security-aware selection of the services. Therefore, instead of contacting the usual directory service, applications request services to the SRF by means of S&D Requests. These requests are extremely flexible and allow developers to choose the degree of freedom given to the SRF in the selection process. S&D Requests can be as generic as requesting an S&D service (such as a Access Control service), more restricted by requesting any implementation of a particular solution (e.g. RBAC), or very specific by requesting a specific implementation (e.g. Oracle RBAC). Please note that, although in the latter case there is no dynamic selection or replaceability, the model still provides advantages since on the one hand provides means for monitoring the operation of the WS and on the other hand the use of S&D Requests constitutes the basis for the verification of compliance of applications with Security policies.

Consequently, in the application code, calls to the discovery service are replaced by calls to the SRF, while direct calls to the WSs are replaced by calls to WS implemented in the form of SERENITY ExecutableComponents.

In order to verify that an application complies with a Security Policy we must consider that Security policies are described on the basis of S&D Properties. Then, S&D Requests contain references to S&D Artefacts, which in turn refer to the S&D Properties they provide. Therefore, in order verify that an application complies with a Security Policy we must verify that all S&D Artefacts used by the application (which can be discovered by inspecting the S&D Requests contained in the application code) refer to the properties defined in the Security Policy.

The proposed model has two important advantages: (i) it supports the dynamic replaceability of the S&D solutions needed by the application, and the monitoring of the correct execution of these solutions; (ii) it allows us to verify the compliance of the application to a policy even if we do not know a priori which WSs are going to be available and selected at runtime.

6. Motivating example – technical view of a specific scene

In order to illustrate the proposed model and the relation of the different components of the system, we take up again the motivating example shown in section 2. As is aforementioned in the scenario description, in the application at several times are requests security properties as confidentiality or authenticity. In this case, as the application is a Serenity-aware application (develop in SERENITY framework) and containing an S&D Request when is required a S&D property. Figure 2 shows an extract of a very simple Serenity-aware application, which required tree S&D properties (no repudiation on received message, on fulfillment of a goal and availability of least a principal) represented by the MedicalAssitance DoctorSearching.Serenity.org S&D Pattern. There are two points to highlight:

- The S&D Request (call to the SRF) to request the selection of an ExecutableComponent implementing the MedicalAssitance_DoctorSearching.Serenity.org S&D Pattern. The S&D Artefact (in this case S&D Pattern) requested is included as a parameter of this S&D Request.
- The calls to the operations of the ExecutableComponent. In this specific application these calls get a medical center key (getMedicalCenterBLOBKey).

In order to verify the Security policy compliance by the application this two points are checked and compared with the information provided by the Security policy.

```
...
public SerenitySimpleApplication() {
        SRF_AP_AccessPoint mySRF = new SRF_AP_AccessPoint(this);

    try {
     myEC = new SerenityExecutableComponent_AP(mySRF, "P:
MedicalAssitance_DoctorSearching.Serenity.org ",null,null);
        argsListE = new TreeMap<String, Serializable>();
    argsListE.put("Text", "This is a string test");

    Thread.sleep(5000);
    eb = myEC.callOperation("getMedicalCenterBLOBKey", argsListE)
...
```

Figure 2. Extract of Serenity-Aware application

The Security policy is expressed on the basis of the S&D Request, which contains an S&D Artefact, which in turn refers to an S&D Property. In order to verify that the application complies with the policy, we need to check the information of the S&D pattern to verify that the S&D property provided is the same as the S&D property mandated by the Security policy. Figure 3 shows the MedicalAssitance DoctorSearching. Serenity.org s pattern specification, in which we can see that the properties provided are: Non repudiation On Receiving Message, Non repudiation On Fulfillment Of A Goal and Availability Of At Least One Principal.

```
  ...
<informationalPart>
    <creator>
        <name>MedicalAssitance_DoctorSearching.Serenity.org
      </name>
        <date>12-12-2008</date>
    </creator>
    <label>provides a security solution that complies with the Medical
Assistance.Serenity.org </label>
    <providedProperties>
        <property>
          <name>NonrepudiationOnRecivingMessage</name>
            <domain> Serenity.org </domain>
            <version>1.0</version>
            <timestamp>-7875</timestamp>
        </property>
        <property>
        <name>NonrepudiationOnFulfillementOfAGoal</name>
            <domain> Serenity.org </domain>
            <version>1.0</version>
            <timestamp>-78756</timestamp>
        </property>
        <property>
          <name>AvailabilityOfAtLeastOnePrincipal </name>
            <domain> Serenity.org </domain>
            <version>1.0</version>
            <timestamp>-787665</timestamp>
        </property>
        </providedProperties>
...
```

Figure 3. Extract of S&D Pattern

Finally, the figure 4 shows the Security policy document, including the information of the rules that the application will have to comply with.

```
...
<Rules>
    <Rule>
    <!-- This rule is applicate over all the object of the      Type: byte[] -->
        <Property>
      NonrepudiationOnRecivingMessage.serenity.org
        </Property>
        <Object>
            byte[]
        </Object>
    </Rule>
</Rules>
...
```

Figure 4. Extract of Security Policy

7. Discussion and conclusion

Recent conclusions of ESFORS group of experts declare, in order to achieve appropriate guarantees for the security of complex systems in open communication environments, the state of the art of development processes for secure systems has to be considerably improved. Such improvements should include methods for precise specification of security policies. In accordance with these recommendations, in this paper we have presented a framework for the verification of the compliance with security policies, based on the SERENITY model of secure and dependable systems and its application to the Service-Oriented Computing. We have exposed a practical scenario of serenity-aware application in which we have shown the technical aspect of this and we can highlight several improved over the traditional methods.

8. References

[1] Maña, A.;Rudolph, C.; Spanoudakis, G.; Lotz, V.; Massacci, F.; Melideo, M. & López-Cobo, J. M.Security (2006). Engineering for Ambient Intelligence. IDEA Group.

[2] Bell, D. E. & La Padula, L.J. (1973). Secure Computer Systems: Mathematical Foundations, MTR–2547, Vol. I, The MITRE Corporation, Bedford.

[3] Brewer, D.F.C & Nash, M.J. (1989). The Chinese Wall security policy. Security and Privacy, 1989. Proceedings., 1989 IEEE Symposium on (pp. 206-214). Oakland, CA, USA.

[4] Ferraiolo, D.F. & Kuhn, D.R. (1992). Role Based Access Control. 15th National Computer Security Conference. (pp. 554-563).

[5] Biba, K. J. (1977). Integrity Considerations for Secure Computer Systems, MTR-3153. Technical Report TR-3153, MITRE Corp, Bedford, MA.

[6] Spanoudakis, G.; Maña A.;, Kokolakis, S:; Rudolph, C.; & Lotz V. (2009). The Serenity Book. Springer-Verlag.

[7] Maña, A. & Pujol, G. (2008). Towards Formal Specification of Abstract Security Properties. Proceedings of The Third International Conference on Availability, Reliability and Security. IEEE Computer Society Press. Barcelona, Spain.

[8] Maña, A.;Sánchez, F.; Serrano, D.& Muñoz A. (2006b). Building Secure Ambient Intelligence Scenarios. Eighteenth International Conference on Software Engineering and Knowledge Engineering (SEKE06). San Francisco Bay, USA.

[9] Pujol, G; Gurgens, S. Specification of S&D Properties. SERENITY model of secure and dependable system. Springer press (2008).

[10] Maña, A; G. Pujol. Modelo para la Formalización Abstracta e Intuitiva de las Propiedades de Seguridad. RESCI, X Reunión Española sobre Criptología y Seguridad de la Información, Springer-Verlag LNCS, Salamanca España (2008).

[11] Maña, A.;Presenza, D.; Piñuela, A.; Serrano, D.; Soria, P. & Sotiriou, D. (2007). Specification of SERENITY Architecture. Serenity Public Report A6.D3.1

Software-Only M-JPEG Performance for Emerging Multimedia Applications

Varun Kulkarni and Jonathan C.L. Liu
Dept. of Computer, Information Science and Engineering
University of Florida, Gainesville, FL 32611
{vak, jcliu }@cise.ufl.edu

***Abstract*—Distributed multimedia applications, especially videos over Internet Protocol (IP) are gaining popularity due to their ability to convey more information than text and images. Video on Demand, Video Conferencing, and surveillance through videos are some of the popular examples. Satisfactory Quality of Service (QoS) of such videos is of paramount importance, and software-only M-JPEG is critical for achieving good QoS. Via the analysis on emerging multimedia-based applications, this paper performs an experimental study to ensure that Motion JPEG (M-JPEG) and Motion JPEG 2000 are reliable choices to ensure quality of service with the scenarios of multiple encoders and decoders. The QoS assurance of M-JPEG is studied with respect to Archiving abilities, IP Network adaptability and Video Processing efficiency. Based on experimental analysis, we further propose a way of improving QoS by reduction in startup delay for play back videos in distributed environment. Further, we also propose feasibility for pure software solution for IP based surveillance cameras.**

Keywords - Motion JPEG, QoS Assurance, Performance Study, Video Surveillance.

I. INTRODUCTION

JPEG is an image compression standard created by the standard committee called Joint Photographic Experts Group in early 1990s. The standard provides lossy and lossless modes for compression. The lossy mode follows a Discrete Cosine Transform (DCT) - based encoding approach, while the lossless mode makes use of predictor to provide better quality than lossy mode. JPEG 2000 is extension to the JPEG standard with some of the vital features like low-bit rate compression performance, spatial access to the bit stream, Pan and Zoom with optimal decompression techniques, Region of Interest encoding. However, since JPEG 2000 follows a complex encoding strategy as compared to JPEG; JPEG has dominance in low complexity applications at suitable bit-rates [4].

It is the human nature to extend the JPEG standard to support video streaming. The idea is to apply JPEG continuously on the consecutive images quickly enough such that motion of video can be observed. Motion JPEG (M - JPEG) is thus proposed as another standard that makes use of JPEG (usually lossy) compression intra-frame encoding. Since the M – JPEG compresses every frame of the video individually and stores them collectively, which means that decoding can be applied to each frame independently of other frames, unlike MPEG, which follows an inter-frame encoding approach. Due to this approach by Motion JPEG, the quality of the image is directly and statically dependent on the nature of complexity of the frames data. This suggests the frames with monotone patterns, indicating smooth-transitions over set of frames, achieves more compression than the frames with complex patterns.

With the rapid improvement of CPU speed in last two decades, M-JPEG implementation has gradually shifted from hardware-only solutions (e.g., Parallax M-JPEG video cards) to software-only solutions. The trend is supported by many fronts of industrial evolutions: (1) computers have moved into multi-core CPUs design; (2) the communication bandwidth available to users have reached tens of Mbps; and (3) M-JPEG indeed delivers higher quality compared to MPEG series of standards. We believe the trend will continue, and more applications will be built on software-only M-JPEG solutions.

We particularly focus on the need of high-quality video communication. For instance, it has been reported that, due to the low quality, the suspect's face on decompressed images of surveillance video cannot be recognized. We then search for the generic models of interaction for the emerging application scenarios. For instances, young generations now go beyond one-to-one video conferencing and experiment the one-to-few video conferencing. It means the computer needs to do encode-one and decode-few on concurrent M-JPEG streams. Similarly, video surveillance applications will require the support of decode-many M-JPEG streams at the same time. On the contrary, the trans-coding servers providing different stream formats need the support of encode-many M-JPEG streams simultaneously.

To support this kind of concurrency on the same computer, there is no doubt that the quality of service of the M-JPEG streams will be affected. In this paper, we start addressing the investigation on two main QOS factors: start-up delay and achieved display frame rates. After studying the characteristics of JPEG and M-JPEG standards and performing enough experiments, we think that (software-only) M-JPEG and Motion JPEG 2000 are fairly good choices for our research investigation.

We thus performed various experiments to ensure the characteristics of JPEG 2000 and Motion JPEG for these two QoS factors. The experiments are conducted based on the open source codec made available at open-jpeg community [7]. Rest of the required experimental setup was made by coding test beds in GNU C language. The servers and client hosts, used to

simulate the distributed environment in the experiments are Ubuntu Linux version 9.10 machines with kernel 2.6. Experiments have revealed facts supporting Motion JPEG as a fair option to be considered for multimedia videos over IP networks.

We found that Motion-JPEG 2000 indeed outperforms MPEG standards for e.g. in complexity of processing and reduction in startup delays for play back videos. It is true that MPEG does require less network bandwidth to transmit the data as compared to M-JPEG. However, M-JPEG is proved to achieve the much lower group delay and get rid of high buffering requirements for groups of frames to operate on. The performance of M-JPEG codec in terms of compression ratios was appreciable. Compression ratios up to 45:1 were observed with tolerable loss of video quality. Although the encoding performance is restricted to approximately 19.1 frames per second (fps), higher compression ratio facilitates enhanced decoding performance up to 55.70 frames per second (fps). The decoder performance in IP based networked environment was notable as it performed with an average of 20.12 frames per second (fps).

The rest of the paper is arranged as follows: Section 2 makes the analysis of QoS guarantees from emerging applications in Archiving, Network adaptability and Video processing efficiency of Motion JPEG aspects. Section 3 proposes a new approach towards reducing the startup delay for play back videos using Motion JPEG. Section 4 provides the performance results for pure software-based solutions for different interaction models. The paper concludes by providing an opinion regarding Motion JPEG as a fair option for videos and possible further research in this area.

II. Analysis of Quality Of Service Assurance By Motion JPEG

This section makes detailed analysis of M-JPEG's ability to provide QoS guarantees from three different aspects namely archival capability, adaptability towards IP networks and video processing. The analysis shows that M-JPEG display good characteristics for all the three aspects.

A. Quality of Service – Archiving

As we know the original data are too large to be stored, the compression algorithm, along with the reduction in original data should also specify the archival characteristics. The compression ratios obtained by M-JPEG encoding are good enough. M-JPEG facilitates efficient archival of lossless compression which cannot be achieved by the MPEG compression strategies. Long term preservation of the video data is also the topic of concern which demands a generic format for storage of video. As far as lossless archival is concerned Motion JPEG remains a strong choice among existing compression techniques. JPEG 2000 family of standards also provides efficient ways for storing metadata, which makes the archival supportive for efficient data processing [6].

The open-JPEG codec allows specifying the rate parameter which helps in setting the desired compression ratio. However, more the compression ratio more is the loss of information and hence lossy compression is observed. Some of the compression ratios achieved on a single machine are tabulated as the following TABLE 1. It is clear that a machine perhaps need to run several M-JPEG encoding processes in order to handle the archiving applications.

TABLE I. Achievable Compression Ratios with M-JPEG codec

Original Size -YUV data (MB)	Size after Encoding - MJ2 data (MB)	Compression Ratio
23.48	2.30	10:1
27.89	1.39	20:1
30.412	1.21	25:1
38.015	1.72	22:1
47.14	2.61	18:1
51.02	3.18	16:1
55.68	2.65	21:1

B. Quality of Service - Network Adaptibility

Most of the prevailing applications today make use of TCP/IP networks for transferring multimedia data. Multimedia applications as we know are sensitive towards transmission errors, since a transmission error requires the sender to resend the required data. This causes jitters and long delays in the play back at the receiving side affecting the QoS adversely. The situation can be worse when the video frames are encoded using inter-frame encoding approach, because the dependent frames are required to be retransmitted by the sender. M-JPEG video data being intra-frame encoded is fairly resilient towards transmission errors. The MPEG data suffers greatly as compared to the M-JPEG in presence of transmission errors. Since M-JPEG supports software based processing solutions, an application level rate adaptive transmission [3] technique can help adapt to the dynamics of the IP networks to provide better quality of service with respect to high quality image transmission. Such a strategy enhances the M-JPEG performance not only over TCP but also over UDP.

Further M-JPEG does not impose strict requirements on the bandwidth with which the video data should be transmitted for good QoS. The flexibility provided by M-JPEG with respect to the frame size, the quality of compression and the frame rate parameter makes M-JPEG adaptable to given network configurations to meet a better quality of service. As M-JPEG follows simple approach towards coding, it supports audio synchronization with use of time stamps over the congested TCP/IP network networks. M-JPEG remains a good choice for better QoS as compared to UDP networks. In M-JPEG, optimal time required for the coding process also ensures better timing guarantees for quasi real time networked applications [1].

M-JPEG can also perform well for the wireless environments. A JPEG 2000 Wireless System Architecture for WLAN has been proposed as JPWL encoding and decoding standard. Special data protection schemes proposed in the standard can help improve the performance of single frame and multi frame applications. A detailed study of the M-JPEG stream over various versions of TCP would be the next step in evaluating the performance of M-JPEG from QoS perspective [1]. We performed experiments with open-JPEG codec over wireless networked environment on the university campus with IEEE 802.11a and 802.11g network topology. The experimental methodology focused on single server - single client architecture. The server machine, upon client request encoded data completely (YUV to MJ2 encoding) and sent the encoded frames over wireless network to client host. The client machine then decodes (MJ2 to YUV decoding) the received data, frame by frame. Client and the server machine were Linux machines with Ubuntu 9.10 distribution and kernel 2.6. Encoding and decoding were the major processes running on these machines. The average frame size is 152.31 KB. Decoder performance was measured in terms of frames decoded per second (fps). Appreciable encoding and decoding performance is observed as shown in Table II.

TABLE II. ACHIEVABLE M-JPEG CODEC PERFROMANCE IN WIRELESS NETWORK ENVIRONMENT

Server - Encoding (YUV - MJ2) Performance (fps)	Client - Decoding (MJ2 - YUV) Performance (fps)
18.4	18.6
18.7	18.4
18.5	18.5
18.9	19.2
19.2	18.4
19.1	18.7

C. Quality of Service – Video Processing Efficiency

Coding the video data efficiently so that it can be further processed efficiently at real time rates is also a good indication of providing QoS. M-JPEG display these characteristics in the encoding and decoding processes. It achieves good coding performance while keeping the encoding and decoding process simpler. Pure software based solutions are feasible to speed up the coding/decoding process [2]. Efficient hardware and software based feasible algorithms have been proposed and implemented in order to make M-JPEG processing faster [2]. JPEG 2000 has a considerable improvement over JPEG as far as the processing performance is concerned. JPEG 2000 supports various types of images such as bi-level, gray-level, multi-component etc. It suits really very well for Distributed Multimedia Applications.

Motion JPEG 2000 which is an extension to the JPEG2000 standard inherently show appreciable performance for video data processing. At also performs well at lower values of the bit rates. Further, it is also immune to the bit-errors that might occur due to lossy network channel and thus is provides considerable advantage over other encoding standards. Motion JPEG 2000 is based on the concept of Discrete Wavelet Transform. A pure software based implementation of Wavelet based video codec has been made feasible and shown that Motion JPEG 2000 gives better video processing performance as compared to Motion JPEG standard. Motion JPEG 2000 out performs Motion JPEG (M-JPEG) and MPEG2 standard. Better compression ratios are also obtained as compared to Motion JPEG.

Using the open-JPEG codec we experimented with the maximum achievable decoding performance on a single host with lossy encoding. The loss of quality was tolerable and we found that decoder can perform well with performance figures in Table III. The host in the experiment was Linux machine with Ubuntu 9.10 distribution and kernel 2.6. Encoding and decoding were the major processes running on these machines. The performance was measured in terms of number of decoded frames per second (fps). Encoding comprised of conversion of YUV data frames to compressed MJ2 data frames. The decoding comprised of conversion of MJ2 to YUV data frames. Note that, while conventional video conferencing only requires one-encoder-with one-decoder interaction, the performance results indicate that multiple-decoder scenarios can be feasible.

TABLE III. ACHIEVABLE M-JPEG VIDEO PROCESSING PERFROMANCE IN SINGLE HOST ENVIRONMENT

Compression Ratio in single host environment	Encoding Performance (frames per second)	Decoding performance (frames per seconds)
1:50	19.25	55.7
1:40	18.9	52.1
1:45	19.13	49.21
1:30	19.11	48.78
1:35	18.8	47.38
1:20	19.02	45.12
1:10	19.21	35.43
1:05	18.8	26.13
1:03	19.13	19.12

III. REDUCTION IN START-UP DELAYS

Conventionally, a host requests for multimedia data to other host in a networked environment. The provider/sever encodes the video data on demand dynamically or sends statically encoded data over the network, upon request from the client. In case of inter-frame compression schemes the pattern and sequence of frames are important from the decoder's perspective and should be sent in corresponding order to help decoder improve performance and deliver required QoS. However, the decoder cannot initiate the decoding process, unless it receives the required set of frames in the order for processing. This introduces a startup delay in the playback application at the client side with standards following inter-frame encoding. Startup delay is measured as time between the client sending request to the server and the actual video play back initiation at the client. This further raises a concern of minimum buffer size at the client side. Usually this buffer size

is large for videos compressed using inter-frame encoding strategies. Motion JPEG encoder performs encoding of picture frames independently. This implies that the decoder can decode each frame independently. Using this property the decoder can begin decoding process as soon as the minimum set of encoded frames is available in the decoder's buffer. Unlike MPEG, the decoder does not have to wait for group of frames with a particular sequence to be present in the buffer. This helps in reducing the startup delay for the video at the client end.

A. *Experimental Design and Methodology*

The server and client hosts were Linux machines with Intel dual cores (CPUs), Ubuntu 9.10 distribution with kernel 2.6.32. All the machines involved in the experiments support Symmetric Multiprocessing (SMP). The client host requests the server host for the video data. Upon receiving the request from the client over the network, the server initiates the M-JPEG encoding process. The type of network used in the experiments was IP based university network. During the encoding, YUV data stream is converted to MJ2 compressed data frames. Once an encoded frame is available in the server buffer, the server spawns a separate thread of execution for sending the data frame to client over the network. Thus the encoding performance of server is not halted due to the transmission event. There is one buffer per remote client at server side. The average buffer size at the decoder was 230 KB. On, receiving the first data frame client initiates the decoding process and can proceed for the further play back. Gradually 10 remote clients demand encoded data from a single server and the server performs concurrently to satisfy the client requests. The average frame was 156 KB in size and average processing time per frame was 53.33 milli-seconds at the server. The network delays measured at client side were negligible since the experiments were conducted during low network load. With ideal QoS in terms of Network delays, we measured the startup times at the client side. The observations of the above experiment are shown below:

TABLE IV. START-UP DELAYS MEASURED AT CLIENT SIDE

Number of Encoding requests at Server	Startup Delay (seconds) measured at Client
1	1.89
2	2.52
3	3.11
4	3.58
5	5.66
6	6.78
7	8.21
8	9.67
9	10.79
10	11.88

B. *Analysis of Start-up Delays and System Performance*

In this section we analyze the relationship of the start-up delay with the number of remote clients associated with single server. We further analyze the relationship of the resource utilization with the number of remote clients and hence with the start-up delay. Figure1 shows the trend with startup delay with increasing number of remote clients connected over TCP/IP networks.

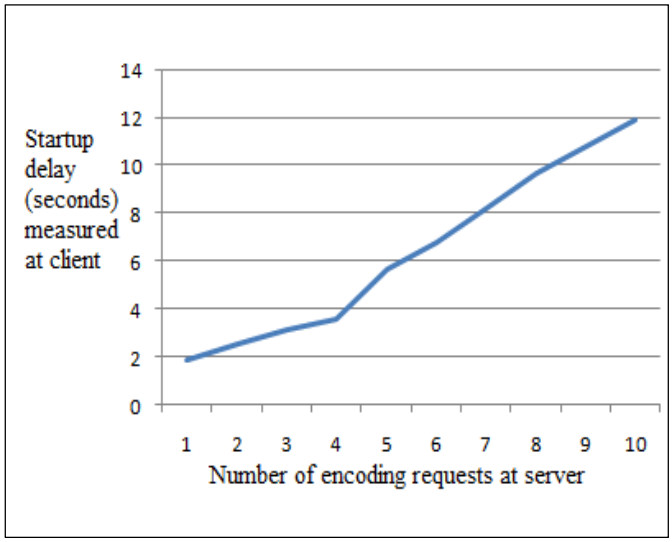


Figure 1. Trend in startup delays with increasing number of clients

Least Start-up delay was recorded with single client. However, the delay increases with increasing number of client requests. Clearly, Single server becomes the bottleneck for increasing client requests for encoding YUV data. If load balanced, as it is mostly done these days with server farms, the startup delay values for multiple clients can be reduced further to deliver performance as that of single client and single server. Although, startup delay increases with increase in number of remote clients, the range of delay values is considerable for providing the QoS assurance over TCP/IP networks. An overall reduction in the startup delay is observed at the client side. Further, the experimental results showed that minimum size of buffer the client required for real time decoding was 230 KB. We further observed the approximate combined system performance of two cores using 'mpstat' - Linux utility while the server machine was busy serving dynamically increasing clients. The %CPU utilization in user mode, system mode, the %IO wait, %interrupts handled increases to high values respectively. Figure 2 shows the %CPU utilization in user mode for two available cores.

The analysis of system performance study revealed that the single machine with 2 cores (CPUs) become processing bottleneck for a single server serving multiple networked connected remote clients. The system memory does not pose a severe bottleneck during the encoding/decoding processes. Load balancing with multiple synchronized servers can help improve the performance.

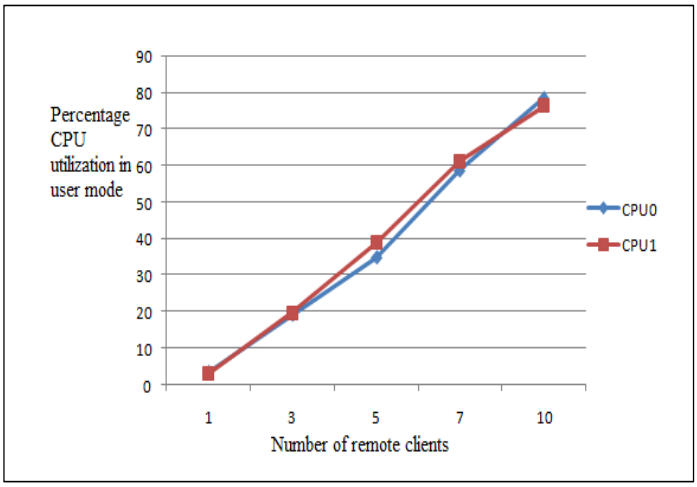


Figure 2. %CPU utilization in user mode when remote clients are served by single server.

IV. PURE SOFTWARE BASED M-JPEG PERFORMANCE

Internet Protocol (IP) based surveillance cameras heavily use the encoding and decoding of video streams. Specialized hardware is necessary to ensure quality of service. Although M-JPEG standard requires higher bandwidth, it facilitates execution and processing with inexpensive hardware equipment, due to its simplicity in the nature of processing. We have performed experiments to confirm whether, pure software based solutions that can execute on inexpensive hardware can be used to provide quality of service with M-JPEG. The experimental methodology focused on two architectures namely clustered clients and clustered servers. Both of these architectures were used with the following methodology - The server sends out the encoded frame to the corresponding client as soon as it is ready in the server buffer and the client initiates a frame decoding as soon it is available in the client buffer. The data from server to client and vice versa was transferred over TCP/IP networks. POSIX compliant multithreading is followed while constructing the test beds.

A. Clustered Clients Architecture

In case of clustered clients, a single host (machine) acts as a client responsible for decoding MJ2 video data streams received from remote servers. Based on the methodology mentioned above, the resources of a single machine were utilized in common by multiple decoding processes. The experiment begins with single remote server sending encoded data as per the methodology, to the client over IP network. Number of servers increase over time. Every remote server makes the client to spawn a separate thread of execution for decoding the data stream. The performance of the client in terms of decoding rate (frames decoded per second - fps) is measured.

The approximate Load averages on the client machine were 0.11, 0.06 and 0.01, which means that the machine had lower loads of allied processes. Every decoding client thread reports its performance with number of frames decoded per second. The average of the performance of currently active clients is measured as average decoding performance. Starting from single remote server we extended the experiment to measure the performance for 16 remote servers and the average decoding performance rate dropped from 19.1 fps for one remote server to 2.5 fps for 16 remote shown in Figure 3.

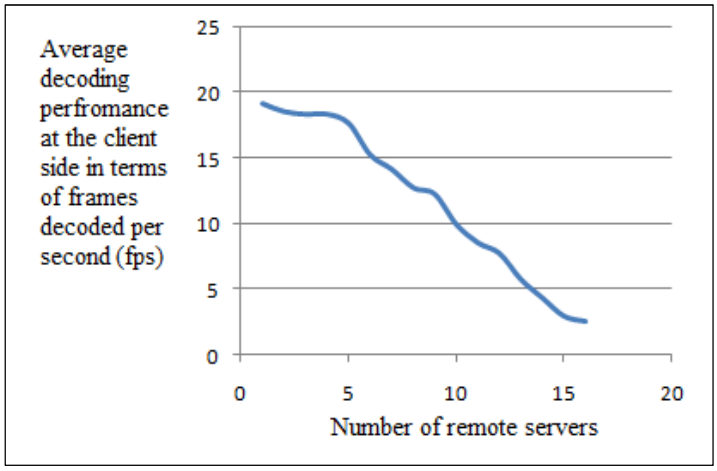


Figure 3. Decoder performance with clustered clients architecture.

From experimental results in Figure 3 it is evident that with increasing number of remote servers, the available processors become a bottleneck for decoder performance at client side. The experiments further revealed that CPU utilizations in terms of number of active processes and the percent of the I/O utilization within the system increased with increase in number of remote servers.

B. Clustered Servers Architecture

In case of clustered servers, the encoding is done on a single machine acting as a server. Here, the experiment begins with a remote client requesting for the encoded data to the server. Server host uses dynamic encoding approach, where the encoding is done on demand upon request from each client. Server randomly chooses a variable sized YUV data on each request. Each request from remote client makes the server to spawn a process for encoding the available YUV data; Server sends out the encoded frame as soon as it is ready in the server buffer. A separate thread for sending the frame to the requesting client is spawned so that the encoding performance remains unhindered. The client machine waits till it receives encoded data from the server. Once the encoded frame is received by the client over IP network, the client begins decoding the received frame as per the mentioned methodology. The decoder performance is then measured as in the previous architecture i.e. number of frames decoded by the client per second. The average decoding performance is the average of the decoding performance reported by each decoding client thread.

During the experiment the client and server machines selected had minimum allied processing load. The load averages at server were 0.07, 0.09, 0.02 and that at client were 0.12, 0.11, 0.13. It is observed that the decoding performance in this scenario decreases from 19.1 fps for single remote client to 0.86 fps for 16 remote clients. The decoding performance for maximum number of clients is poor than that of clustered client architecture as evident in Figure 4.

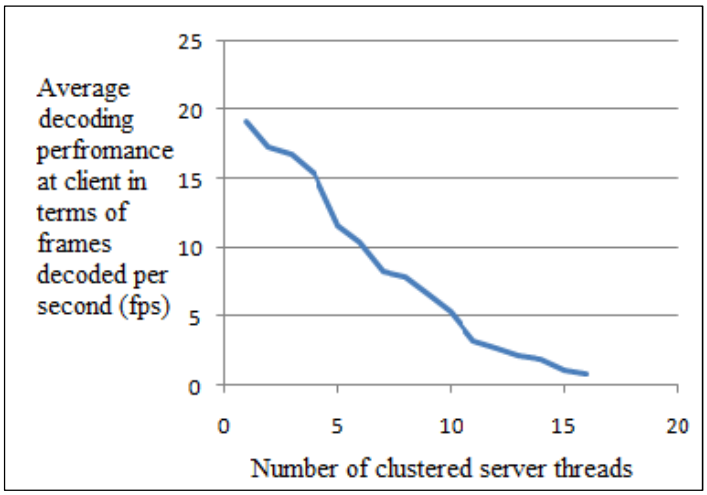


Figure 4. Decoder performance with clustered servers architecture.

With increase in the number of remote clients, the available processors become a clear bottleneck for encoder performance. The experiments further revealed that CPU utilizations in terms of number of active processes and the percent of the I/O utilization within the system for the server machine increased with the increase in number of remote clients. The server clearly has a limit in terms of system resources while encoding and decoding.

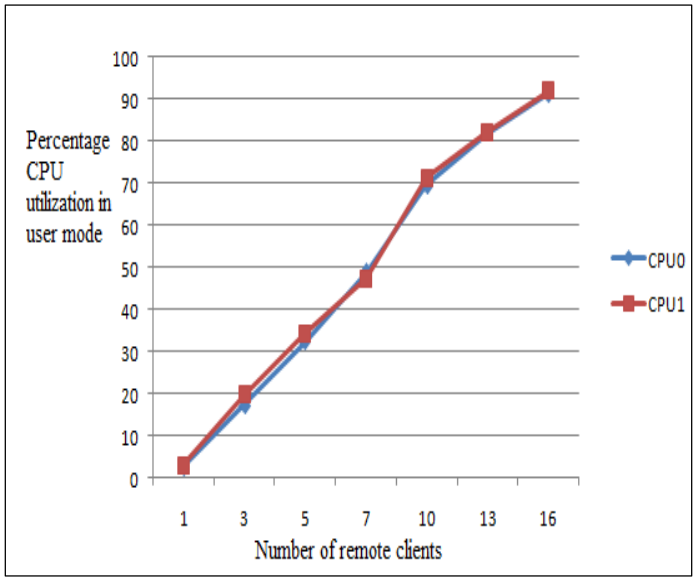


Figure 5. %CPU utilization for clusetered servers

The system performance is observed closely to find the effect on the resource utilization at system level for a pure software based solution. The %CPU utilization in user mode and system mode along with the %IO wait and %interrupts handled increases to high in the system. Figure 5 shows the %CPU utilization in user mode for two available cores.

The CPU and memory utilization increased with increase in number of processes responsible for encoding or decoding. The processes responsible for sending or out the encoded frames at the server or receiving the encoded frames received at the client add negligible overhead to the decoder performance. Scene analysis techniques can further help reduce the requirement of high bandwidth for M-JPEG [5].

V. CONCLUSION

After performing a careful experimental study of JPEG, Motion JPEG, JPEG 2000 and Motion JPEG 2000 standards for QoS for videos and the effect of software based implementation of these standards on system performance, we think that QoS for distributed multimedia video data over IP networks, can fairly be achieved by Motion JPEG standard. Pure software based solution for IP based surveillance cameras is feasible with Motion JPEG [5].

For lower bandwidth requirements Motion JPEG 2000 remains a better option than Motion JPEG. The performance study does show that some QoS factors that M-JPEG is capable of providing, cannot be satisfied by MPEG formats. Future study can be extended to evaluate the emerging domain of Parallel and High Performance Computing for better QoS delivery for Multimedia applications. Exploiting the parallelism introduced in the coding and decoding methodologies by the Motion JPEG standards can be applied at least in two ways. One way is to exploit the parallelism with the help of GPU Architecture, and the other is to scale the multi-core CPU design to a large number of processors.

REFERENCES

[1] Bing Wang, Jim Kurose, Prashant Shenoy, Don Towsley, "Multimedia Streaming via TCP: An Analytic Performance Study", ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), Volume 4 , Issue 2 (May 2008) Article No.: 16 , 2008.

[2] Brian C. Smith, "Fast Software Processing of Motion JPEG Video", Proceedings of the second ACM international conference on Multimedia, San Francisco, California, United States, Pages: 77 - 88, 1994.

[3] Duminda Nishantha, Yukuo Hayashida, Toshihiro Hayashi - "Application Level Rate Adaptive of Motion-JPEG Transmission for Medical Collaboration Systems", Proceedings of the 24th International Conference on Distributed Computing Systems Workshops, Volume 7, Pages: 64 – 69, 2004 .

[4] Gregory K. Wallace, "The JPEG Still Picture Compression Standard" , Communications of the ACM Archive, Volume 34 , Issue 4, Pages: 30 - 44, 1991.

[5] Jérôme Meessen, Christophe Parisot, Xavier Desurmont and Jean-François Delaigle, "Scene Analysis for Reducing Motion JPEG 2000 Video Surveillance Delivery Bandwidth and Complexity" - Book: ICIP (2005), pages: 577-580

[6] Pearson, Glenn and Michael Gill, “An Evaluation of Motion JPEG 2000 for Video Archiving”, Proc. Archiving 2005 (April 26-29, Washington, D.C.), IS & T (www.imaging.org), pp. 237-243.

[7] http://www.openjpeg.org

A 3D Curve Chain Code Representation for the Analysis of Human Motion Similarity

Francisco Torres[1], Edward Chow[2]

Computer Science Department, College of Engineering and Applied Science, University of Colorado at Colorado Springs
1420 Austin Bluffs Parkway, Colorado Springs, CO, USA
[1]ftorres@eas.uccs.edu
[2]chow@eas.uccs.edu

Abstract*— Performing human motion analysis requires the interaction of multiple disciplines, and its range of applications is extensive: in clinical gait analysis, the analysis of human motion goes from understanding the multiple mechanisms that translate muscular contractions about articulated joints into functional accomplishments, i.e. standing, walking, to a best appreciation of the relationships between the human motion control system and gait dynamics for the planning of treatment protocols, i.e. surgical intervention, orthotic prescription; in sports, analysis techniques like osteopathic evaluation and treatment are used to improve performance while avoiding injuries; in dance, the analysis and identification of motion qualities to extract emotion from a given performance by establishing relations between effort factors -space, weight, time, flow-, and their corresponding dynamic factors -force, stiffness, inertia, gravity-. This paper considers the motion of the human body as a generator of 3D curves, using the orthogonal changes of direction of those curves as a chain code representation for the human motion. Mapping human motion into chain codes allows the use of sequence alignment algorithms, i.e. Needleman-Wunsch algorithm, to implement the analysis of human motion similarity.

Keywords— **Human motion similarity, chain codes.**

I. Introduction

Performing human motion analysis requires the interaction of multiple disciplines, and its range of applications is extensive: in clinical gait analysis, the analysis of human motion goes from understanding the multiple mechanisms that translate muscular contractions about articulated joints into functional accomplishments, i.e. standing, walking, to a best appreciation of the relationships between the human motion control system and gait dynamics for the planning of treatment protocols, i.e. surgical intervention, orthotic prescription [1]; in sports, analysis techniques like osteopathic evaluation and treatment [2] are used to improve performance while avoiding injuries; in entertainment, an animation technique named rotoscoping has been used to give life to non-human characters based on human motion [3] ; in dance, the analysis and identification of motion qualities to extract emotion from a given performance by establishing relations between effort factors -space, weight, time, flow-, and their corresponding dynamic factors -force, stiffness, inertia, gravity- [4].

To perform the analysis of human motion in such different scenarios, different approaches are used and they have been developed to meet their particular requirements: high data acquisition for sports [5], or real time tracking for virtual reality applications [6]. In general, these approaches include motion analysis data collection protocols, measurement precision, and data reduction models [7].

Even knowing that human body parts are not rigid structures, human body is often considered as a system of rigid links connected by joints when analysing human motion: lengths of links are measured and positions of joints are calculated as a pre-processing step. After that, the human motion analysis begins: the calculation of forces, velocities, and accelerations for specific parts of the human body takes place and those are then compared against the expected results. A drawback of this approach is the fact that during human motion markers worn by the performer will not remain in their original position due to the flaccid condition of the human body. This work avoids the pre-processing step and, consequently, the human body representation as rigid links connected by joints.

The organization of the paper is as follows: Section II covers the chain code representation for 3D curves and its implementation for human motion. Section III explains the definition of similarity used in this work. In Section IV, the algorithm that implements the analysis of human motion similarity is described. Section V shows the experiments and results. The conclusions and future work are discussed in Section VI.

II. Chain Code Representation for 3D Curves

Markers worn by the human body generate 3D curves during human motion, as shown in Fig. 1.

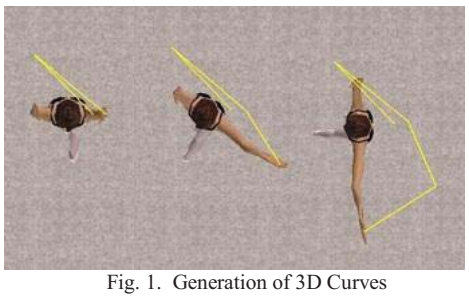


Fig. 1. Generation of 3D Curves

Markers set on the same position in the human body generate 3D curves of similar shape for similar human motions.

Object shape is an important feature of images and it is used in content-based image retrieval [8]. A suitable shape representation and similarity measure for content-based image retrieval should meet the following criteria [9]:

- The representation of a shape should be invariant to scale, translation, and rotation
- The similarity measure based on the shape representation should conform to human perception; i.e., perceptually similar shapes should have high similarity measure
- The shape representation should be compact and easy to derive, and the calculation of similarity measure should be efficient

Several techniques for shape description or representation have been proposed and used: chain code [10], polygonal approximations [11], curvature [12], Fourier descriptors [13], moment descriptors [14], etc. Reference [15] describes a chain code representation that is invariant to scale, rotation, and translation. This representation is also unique for every 3D curve, in the logic that a 3D curve is represented by only one chain code.

In the Motion Capture session, the 3D position for each marker worn by the user is recorded on each sampling time. The set of all 3D positions for each marker represents the shape of the 3D curve generated by them. Reference [15] assumes a length of 1 for each vector that represents the orthogonal change of direction, as shown in Fig. 2.

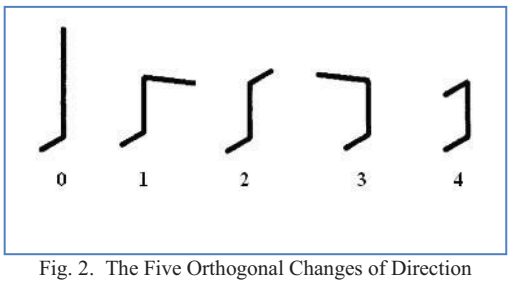


Fig. 2. The Five Orthogonal Changes of Direction

Formally, if the consecutive sides of the reference angle have respective directions u and v, and the side from the vertex to be labeled has direction w, then the chain element is given by the following function,

$$\textbf{chain element } (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \begin{cases} \textbf{0}, \text{ if } w = v; \\ \textbf{1}, \text{ if } w = u \text{ x } v; \\ \textbf{2}, \text{ if } w = u; \\ \textbf{3}, \text{ if } w = -(u \text{ x } v); \\ \textbf{4}, \text{ if } w = -u; \end{cases}$$

where x denotes the cross product.

In our implementation, the length of these vectors is a variable labeled as ***step size***. Also, an extra character that represents the lack of motion was included, for the case of having a distance -between two consecutive markers- less than ***step size***.

The following pseudo code describes the algorithm that generates the chain codes for each marker:

```
GenerateChainCode(list3Dpoints)
  Create orthogonal vectors u and v
  While there are 3D points in list3Dpoints
    A = this 3D point
    B = next 3D point
    If distance between A and B < step size Then
      chainCode = chainCode + '.'
      While there are 3D points in list3Dpoints
        B = next 3D point
        If distance between A and B > step size Then
          break loop
        EndIf
        chainCode = chainCode + '.'
      EndWhile
      If no 3D points in list3Dpoints Then
        break loop
      EndIf
    Else
      While intersection of AB in any of the current cube's planes
        Create vector w
        If w = v Then
          chainCode = chainCode + '0'
        Else If w = u x v Then
          chainCode = chainCode + '1'
        Else If w = u Then
          chainCode = chainCode + '2'
        Else If w = - (u x v) Then
          chainCode = chainCode + '3'
        Else If w = - u Then
          chainCode = chainCode + '4'
        EndIf
        u = v
        v = w
        A = point of intersection of line AB and the plane
      EndWhile
    EndIf
  EndWhile
```

The following is an example of a chain code representing a segment of a 3D curve for a given marker:

2003..200000..22010220..

III. Definition of Similarity

Similarity is an important concept in our work as it is in several other fields. Several similarity measures have been proposed: information content [16], dice coefficient [17], etc. Michael McGill surveyed and compared 67 similarity measures used in information retrieval [18].

A problem with previous similarity measures is that each one of them is tied to a particular application or assumes a particular domain model. An information-theoretic definition of similarity [19] that achieves universality -as long as the domain has a probabilistic model- and theoretical justification -the similarity measure is not defined directly by a formula, but from a set of assumptions about similarity- states the follows:

- **Intuition 1**. The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.

- **Intuition 2**. The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.
- **Intuition 3**. The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

We follow this definition of similarity to analyse human motion similarity, as explained in the next section.

IV. Analysis of Human Motion Similarity

After the step described in the previous section, we have the human motion represented as a set of chain codes. If human motions are intended to be similar and markers are set on the same position on the human body, chain codes of markers placed on the same position must be similar.

Needleman-Wunsch global alignment algorithm is used to analyse the human motion similarity. This algorithm aligns two sequences by inserting some gaps at locations that are unknown. We used the Needleman-Wunsch algorithm with values of 1 for match, 0 for mismatch, and 0 for gap penalty. The following is the algorithm in pseudo code used to analyse the human motion similarity.

```
HumanMotionSimilarity(hm1, hm2)
  For each marker on hm1 and hm2 Do
    seq1 = chain codes for this marker on hm1
    seq2 = chain codes for this marker on hm2
    Needleman-Wunsch(seq1, seq2, seq1_al, seq2_al)
    /*Intuition 3*/
    If seq1_al = seq2_al Then
      similarity_value = 1.0
    Else
      /*Intuition 1 and 2*/
      seq_length = length_of(seq1_al)
      For i = 1; i <= seq_len Do
        If seq1_al[i] = seq2_al[i] Then
          sim = sim + 1
        EndIf
      EndFor
      similarity_value = sim / seq_length
    EndIf
    sum_sim_value = sum_sim_value + similarity_value
    counter = counter + 1
  EndFor
  sum_sim_value = sum_sim_value / counter
```

If chain codes are the same, the human motion similarity value is 1. This corresponds to Intuition 3.

If chain codes are not the same, Needleman-Wunsch algorithm will insert gaps as needed until the optimum sequence alignment is found. This corresponds to Intuition 1 and 2.

V. Experiments and Results

The following four key rehabilitation exercises were captured in a Motion Capture Laboratory facility:

- Mini-squats: muscle group - hips
- Posterior pelvic tilt: muscle group - abdominal, pelvis
- Shoulder elevation with rotation: multiple muscles around the scapula and shoulder
- Standing hip ABD (abduction): muscle group - hips

Each one of those key rehabilitation exercises was performed three times. Data sets for those rehabilitation exercises were provided in C3D format. We implemented a software application that takes those data sets and simulates the sensors' motion, along with the generation of chain codes for them. This software application also plays human motion videos, if any, for a better understanding of the sensor's simulation as shown in Fig. 3.

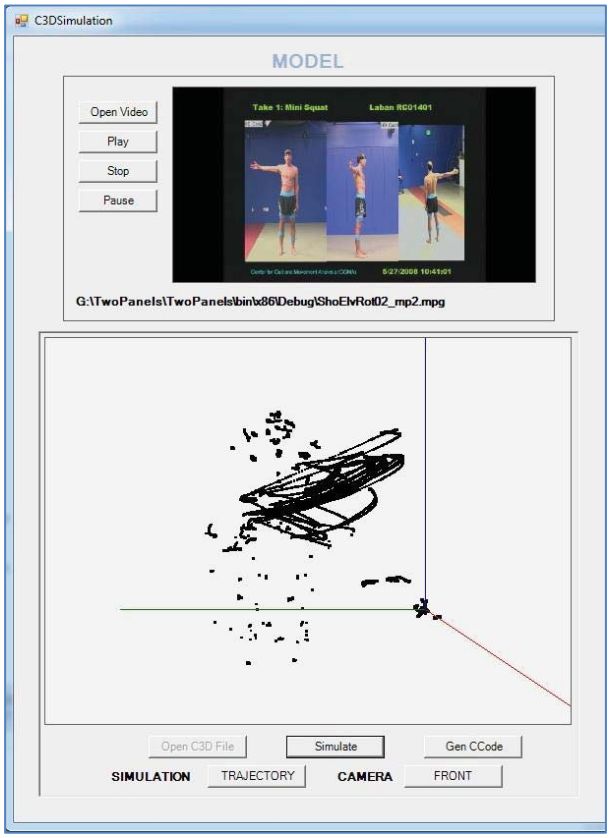


Fig. 3. Chain Code Generator

Some of the results obtained by using our software application with different ***step size*** values (0.2, 0.4, 0.6, 0.8, 1.0) are shown in Table 1. Labels used in the table are:

MSx = Mini-squat x
PPx = Posterior Pelvic Tilt x
SEx = Shoulder elevation with rotation x
SHx = Standing hip abduction x

where x represents the session number (1-3), and numbers in **bold** face represent the ***step size***

Table 1 Human Motion Similarity values for the four Key Rehabilitation Exercises using different step sizes

0.2	MS1	MS2	MS3
MS1	1.0	0.577	0.640
MS2		1.0	0.601
MS3			1.0

1.0	SE1	SE2	SE3
SE1	1.0	0.868	0.839
SE2		1.0	0.855
SE3			1.0

0.4	PP1	PP2	PP3
PP1	1.0	0.809	0.739
PP2		1.0	0.714
PP3			1.0

0.8	SH1	SH2	SH3
SH1	1.0	0.778	0.730
SH2		1.0	0.750
SH3			1.0

Human motion similarities for three different sessions for each exercise are shown in Fig. 4, 5, and 6 respectively.

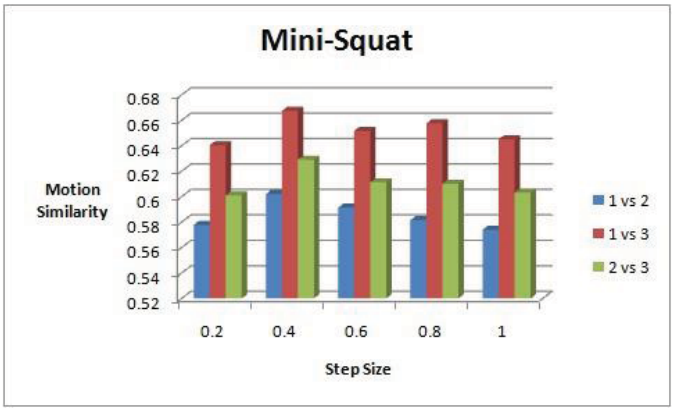


Fig. 4. Human Motion Similarity for Mini-Squat exercise

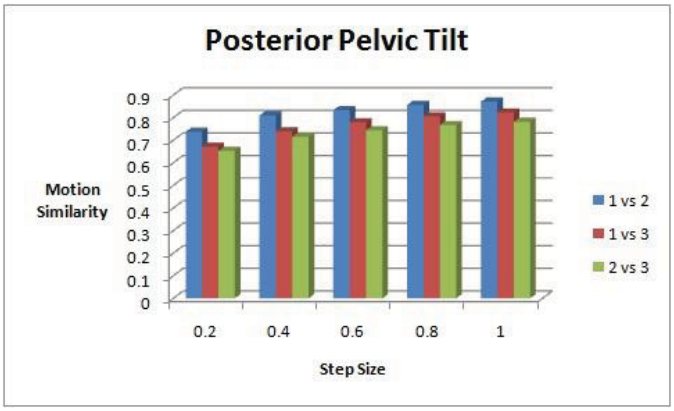


Fig. 5. Human Motion Similarity for Posterior Pelvic Tilt exercise

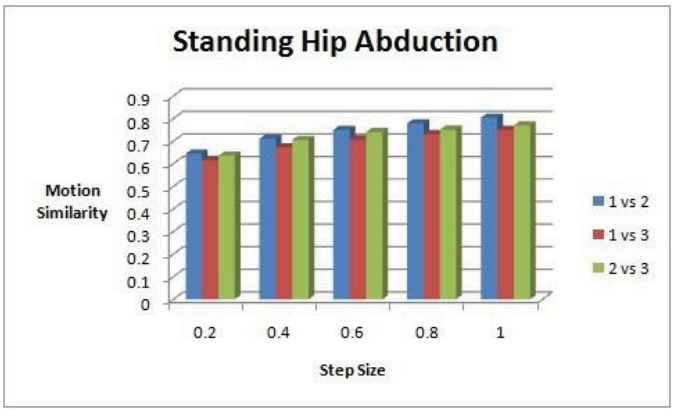


Fig. 6. Human Motion Similarity for Standing Hip Abduction exercise

VI. Conclusions and Future Work

We have presented the use of chain code representation for 3D curves which allows the analysis of human motion similarity. Analysis of similarity among different human motions is done by searching for global alignment between chain codes using Needleman-Wunsch global alignment algorithm, and then by applying an information-theoretic definition of similarity on those alignments.

Chain code representation for 3D curves proposed in [15] was enhanced to include a representation for the lack of motion, or the periods of time where body parts are immobile during human motion.

Different values for ***step size*** were used in the analysis of similarity for the four key rehabilitation exercises. Although graphs in Fig. 4, 5, and 6 show an increase of human motion similarity if ***step size*** is larger, that is not the case. Larger ***step size*** generates more lack of motion characters in the chain codes even if the body parts are moving, and the ratio of sensors placed on body parts that barely move in a particular human motion add to the final value of human motion similarity. As an example, Mini-squats key rehabilitation exercise shown in Fig. 4 requires the movement of almost all of the markers set on the human body, whereas in Posterior pelvic tilt key rehabilitation exercise requires the movement of very few markers, i.e. those in the abdominal region. We can see that in those motions where most of the markers are immobile contribute to have a more similar human motion, which increases if ***step size*** is larger.

A better analysis of human motion similarity must include time and space in the equation: being able to select the limbs (set of sensors) and the periods of time for a particular human motion should generate a more accurate analysis for human motion similarity.

An approach that considers the human motion as a human body generating sets of 3D curves, and the description of human motion by using a human movement notation is described in [20]. The motion analysis takes those sets of curves and performs a measure of similarity between them, taking time and space domains into consideration.

Finally, the use of a global sequence alignment algorithm serves well in an approach where the analysis of similarity is meant for the whole human motion, but if we need a more granular analysis we may need an algorithm that doesn't modify the chain codes in order to provide motion similarity values. Reference [21] proposes a measure of dissimilarity for 3D curves that can be adapted for the purposes of this work.

Acknowledgments

We want to thank Dr. James Carollo for the use of The Children's Hospital Center for Gait and Movement Analysis facility in Denver, CO, USA, where the Motion Capture sessions for the four key rehabilitation exercises were made.

Also, we want to thank Taipei National University of Arts for their sponsorship, in particular Professor Yunyu Wang.

REFERENCES

[1] Clinical Gait Analysis and its Role in Treatment Decision-Making; Roy B. Davis, III, PhD, Sylvia Õunpuu, MSc, Peter A. DeLuca, MD, Mark J. Romness, MD. [Online]. Available: http://www.medscape.com/viewarticle/440148

[2] Sports performance. [Online]. Available: http://www.osteopathiccenter.org/sportsdir.html

[3] Waking Life. [Online]. Available: http://www.youtube.com/watch?v=uk2DeTet98o

[4] Antonio Camurri, Ingrid Lagerlöf, Bualtier Volpe, "Recognizing Emotion from Dance Movement: comparison of spectator recognition and automated techniques," *Int. J. Human-Computer Studies*, 59, 2003.

[5] Performing Measurement and Analysis of a Ball's Impact Characteristics using LABView and NI Vision Hardware and Software; Fiachra Collins, Dr. Kieran Moran, Dr. Dermon Brabazon; National Instruments. [Online]. Available: http://sine.ni.com/cs/app/doc/p/id/cs-12479

[6] Javier Varona, José M. Buades, Francisco J. Perales, "Hands and Face Tracking for VR applications," *ACM 2004*.

[7] Roy B. Davis, III, , Sylvia Õunpuu, Dennis Tyburski, and James R. Gage, "A Gait Analysis Data Collection and Reduction Technique," *Human Movement Science 10, 1991*.

[8] John Eakins, Margaret Graham, "Content-Based Image Retrieval," *JISC Technology Applications, Report 39, October 1999*.

[9] Guojun Lu, "Chain Code-Based Shape Representation and Similarity Measure," *Gippsland School of Computing and Information Technology, Australia*.

[10] Chi-Min Oh, Chil-Woo Lee, "Bayesian Shape Recognition using Principle Component Analysis and Modified Chain Codes," *Int. Conf. on Control, Automation, and Systems (COEX), Seoul, Korea, 2008*.

[11] Filip Krolupper, Jan Flusser, "Polygonal Shape Description for Recognition of Partially Occluded Objects," *Pattern Recognition Letters 28 (2007) 1002-1011*.

[12] Chi-Keung Tang, Gérard Medioni, "Robust Estimation of Curvature Information from Noisy 3D Data for Shape Description," Institute for Robotics and Intelligent Systems, Univ. of Southern California, Los Angeles, CA.

[13] Tomislav Petkovic, Josip Krapac, Technical Report: Shape Description with Fourier Descriptors, 2002.

[14] Rafal Kapela, Andrzej Rybarczyk, "Real-time Shape Description System based on MPEG-7 Descriptors, " *Journal of Systems Architecture 53 (2007) 602-618*.

[15] Ernesto Bribiesca, "A Chain Code for Representing 3D Curves," *Pattern Recognition 33 (2000) 755-765*.

[16] Philip Resnik, "Disambiguating Noun Groupings with Respect to WordNet Senses," Sun Microsystems Laboratories.

[17] Frakes, W. B. and Baeza-Yates, R., Information Retrieval, Data Structure and Algorithms, Prentice Hall.

[18] Michael McGill, Matthew Koll, Terry Noreault, "An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems," School of Information Studies, Syracuse University, 1979.

[19] Dekang Lin, "An Information-Theoretic Definition of Similarity," Department of Computer Science, Univ. of Manitoba, Canada.

[20] Francisco Torres, J. Knight, Yunyu Wang, James Carollo, Edward Chow, Semwal Sudhansu, "Improve Rehabilitation and Patient Care with Laban Specification and Wireless Sensor Tracking," *Gait & Clinical Movement Analysis Society Conference, GCMAS 2009*.

[21] Ernesto Bribiesca, Wendy Aguilar, "A Measure of Shape Dissimilarity for 3D Curves," *Int. J. Contemp. Math. Sciences, Vol. 1, 2006, no. 15, 727-751*.

*i*Rock Interactive Multimedia System

Chris Ellison, Bee Ong and Kia Ng
ICSRiM - University of Leeds
School of Computing & School of Music
Leeds LS2 9JT, UK
kia@icsrim.org.uk, www.icsrim.org.uk

***Abstract*— This paper presents a technology-enhanced learning system for geology using musical sound from rocks. It discusses the system design, detailing multi-channel digital acquisition, audio-visual augmentation, real-time feedback together with a physical interface design that mimics the traditional xylophone/lithophone. The paper reports on several conceptual designs and rationale for the final design choices implemented. Using the combination of natural sound and multimedia interface, iRock (the interactive Rock instrument) hopes to stimulate curiosity and interest in the relationship between the geological formation processes and acoustic qualities of rocks.**

Keywords- interactive; multimedial performance; lithophone; geology; edutainment; sound visualisation; audio-visual feedback

I. INTRODUCTION

This paper presents the design and development of an interactive rock instrument for a permanent public installation at Brantwood, the former home of John Ruskin (www.brantwood.org.uk). The overall aim of the iRock project is to use the musical characteristics of mineral rocks to introduce knowledge about sound, and its relationship to the rich geology and natural resources of the fells and dales. The project is conducted in collaboration with aggregate quarries and schools local to the Cumbria region.

The project aims to create a new and innovative musical interface, by designing a stone instrument fit for the 21st century. The instrument is designed for an interactive learning exhibit to introduce all age groups to geological knowledge and musical potential of rocks from the region. Though resembling a traditional lithophone in its physical appearance, the design incorporates interactive multimedia technologies and digital signal processing to provide real-time audio-visual feedback. The resulting interactive rock instrument (iRock) offers an array of functionalities including: visual representations of geological structure using Scanning Electron Microscopy (SEM), associated textual geological information such as rock formation processes, and digital signal manipulations, both for audio augmentation and to provide the performer visual feedback. iRock is designed using a variety of rocks types common to the region to illustrate differences between igneous, sedimentary and metamorphic rocks.

As part of the project, there are two separate instruments: (i) a larger four octave instrument tailored towards acoustic performance incorporating an elliptical key layout with optional sensor/electronics attachments; and (ii) a smaller setup containing up to sixteen keys arranged linearly, with a different rock type for each key of the instrument, specifically designed for interactive multimedia edutainment. The multimedia attachment from the smaller setup can be transferred and used (with additional sensors) on the main instrument for augmented multimedia performance.

II. SYSTEM DESIGN

The basic architecture of the iRock interactive multimedia system is summarised in Figure 1. Vibrations emanating from tuned stone keys are captured using an individual piezoelectric transducer for each note. The resulting analogue signals are connected to separate channels of an audio interface where analogue to digital conversion takes place prior to routing the signals to a computer. Here custom-designed software modules manage the incoming signals, and use a variety of signal processing techniques to generate both audio and video outputs for interactive multimedia feedback.

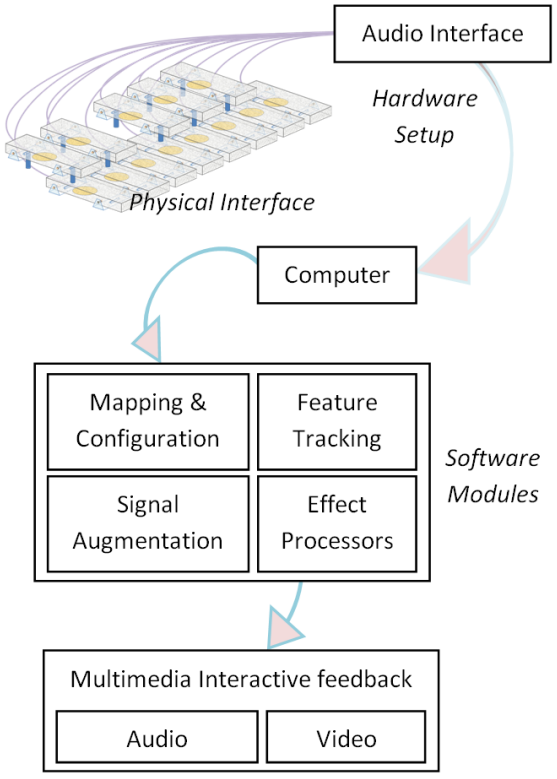


Figure 1. System overview.

A. *Analogue Instrument*

The analogue instrument provides a physical interface through which the performer may create sound, and interact with the multimedia elements of the iRock system. It is in essence a lithophone, a form of tuned percussion instrument comprising of stone keys tuned to an equally tempered scale. The key layout adopts a traditional keyboard style, as one would expect to find on traditional classical instruments such as the piano or xylophone, making the instrument familiar and easy for the musician to learn.

The design of the acoustic instrument explores user-centred concepts, aiming to minimise the amount of physical energy expended during performance. Two major changes to traditional horizontal keyboard layout design are proposed. The first of these is to adopt an elliptical layout, allowing the performer to remain in a central position, as illustrated in Figure 2 (left). By choosing this layout, the amount of arm extension required for the performer to reach upper and lower ranges of the instrument is minimised by comparison to a conventional linear arrangement. It is also believed that strike accuracy can be improved as the forearm of the player, and longitudinal axes of the beaters remain approximately parallel to the longitudinal axes of bars throughout the pitch range of the instrument. Further studies are required to test this hypothesis, involving the use of a motion capture system to evaluate performance technique.

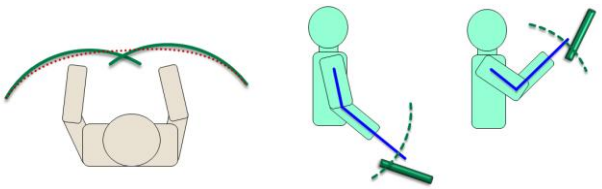


Figure 2. Elliptical layout with adjustable playing surface height and inclination.

In addition to incorporating an elliptical layout, the height and inclination of the instrument's playing surface is adjustable. Figure 2 (middle and right) above illustrates how optimum inclination angle for a performer is dependent on the playing surface height. This design modification has the additional benefit of allowing the playing surface to be tilted towards the audience, heightening the visual experience.

The design requirements specify that rock samples with a wide range of geological compositions, sourced from a variety of local quarries, be incorporated into a single instrument, to encourage learning about the geology and aggregates industry of the region. This poses a unique challenge, due to the large variation in acoustical and mechanical characteristics exhibited by the different rock types. A set of samples for each rock type has been collected, to allow recording and measurement of their acoustical properties as discussed below.

1) Transverse Vibrations of a Uniform Bar: To make meaningful quantitative comparisons between rock samples a standard theoretical model for the vibrating bar is utilised. A rectangular bar may produce transverse, longitudinal or torsional vibrations, though predominantly transverse modes are excited when striking the bar on its upper surface. The equation of motion describing transverse displacement *(y)* along the bars length *(x)* is given below [1, 2, 3]:

$$\frac{\partial^2 y}{\partial t^2} = -\kappa^2 c^2 \frac{\partial^4 y}{\partial x^4}$$

$c = \sqrt{Y/\rho}$ *(Y =Young's modulus,* ρ *=density)*

$\kappa = T/\sqrt{12}$ *(Rectangular cross-section, thickness T)*

The constant κ may be thought of as the radius of gyration of the bar's cross-sectional area S, defined by $\kappa^2 = \int r^2 dS/S$

The value of this integral for a rectangular cross-section is $\kappa = T/\sqrt{12}$ as provided above.

Applying boundary conditions to our differential equation of motion, for a bar free to move at both ends, leads to an important relationship, which predicts the frequency (*f*) of each transverse mode of vibration in terms of bar's length (*L*), and constants *c* and κ defined above:

$$f = \frac{\pi c \kappa}{8L^2}(3.0112^2, 5^2, 7^2, 9^2, ...)$$

It is important to note in the above formula that the higher modes of vibration are not integer frequency multiples of the fundamental, and therefore not true harmonics as with the vibration of a guitar string, for example. The equation of motion differs from that of the vibrating bar as it contains a fourth partial derivative with respect to *x*, rather a second order term. This is significant in that it indicates the vibrating bar is a dispersive medium for transverse waves – different frequency components of complex wave solutions to the equation of motion travel through the bar at different speeds, causing dispersion and the shape of the wave to change over time. The resulting anharmonic modes of vibration are an important factor in the distinctive sound of the rectangular bars, in addition to the transient attack characteristics produced by striking mallet on stone.

The decay rate of each mode of vibration depends on a number of factors, such as internal energy losses, energy losses via the bar suspension, and frequency of vibration. Amplitude decay curves for each mode of vibration are exponential to a good approximation, allowing use of the logarithmic decrement δ (natural log of the ratio between successive maxima A_n and A_{n+1}) as an indicator of decay rate:

$$\ln \frac{A_n}{A_{n+1}} = \delta$$

2) Bar Suspension and Bracket Design: An ideal mechanism for suspending the instrument bars must allow freedom of movement, thus preserving acoustic vibrations and producing minimal contribution to vibration energy losses. Equally, the suspension must be robust and the bars well secured, as the instrument will live in a public space. These competing design goals have led to a number of possible bracket designs being tested, both for their mechanical strength and impact on decay rate. A traditional xylophone suspension involves drilling horizontally through the bars parallel to the nodal plane, and at the neutral axis for transverse bending of the bar. Due to the more fragile nature

of the rocks this type of drilling is impractical, and an alternative solution must be found. Another method is to drill vertically through the bar at the intersection of nodal plane and longitudinal centreline. During experiments, it was determined that this may slightly alter the pitch as the material removed by drilling is not along the neutral axis for bending of the bar. It has been found that fixing pins into the side of the bar using an epoxy (parallel to nodal plane, at neutral axis, pin diameter $\approx$ 3mm, drilling depth $\approx$ 6mm) has a minimal effect pitch and decay rate. These pins may then be mounted into a machined bracket, using rubber grommets to allow some freedom of movement. Figure 3 shows the current bracket design.

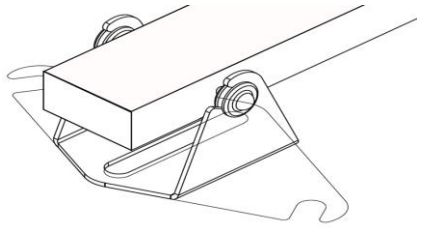


Figure 3. Bracket design for iRock keys.

3) Rock Tuning and Layout: To determine the optimum bar layout, experimental data has been gathered during the testing of over 90 rock samples of varying type and dimension. Firstly, the rocks have been categorised according to their ability to sustain vibrations i.e. their decay rates. Decay rate has a critical dependance on the rock's crystalline structure, which must transmit vibrations with a minimum of internal losses. For example a rock which is full of pore spaces does not ring well, as the air or water in the pores produces highly effective damping. Following calculation of the decay rates (dB/s) from audio recordings, the chosen rocks are ordered within the instrument with highest decay rate at the lower end of the pitch range. By ordering in terms of decay rate, differences in sustain time when moving between keys of a different rock type are minimised. Having the determined the overall layout of rock types, tuning of individual rock keys may commence.

The Young's modulus and density of each rock type are colculated, and predictions made for each bar length and thickness required to tune the key to a particular pitch. In a traditional xylophone, constructed with bars of a single uniform material, the inverse square relationship between frequency and bar length produces a natural curvature when bars are laid in ascending order. We wish to preserve this natural curvature, and have achieved this aim by adjusting bar thickness for each rock type, thus compensating for material dissimilarities.

Since higher frequency vibrations decay more rapidly for a given logarithmic decrement, some compensation for varying decay rates is inherently provided by positioning rock types with longer sustain times in the upper ranges of the instrument. This has the additional advantage of allowing shorter bars to be constructed from the weaker sedimentary limestones which have the greater sustain times, while longer key lengths in the lower pitch range will be constructed from harder and more robust rock types. Although the required rock lengths could be estimated to a reasonable degree of accuracy, non-uniformities in geological composition and thickness dictated that an iterative fine tuning process be adopted, a lengthy undertaking where small amounts are removed from the bar length until the desired pitch is attained. Tuning accuracies of <5 cents (the approximate 'just noticeable difference' in pitch for the human ear) have been achieved using this process.

As a prototype to validate both the tuning processes, and hardware / software interface, an eighteen-semitone instrument has been constructed from limestone with a simple linear key arrangement. This has provided valuable insight into the practicalities of tuning rock bars, and allowed testing of interactive multimedia elements of iRock.

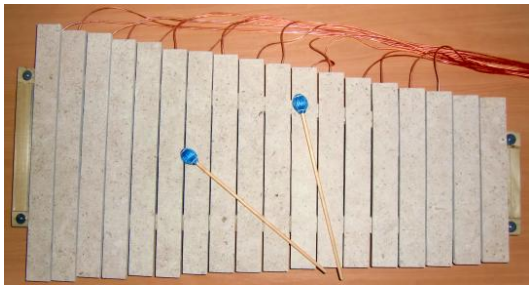


Figure 4. iRock prototype.

B. *Augmented Instrument*

1) Data Acquisition: Contact microphones constructed from a ceramic piezoelectric material are used to translate acoustic vibrations into voltage signals for analogue to digital conversion. The frequency response of a typical piezoelectric crystal is summarised in Figure 5. The useful frequency range of the microphone corresponds to the region of the graph over which the gain is approximately linear, avoiding both the low frequency RC roll-off and high frequency resonance regions.

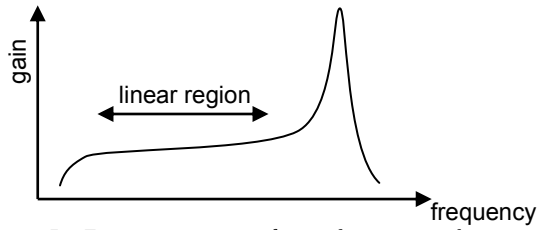


Figure 5. Frequency response of piezoelectric microphone.

A further consideration is input impedance matching due to the high (typically megaohm) impedance presented by a passive contact microphone. Improved impedance matching may be achieved by use of a buffering pre-amp, though use of an audio interface with sufficiently high input impedance makes direct connection viable.

Microphone options which have been tested include low cost ceramic discs (connected directly or via a buffering pre-amp built in-house), prefabricated contact microphones (both passive and active 48V phantom powered), and directional microphones which are non-contact and do not contribute to bar damping. Microphone positioning and method of

attachment is also found to have a critical effect on bar damping. Of the microphones tested, a prefabricated passive microphone with XLR connector provided the best solution in terms of frequency response, signal-to-noise ratio, bar damping, physical size, robustness and overall cost.

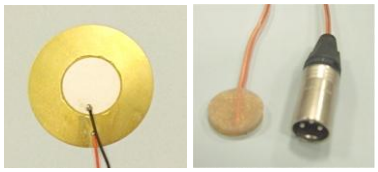


Figure 6. A piezoelectric ceramic disc (left) and a piezo-based passive contact microphone with XLR connector (right).

2) Audio Interface: The audio interface receives analogue voltage signals from the contact microphones, and converts to a digital signal that can be utilised within the Max/MSP/Jitter programming environment. Our first critical consideration is the number of available input channels. In an idealised setup, both the four-octave performance instrument, and the smaller interactive (iRock) instrument would provide separate input channels for each instrument key. This configuration allows separate processing of signals from each key, greatly simplifying software key detection when the performer strikes a note. For the 49-key acoustic performance instrument this raises a number of concerns:

- sourcing a portable audio interface with provision for such a large number of input channels
- simultaneous 49 channel signal processing places a heavy burden on the computer's CPU

In the current prototype, two optical ADAT linked audio interfaces each provide eight input channels, allowing direct connection of sixteen keys via XLR inputs. Communication between audio interface and computer is presently via a firewire interface with incoming channels routed to the Max/MSP/Jitter graphical programming environment using an [adc~] object. This configuration provides adequate channels for the final iRock edutainment system.

3) Signal Processing: Digital signal processing (DSP) is performed using patches developed in Max/MSP/Jitter. The patches are designed to allow control over sonic characteristics such pitch, sustain, and timbre. For example in the analogue instrument, pitch of the lowest note is physically constrained by maximum bar length and minimum bar thickness. Real-time DSP is applied to allow frequency domain pitch shifting to extend the attainable pitch range of the iRock instrument.

Individual signals are first filtered to remove low-frequency noise and 50Hz hum, then sent to the sub-patch responsible for key detection. At present key detection is achieved by an edge detection algorithm that detects the leading edge of the attack when a key is struck. This method of key detection adds minimal latency with triggering delays unnoticeable to the performer. Trigger sensitivity is adjustable to allow for differences in the playing environment, mallet choice, and performer preference.

The simple edge detection algorithm performs well across a wide range of strike forces, allowing a good dynamic range for expressive performance. However, if adjusted to provide maximum trigger sensitivity, with extremely loud key strikes (out with the dynamic range for normal playing) it is possible for a bar strike to be picked up by the piezoelectric sensor of an adjoining bar giving false key detection. We aim to resolve this in future versions by using a comparative ordering of the edge gradients to eliminate false key detections whilst continuing to recognise polyphonic playing.

Alternative key detection methods, such as pitch tracking or attack detection [4] of merged input channels, are being tested for use on the four-octave performance instrument. Whilst pitch based key detection has been achievable, use of an FFT based detection algorithm inherently adds undesirable latency to the system which is noticable to the performer (to give the required frequency resolution for key detection the FFT must analyse a significant number of samples to perform the calculation).

Detected key strikes are converted into MIDI messages comprising of MIDI note number, and MIDI velocity indicating how forcefully the key is struck. User adjustment of the velocity sensitivity is provided to allow the performer to correct for variation in attack characteristics produced by a different beater hardness or weight [5]. In essence the iRock may be used as a fully functional, velocity sensitive MIDI interface used to trigger any audio sample, with stored sample banks selectable from the user interface.

After filtering and key detection, the merged signal is temporarily held in a cyclic (circular) buffer [6], which stores a fixed number of digital audio samples. The contents of the cyclic buffer may then be retrieved and processed to generate an array of audio effects, which we broadly divide into two categories – enhanced natural sound and FX.

4) Augmentation: The educational aspects of the project require the iRock instrument to contain a variety of rock types. These different geological compositions vary greatly in their acoustical qualities, yet must be incorporated into an instrument that remains musical. One of the primary disparities in rock acoustics is the decay rate, a measure of the duration for which bar vibration is sustained. To preserve the musicality of the instrument it is desirable to have flexible control over the decay rate of each rock type, minimising acoustic disparities by augmenting the sustain time of rocks with inherently poor acoustics.

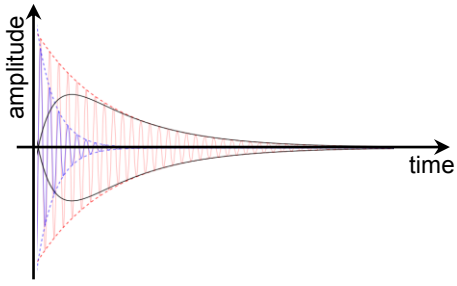


Figure 7. Exponential amplitude decay.

Variation in sustain time is depicted in Figure 7, where blue and red curves represent exponential amplitude decays of two rocks with differing acoustic qualities. To equalise the sustain time of the two rock types, the amplitude difference (drawn as

a black curve) may be added to the amplitude of the blue curve at each instant in time. Since frequency ratios of transverse vibration modes are known, it is possible to augment sustain times by adding sinusoidal components and shaping amplitude envelopes governing the decay for each component frequency [7]. A patch has been created in Max/MSP to algorithmically generate the required sinusoidal components, with amplitude envelopes shaped by adjustment of exponential time constants. Implementation of sustain control with the prototype iRock instrument is now in progress, with real time measurement of peak amplitude following a bar strike, and manipulation of decay curves for individual frequency components the key features.

In addition to augmentation of the sustain times, the pitch range is extended by use of the [gizmo~] MSP object for frequency domain pitch shifting [8]. When an iRock key is struck, the resulting audio buffer may be sent to a single [gizmo~] object for pitch shifting, or multiple instances of [gizmo~] can be used to generate chords of up to eight notes at any chosen intervals. By introducing variable signal delays between the cyclic buffer and various [gizmo~] objects, arpeggio chords are produced, with arpeggio rate user defined. Modulating the degree of pitch shift produces a range vibrato effects, from subtle colouration of natural sound, to more extreme and synthetic sounding timbres. Similarly, amplitude modulations may be used to introduce tremolo if desired.

These augmentations to the iRock instrument equip the performer with a broad palette of sounds, all originating from the natural sound emanating from stone, and all produced in real time. The same array of effects may also be applied to triggered samples, when the iRock is used as a MIDI controller. Taking a MIDI output from the audio interface allows external MIDI instrument to be controlled directly from the iRock. To fully utilise the available sound palette, it is planned to implement some multi-track looping functionality allowing the performer to adjust instrument parameters to find a sound, record an audio (or MIDI) clip, then loop playback and record additional tracks, constructing a multi-layered performance.

5) User Interface: A suitable interface is required to allow the user control over available functionality and interactive educational content. A number of options are under consideration, such as use of an 'off the shelf' MIDI footpedal / touchpad system, or mouse interaction with the wall mounted display screen. At present ASCII codes from a simple numeric keypad are used to select between functional modes. We plan to integrate a tactile interface in future versions to improve accessibility.

6) Audio Output: Prior to converting back to analogue audio for amplification it is intended to have an additional effects stage, to provide standard effects such as reverberation, echo, chorus, flanging and digital delay. Although relatively simple to implement effects of this type within the available software, it may become necessary to reduce CPU load by delegating standard effects to some additional external hardware. Following this final effects stage in the signal path, audio is output for amplification in two-channel stereo. Later versions of the iRock may incorporate multi-channel surround sound, allowing some spatialisation and directionality to be added to the sound, though at present this is beyond the scope of the project.

7) Visualisation: To provide a full multimedia experience, visual feedback from the iRock instrument is displayed on a wall-mounted high-definition screen. Visual aspects to the project must fulfil three main criteria – to provide geological information on the rock type as a key is struck, visualisation of the sound emanating from the instrument, and information to guide the user through the instrument's features and allow control and adjustment of parameters.

Using the key detection features already outlined, the resulting MIDI note-number messages are used to control the display of geological information. As a new key is struck, information as to the rock type and geological composition of the rock are brought up on the display. This geological information includes images taken using a scanning electron microscope (SEM) at a number of magnification settings, which give a visual representation of the rocks geological structure (see Figure 8 left). To enhance interactivity, key strike velocity is used determine the magnification – increasing the force with which a key is struck increases the magnification allowing the user to zoom in on the crystalline rock structure. Simultaneously, textual information is displayed, describing a rocks geological type, mineral composition, formation processes, and source location. This feature of iRock is designed to introduce learning about the region's diverse geology, and reinforce the link between each rock's crystalline structure and its ability to sustain vibrations.

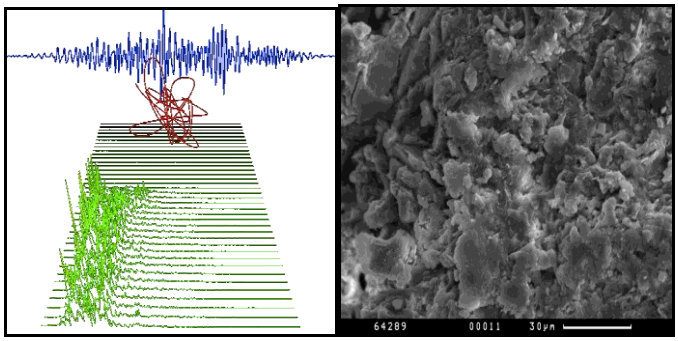


Figure 8. Real-time visualisation of iRock. Audio visualisation (left). SEM picture (right) of a rock key that has been hit.

In addition to learning interactively about the geology of the rocks, the display provides visualisations of the sound, giving an instantaneous visual response to changes in the audible sound. The visualisations reflect the sound which is output from the iRock software to speakers or headphones. With no augmentation we directly visualise the natural vibrations of the rock keys. If an augmentation is selected, we can hear how the sound has been altered and see these differences in the visual feedback.

Firstly, a 2D (amplitude-time) waveform display (see Figure 8, top) shows real-time changes in the audio signal, similar in form to an oscilloscope plot. When a key is first struck, the dense frequency spectra creates an irregular, jagged waveform, which resolves into a smooth sinusoid as damping attenuates transients and higher vibrational modes, until only the fundamental frequency remains. Differences in wavelength

may be clearly seen when selecting high pitched and low pitched keys, thus highlighting the inverse relationship between frequency and wavelength as taught in school physics lessons.

Secondly, a scrolling 3D (frequency-amplitude-time) plot indicates the time-variant spectral composition of the sound (see Figure 9, bottom), by successively plotting the results of each in a rapid series of Fast Fourier transforms (FFT) to create a scrolling display [9]. In contrast to the two-dimensional waveform, here it is possible to identify individual frequency components which correspond to particular modes of vibration excited by a bar strike. Different strike positions on the bar excite these modes to varying extents dependant on proximity to a node or anti-node. Decay rates for individual modes of vibration can be seen, with the fundamental always outliving the higher modes due to the positioning of bar supports at nodal points of the first mode of vibration.

Finally, a more abstract, but fun visualisation is provided by the addition of a Lissajous plot (Figure 8, middle), which resembles a twine ball reacting to the changing harmonic content of a sound by continually morphing in shape. During transient attacks when spectral content is dense, the plot forms into complex knots, then decays to smooth elliptical curves as initial transient attacks decay to steady-state vibrations. These shapes are generated by plotting audio signal sample values along three orthogonal axes, with delays on two of these axes to introduce a phase shift between the signals.

The audio visualisations provide a valuable link between what is being heard, and how this relates to the time-variant frequency content – or timbre - of the sound. With the addition of augmentations and FX, often the process by which the sound is altered can be readily seen. For example if a tremolo (amplitude modulation) effect is used, the waveform display and lissajous plot can be seen to pulse as the amplitude grows and diminishes at the same modulation rate as the heard sound. In the 3D scrolling display the sinusoidal modulations can be seen superimposed on the decaying fundamental. A vibrato (frequency-modulation) causes the waveform display to pulse in wavelength in a similar manner. Increased sustain times can be observed by a slower decay on the visual displays. Even triggered audio samples can be observed allowing visualisation of any sound, from a drum hit to a vocal sample.

III. Conclusions and Further Work

The paper presents the iRock interactive multimedia system, focusing on key elements influencing the design process. Adapting traditional percussion instrument design, an innovative physical interface, the iRock instrument has been outlined, along with its integration with new technologies to provide an interactive multimedia system offering new musical possibilities.

The physical instruments have been realized recently and tested by several players with positive feedback on playability and acoustical qualities. Further evaluations on hardware and software are being carried out prior to final installation in the coming weeks.

We continue to evolve new design features, building on current system to extend its multimedia functionality and accessibility, with further educational content.

Further explorations of the project include composition, performance technique and performance interface studies centred upon the iRock instrument interface utilising user-centre design concept, gesture interface and multimodal augmentation.

Acknowledgment

Thanks to all partners and collaborators of the Ruskin Rocks project (www.ruskinrocks.org.uk, www.leeds.ac.uk/ruskinrocks) for their interest and contribution, and funding support from Natural England, through DEFRA's Aggregates Levy Sustainability Fund.

References

[1] Fletcher N.H., Rossing T.D. (1990) The Physics of Musical Instruments. Springer-Verlag, New York.

[2] Kinsler, L.E., Frey, A.R., Coppens, A.B., Sanders, J.V. (1982) Fundamentals of Acoustics. John Wiley & Sons, Toronto.

[3] Rossing, T.D. (2001) Acoustics of Percussion Instruments. Acoust. Sci. & Tech., 22(3), 177-188.

[4] Puckette, M.S., Apel, T., Zicarelli, D.D. (1998) Real-time audio analysis tools for Pd and MSP. International Computer Music Conference (ICMC), University of Michigan, Ann Arbor, International Computer Music Association.

[5] Bork, I. (1990) Measuring the Acoustical Properties of Mallets. Applied Acoustics, 30, 207-218.

[6] Charles, J.F. (2008) A Tutorial on Spectral Sound Processing Using Max/MSP and Jitter. Computer Music Journal, 32(3), 87-102.

[7] Puckette, M., (2005) Some Mathematical Tools for Music. <http://crca.ucsd.edu/~msp/publications> (March 2010).

[8] Dudas, R. (2002) Spectral Envelope Correction for Real-Time Transposition: Proposal of a "Floating-Formant" Method. In Proc of the Int. Computer Music Conference, Gothenburg, Sweden, 16-21 September.

[9] Kim-Boyle, D. (2004) Spectral Delays with Frequency Domain Processing. 7th Int. Conference on Digital Audio Effects (DAFX-04), Naples, Italy, 5-8 October.

Towards Hybrid Index Structures for Multi-Media Search Criteria

Richard Göbel, Carsten Kropf
University of Applied Sciences Hof
Alfons-Goppel-Platz 1
95028 Hof, Germany

Abstract-[1] *Modern multimedia databases support the description of objects by different data types, including relational, audio, video and vector data as well as texts and images. Efficient access to this data is supported by index structures via various search criteria. At least two very different principles are used for the construction of an index structure. An index structure for an ordered data type with a single value will probably use a divide and conquer strategy resulting in some type of search tree. An index structure for a data type with many values (e.g. words in texts or features of images) has to use a different strategy. A common approach associates values with all objects containing these values. Optimal support for searches combining criteria for different data types requires special index structures facilitating simultaneous checking of all involved criteria. This paper discusses the challenges for the definition of such an index structure. Based on this discussion we introduce a new index structure combining these principles. A theoretical and experimental investigation of this index structure demonstrates the feasibility of the proposed concept.*

I. INTRODUCTION

Modern database management systems support multimedia applications by providing complex data types for table columns. Early approaches introduced data types for binary large objects providing only very limited support on the contained data. More recent developments introduce textual and/or vector data with specific operations on these data types. In future additional data types will be probably integrated for images, audio and video data including functions for content based searches. For this purpose features need to be extracted from this type of data and queries should define constraints for these features.

Databases support efficient access to the data of a table by index structures. At present different types of index structures are used for different types of data. Index structures for simple data types like numbers or small strings (single-valued data types) often apply a divide and conquer strategy using the natural order on column values to split table entries into disjoint sets. This strategy is usually implemented as a tree structure such as the B-Tree [1] or the R-Tree [8]. The B-Tree and its variants ensure logarithmic search times for queries specifying a value of interest for one or more columns or a range of values for a single column. R-Trees are designed to support range searches for more than one column making them an interesting index structure mainly for spatial searching.

A different strategy needs to be applied for columns with a large number of values ("multi-valued attributes"). As an example, a longer text is usually comprised of a significant number of words. Other examples are images, video and audio clips which are described by multiple features. In the case of multi-valued attributes a divide and conquer strategy is difficult to apply since most subsets of entries usually share a larger number of common values.

A common approach is the association of a value with all entries containing the value in the corresponding column (posting). In the case of texts a posting will contain a term and references to all texts containing this term. The list of all postings is usually called an inverted index here.

Both types of index structures are available in many modern database management systems. With these structures it is directly possible to support simple search requests specifying one constraint for a relational, textual or vector data column. The support of search requests specifying multiple constraints for different types of data is not so easy. Existing database management systems usually support two options here:

- The system selects the most specific search criterion and retrieves all entries satisfying this criterion using the corresponding index structure. Afterwards all entries are removed which do not satisfy the remaining search criteria.
- The system performs separate searches for every search criterion using the corresponding index structures. Afterwards the system computes the intersection of the intermediate search results.

Both options do not ensure efficient searching if very large numbers of entries satisfy all individual search criteria but the combination of these criteria is only satisfied by few or even no entries. In this case the preparation and processing of the intermediate result sets will take a significant amount of time even though the database will return few or even no results.

Database management systems support index structures combining values from multiple columns to deal with this situation, if these columns refer to the same type of data. So far existing systems do not provide index structures combining columns with both single- and multiple-valued attributes.

The next section will discuss the options for a new index structure supporting both types of data. Based on this discussion a new hybrid index structure is proposed in section

1 This work has been funded by the Deutsche Forschungsgemeinschaft and the Oberfrankenstiftung.

III. A theoretical analysis and the results of first experiments with this structure follow in section IV and V. Section VI summarizes the results of this paper.

II. CONCEPTS FOR HYBRID INDEX STRUCTURES

As already discussed in the previous section, the definition of a combined index structure for different data types is not obvious if these data types require very different index strategies. In this paper we consider the combination of columns containing a simple date type on the one hand (single-valued attributes) and on the other hand a single column containing complex data types described by a larger number of features (multi-valued attributes). Commercial database management systems (e.g. Oracle) and freeware systems (e.g. PostgreSQL) already provide some non-relational data types for large texts or vector data (spatial data). Also multi-valued data types (e.g. arrays or lists) are supported facilitating the representation of a larger number of features in a single column.

A combined index may adopt one of the following four strategies:

1. Values from all columns are integrated in an index structure for single-valued attributes using the divide and conquer strategy.
2. Values from single-valued attributes are considered in an index structure for multi-valued attributes (e.g. an inverted index).
3. Index structures for single-valued attributes and multi-valued attributes are concatenated.
4. Both criteria are combined into a single index structure using a deeper integration than option 3. We call index structures following this principle "hybrid index structures".

The integration of multi-valued attributes in an index structure for single-valued attributes (e.g. a B-Tree) requires the derivation of multiple aggregate values from a single entry in a database. As an example all terms would need to be derived from a text and combined with a single combination of values from the other columns (e.g. publisher and publication date). In this case more than 1000 aggregate values would need to be derived from a single entry if the text contains 1000 or more different terms. This approach will probably generate a very large index structure not only at the costs of high memory usage but also at lower search speed. Even worth it is not obvious how to support searches defining multiple features (e.g. terms) efficiently, which are common for queries addressing these complex data types.

Option 2 associates values for simple data types directly with corresponding entries. This approach is already used by document management systems to store information like the publisher or the publication date in an inverted index. The support of multiple search criteria with this approach still requires the generation of intersections for the entry lists associated with values from the search condition. In many (but not all) cases this intersection can be efficiently computed, if all entry lists are sorted. A key issue with this approach is the support of range searches specifying upper and/or lower bounds for a value. A potential solution is the introduction of additional values into the inverted index which represent ranges [3]. The efficiency of this approach is limited, since a single range needs to be mapped to several of these new values making the initial search and the composition of the search result more complex.

The concatenation of index structures for single-valued and multiple-valued attributes (option 3) was described in [10] for combining textual and spatial search conditions. As a consequence an inverted index and an R-Tree were used. The two sub-options are either R-Tree first or Inverted Index first. With R-Tree first (option A), the R-Tree will contain a *d*-dimensional point representing the values of the *d* relevant relational columns. Every point will reference an inverted index organizing all related documents occurring in entries with the value combinations of the point.

With Inverted Index first (option B), every term will point to a separate R-Tree. This R-Tree will be generated for all entries which are related to a text containing this word.

The paper [10] compares these two structures by applying them to some test data. These first tests seem to indicate that searches can be processed faster for option B. This paper does not yet provide a worst case analysis for time and space. In particular the space requirements are not considered.

A worst case analysis is given in [4]. In option A the worst case occurs for a large geographical region and a rare term. In this case most or even all nodes of the R-tree need to be searched, returning few or even no search results. Although this case does not occur for option B, this option uses significantly more space than the other option because many terms will occur in a large number of documents. This may result in space requirements which grow significantly faster than linear. As a consequence of this analysis, it is probably difficult to apply the efficient option B for larger databases due to its space requirements and option A is probably not sufficiently efficient in all cases.

A hybrid index structure (option 4) needs to integrate the divide and conquer strategy and combine it with methods for managing multi-valued attributes. This means that the skeleton of such an index structure will be a tree structure. Every node of this tree structure needs to provide information about both the values of single- and multi-valued attributes of entries below this node. The single-value attributes can be directly integrated into this skeleton resulting in a conventional index structure like the B-Tree or the R-Tree. The key challenge is the integration of a multi-valued attribute into the tree. For this purpose two obvious solutions exist:

- A node may directly contain a list of features representing all entries below a node.
- The set of features below a node may be represented by a bit list. Every bit in this list stands for the presence (1) or the absence (0) of a feature.

So far only few papers address concepts combining these criteria in one node. [6] proposes the KR*-tree assigning a list

of terms to every node. Bit list approaches are proposed by [3] for answering geographic nearest neighbor queries including search terms and [9] for retrieving a set of *k* coordinates with the least spatial extension for *k* entries with alternative coordinates. For these papers space requirements are not yet sufficiently addressed. It is obvious that the bit list will probably require less space. But even bit lists may become too large to fit into a disk block representing the node of a tree.

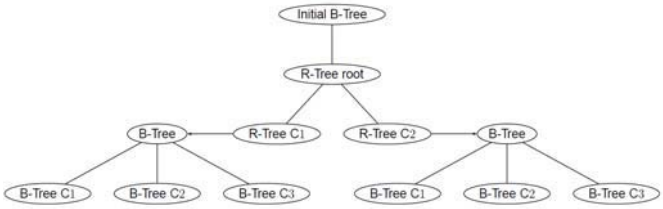
Reducing the required space for bit lists in nodes is investigated in [5]. The size of these bit lists is limited by including only frequent terms in it. Rare terms are stored in a conventional inverted index. The documents associated to these terms can be sequentially searched in reasonable time to check other search criteria. This approach however requires a relatively static R-Tree.

This paper introduces a more flexible solution facilitating an incremental generation of the index structure. It combines an R*-Tree ([2]) for an arbitrary set of single-valued attributes and relatively small inverted indexes for rare terms associated with nodes of this tree. Although we consider texts as multi-valued attributes, the proposed index structure can serve as a framework for arbitrary features derived from images, audio and video data.

III. A NEW HYBRID INDEX STRUCTURE

In this section we propose a new index structure which can be applied to a table containing several relational and **one** textual column. The index will be also applicable to a highly dynamic database content starting with a small number of entries (rows) and growing continuously over time.

The definition of the hybrid index will automatically generate a combined index for the relational columns and a



separate inverted index for the textual column. The relational index (for example a B-Tree or an R-Tree) will be prepared to accept inverted indexes at every node. In this paper we will focus on the case of an R-Tree without loss of generality. All inverted index will be organized as B-Trees. Initially the root node refers to the inverted index of the text column and the indexes of all other nodes are empty. Figure 1 summarizes the structure of the hybrid index.

The words of an inverted index will point to a set of entries with documents containing this word (and not to the documents directly). This approach makes it easy to check the relational conditions for entries related to a search word.

A user defined parameter *l* exists which specifies a limit for the maximum number of entries referenced by a word in an inverted index. As soon as a word references more entries, then the word is removed from the current inverted index and inserted in all inverted indexes of the child nodes. In every child node the word refers to a subset of entries satisfying the relational constraints of this node (point is inside region of node). This process may continue until the number of referred entries is below the specified limit or a leaf node is reached.

On the leaf level a word will be also removed for the inverted index of the node if its set of entries exceeds the specified limit. Then the word is assigned to the inverted indexes of all entries (represented by their point) in the node. Inverted indexes of points are the final destination of words and the size of document sets may exceed the limit here.

On the root level, an initial inverted index is constructed that contains all document entries and adds the functionality to map from a certain word to a code, which is generated sequentially.

This code is subsequently used in each inverted index that resides in the relational part of the hybrid tree. That means that

```
1  procedure insert (doc document)
2  begin
3    points  := doc.points;
4    words := doc.words;
5    insert words into inverted index
6    if entries overflow l there
7    do
8      insert points of entries to relational index
9      assign entries to inverted indexes of relational nodes (where they fit
           in relational context)
10     if again entries have overflows there
11     do
12       remove from this inverted index
13       distribute entries to all elements one level deeper in relational
           index
14     end if;
15   end if;
16 end;
```

Figure 2: Insertion Algorithm

the inverted index in each node element of the hybrid part of the tree will no longer map from a certain word to all the entries with documents that contain the word but will perform

rted Index

(B-Treea mapping from the particular word code to the entries. This approach allows for storing more elements in these nodes and it is easier to estimate the number of elements to be stored in one node of the respective inverted index. Figure 2 summarizes this approach.

In addition to the inverted index, a new storage structure containing certain word codes is attached to the nodes in the relational part of the index structure. This new structure is called “id list” and contains a subset of the word codes residing in the hybrid part of the index. If a certain posting in one particular inverted index exceeds the parameter *l*, mentioned above, this posting will be moved down one level into the hybrid part. All entries referenced by this posting will subsequently be distributed to the child nodes of the current node in order to reduce the number of entries referenced by this posting. Using this knowledge, the “id list” is generated in two stages. First, it will be filled using the codes of words from postings that were moved down one level during the insertion operation. However, this strategy cannot be used all the time, because of the limited size of a block on secondary memory. If the “id list” has reached its maximum extent and cannot be filled by putting more elements to it, a second strategy is

chosen, which is similar to the least recently used (LRU) paging strategy. That means that whenever one posting is inserted in one of the child nodes of the currently inspected node, the word code of the particular posting will be placed in the first entry of this list, probably eliminating some entry at the end. In this way, a LRU stack is constructed which contains the codes of words that have been used most recently eliminating entries that have been used least recently. This stack may be updated on searches as well in order to fulfill searches for frequently used terms even faster. Thus, this list serves for storing frequent items that do not reside in the inverted index of a node and avoids loading this inverted index in some cases.

A combined search would check the initial inverted index of the textual column first. If the number of referenced entries in this index exceeds the limit *l* there for one particular word, it will be marked for having to be searched in the relational part of the index structure. Otherwise, the relational criteria may be checked directly at this point. The search in the initial inverted index produces a list of entries overflowing the *l,* which will be searched in the relational index as a next step. Searching the relational part will first check the "id list", if any of the word codes found in the initial search is contained there. If it is not contained in the id list, it will be searched inside the inverted index attached to the particular relational node. If the entry has been found in the "id list", it is obvious, that the particular entry has to reside in some lower level. The relational part of the search query has to be checked simultaneously, as well in order to retrieve only those items that fulfill all of the given distinct criteria. If some items cannot be found in the current inverted index, these might be referred to in some lower level, which makes the search continue recursively. In this way the tree is traversed and all candidate items are generated, which meet all given search criteria. As a final step, the generated set of entries that drop out of the combined search in the hybrid part of the index structure has to be merged with the initial set to produce the final set of entries which fulfills all the search criteria.

IV. ANALYSIS OF THE NEW INDEX STRUCTURE

This section analyses key properties of the new index structure. Due to the complexity of a hybrid index structure a detailed worst case analysis would result in loose upper bounds of space and time at present. Nevertheless the analysis will provide some constraints indicating a reasonable space complexity and an acceptable search time complexity for most but not all cases.

An inverted index associated to a node contains a posting for every word occurring in some text and references for the list of entries (documents). According to Heaps´ Law ([7]) the number of distinct words *w* in a text grows with the formula $k \bullet n^b$ where *n* is the number of documents and the parameters *k* and *b* depend on the average size, source and language of a text. The parameter *b* is usually close to 0.5. Therefore this formula can be approximated by $k \bullet \sqrt{n}$.

The inverted index for a node below the root will only contain terms which do not reference more than a given limit *l* of entries. With Zipf's Law [11] the frequency of a term is given by $f = n \bullet c/i$ where *i* denotes the *i*-th frequent term and *c* is a text dependent constant. The number of terms occurring in more than l documents can be deduced from this formula:

$$i = n \bullet c/l$$

As a consequence the total number of terms in such an inverted index is given by the following formula:

$$terms = k \bullet \sqrt{n} - n \bullet c/l \qquad (1)$$

In this formula the first term with the square root of *n* grows slower than the second linear term. The function value increases up to a maximum and falls to zero afterwards meaning that all terms are moved to the child nodes for a large number of documents. The value *n* for which this formula reaches its maximum is given by this formula:

$$n = ((k \bullet l)/(2 \bullet c))^2 \qquad (2)$$

Inserting the *n* from formula (2) into formula (1) provides a fixed upper bound for the number of terms, depending only on the constants *c*, *k* and *l*:

$$terms \leq k^2 \bullet l/(4 \bullet c)$$

In addition to the inverted indexes assigned to nodes we need to consider the inverted indexes assigned to points in leaf nodes. In general the number of entries (and documents) grows probably proportional with these points. Therefore the number of entries assigned to points in the tree is constant (probably very small). As a consequence the size of all inverted indexes is limited by a constant and the required space for the hybrid index grows linearly with the number of nodes in the tree. It is a well known property of tree structures that the number of nodes grows also linearly with the number of entries in a table. Therefore the space of the hybrid index structure grows linearly with the number of entries in the table.

The time complexity is not so easy to analyze since very special situations may occur where the search region overlaps with a large number of nodes and all of them include the search term in their associated index. In this case it may happen that all points from these nodes including this term are not present in the area of overlap. Examples for this situation require a very specific distribution of the points in multi-dimensional space which is usually not given in real applications.

A logarithmic search time complexity is probably obvious, if the search condition overlaps with one or only few leaf nodes (one or few paths need to be searched). This situation is also given for other existing index structures.

The key strength of our structure is that a search is continued below a node only if the search term occurs frequently below the node. Otherwise the associated inverted index would be visited and the search would end here.

With a high frequency of a term below a node it is likely that many child nodes overlapping with the search region contribute at least one entry to the search result. If we assume that every node on a given level of the tree contributes one

entry to a result set (which is not guaranteed), then the number of visited nodes on this level is limited by the size *m* of the result set and not the size of the table. Since the number of levels in the tree grows logarithmically with the number of entries the search time would be limited by O(log(*n*) •*m*).

V. EXPERIMENTAL RESULTS

Several tests were performed to prove the efficiency of the newly proposed hybrid index structure. Two different datasets were used as inputs for these tests. One of them are geo-tagged Wikipedia articles ([2]) and the other one is a synthetic dataset that is constructed based on several rules.

The index structures used in the tests are B-Trees as implementations of the inverted index part and an adopted version of the R-Tree containing the extensions, mentioned in the previous section, as the relational index. The size of the "id lists" was **not** limited in the tests, because they were performed in main memory, instead of on disk.

The text contained in the Wikipedia articles was preprocessed, that means a stemming and normalization of the particular texts took place.

The synthetic dataset primarily consists of a set of randomly generated documents with some rules applied to the generation. These rules refer to the generation of the words, the assignment of words to documents and the point coordinates, respectively. The distribution of words to documents has been done by assigning a probability value of a word appearing in a certain document based on the distribution of Heaps´ Law ([7]). The generation of points follows, generally, a Gaussian distribution. A configurable number of points are assigned to every document.

A query for the hybrid index structure contains two parts:

1. A relational search criterion that defines an upper and a lower bound for each relational column, in this case a two dimensional range
2. A textual criterion that consists of a set of search words to be found in the textual column

Each dataset was inserted to a predefined level ranging from 1000 to 64000 documents and a workload of 200 random queries was produced for each generated structure. The queries, again, followed some predefined rules. These rules are set up to guarantee fair conditions for the structures ensuring similar sizes for all result sets. This means that the words are scanned in advance to determine some lower, median and upper bound of occurrence of the distinct words. The location and size of the search region are generated by using a uniform distribution. The sizes of the regions are extended by additional deltas in all dimensions for the different datasets. The sizes of these deltas are chosen inversely proportional with the sizes of the datasets to ensure similar result set sizes.

The parameter *l* mentioned previously was set to an arbitrary number, whereas 100 turned out to be a reasonable amount for these datasets. Generally this parameter should be in a range from 100 to 200 in order to make it possible for the structure to organize properly while limiting sequential scans in an inverted index to a reasonable amount.

Figure 3 illustrates the search time complexity by summarizing the number of all hybrid nodes for all 200 queries relative to the number of documents in the dataset. This figure uses the Wikipedia articles. The amount of documents inserted can be seen on x-axis whereas the y-axis shows the total

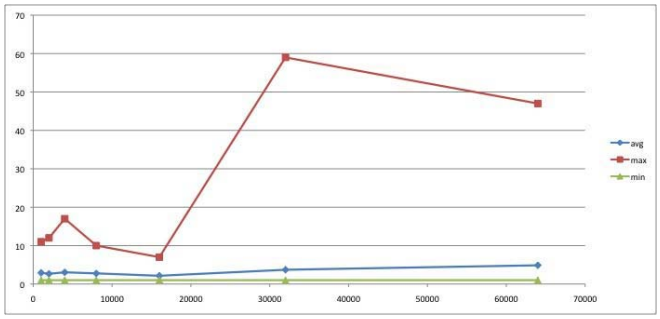


Figure 3: Aggregate Numbers of Hybrid Nodes Searched

amount of hybrid nodes loaded during one search. The curve of the minimum nodes searched is constant to one. The amount of average nodes searched only grows slowly when inspecting a structure having more documents inserted. Regarding the maximum nodes loaded during one search, there is no clear tendency, but it also grows with an increasing number of documents. However, the average number of hybrid index nodes searched only grows slightly (below linear in this exponential scale) with the amount of documents inserted. The fact that the maximum number of hybrid nodes searched grows this fast on 32000 documents is based on the certain predefined randomness of the search query. But the tendency of the average amount of inner nodes loaded during one query shows, that the new structure ensures a logarithmic time complexity for this example.

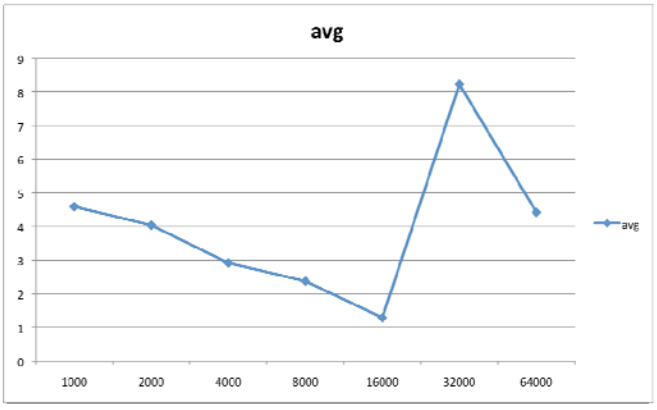


Figure 4: Average Amount of B-Tress loaded

Figure 4 refers to the same set Wikipedia articles. In this picture, it is obvious that only a small portion of the total amount of B-Trees present in the entire tree has been loaded in order to generate the final result set. The maximum of the B-

[2] http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt Georeferenzierung/ Wikipedia-World, last visit: 2.6.09

Trees in the average case here is about eight during one particular search. This means that using the approach of the "id lists" reduces the need of loading many secondary indexes from disk by filtering out unneeded search operations based on the contents of the LRU stack. However the deflection in the case of 32000 documents results, again from the randomness in the relational search criteria, which can also be noticed in Figure 3.

Figure 5 shows the length of the "id lists" for the different

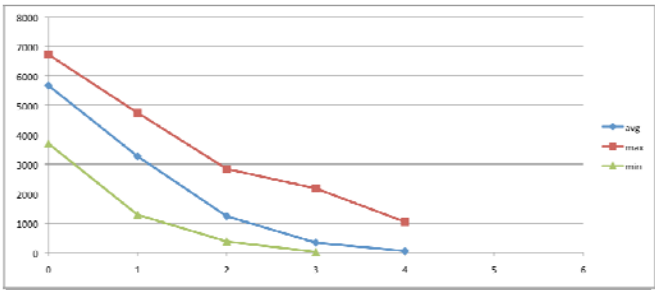


Figure 5: Distribution of "id lists" in separate levels

levels of the relational part of the tree. The generated hybrid R-Tree has five levels, whereas there are (obviously) no lists at the leaf nodes, because these lists will only be generated when items are pushed down to a level below. However, the distribution of the lists again shows that the distribution of words and documents is strongly related to Heaps Law ([7]).

Figure 6 analyses the size of the hybrid index structure

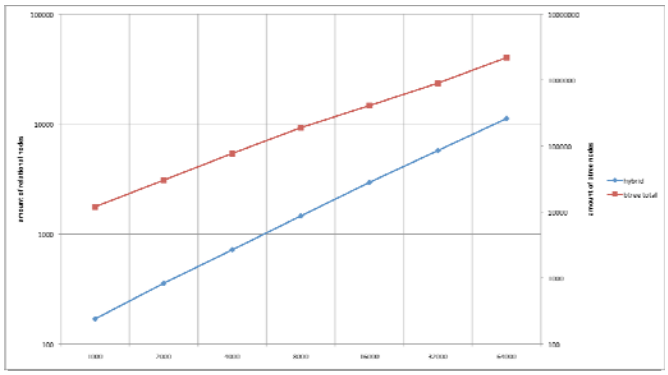


Figure 6: Relational and B-Tree nodes

relative to the number of entries. The scale of the left axis is valid for the relational nodes (R-Tree), whereas the right axis is valid for the B-Tree nodes inside the tree structure. It can be seen that the total amount of nodes inside the tree grows linearly with the number of entries. This confirms the theoretical analysis of the previous section.

An inspection of the synthetic dataset showed similar results. The experimental results of the synthetic datasets only differ slightly from the ones given above and thus, there is no need to show them here.

VI. CONCLUSION

The proposed index structure overcomes problems of previously introduced hybrid index structures. In comparison to existing index structures it provides efficient support for a wide range of queries together with reasonable space requirements. This could be demonstrated by our analysis and first tests.

A very interesting property of the index structure is the small number of entries referenced by words in the inverted indexes. This approach facilitates efficient support for queries with a larger number of terms since only small intermediate result sets need to be processed for generating the final result. The reason is that the relational part of this index structure facilitates the division of an initial inverted index into smaller parts. Therefore this hybrid index could even ensure a better efficiency for combined searches than an inverted index for conventional full text searches.

The general structure of our approach is not limited to the combination of relational and textual search criteria. Postings in our inverted indexes could also assign features to objects like images containing these features. With this approach the index structure may also serve as a general framework for constructing further index structures supporting general multi-valued attributes.

REFERENCES

[1] R. Bayer, C. McCreight: "Organization and Maintenance of Large Ordered Indexes," Acta Informatica 1(3), pp.173-189, 1972.

[2] N. Beckmann, H. P. Kriegel, R. Schneider, B. Seeger: "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles", Proc. ACM-SIGMOD International Conference on Management of Data, pages 322 - 331,,Atlantic City (NY), May 1990

[3] I. Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. Proceedings of 24th International Conference on Data Engineering, pages 656 - 665, Cancun, Mexico, 2008.

[4] R. Göbel, A. de la Cruz, "Computer Science Challenges for Retrieving Security Related Information from the Internet," GMOSS Publication, 2007

[5] R. Göbel, Andreas Henrich, Raik Niemann, Daniel Blank: "A Hybrid Index Structure for Geo-Textual Searches", Proc. 18th ACM Conference on Information and Knowledge Management (CIKM), Honkong 2009

[6] R. Hariharan, B. Hore, C. Li, and S. Mehrotra. Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems. Proceedings of 19th International Conference on Scientific and Statistical Database Management, Banff, Canada, 2007.

[7] H. S. Heaps, "Information Retrieval: Computational and Conceptual Aspects", Acadamic Press, 1978

[8] A. Guttman, "R-Trees, A Dynamic Index Structure for Spatial Searching," Proc. ACM SIGMOD Conference, Boston, pp.47-57, 1984.

[9] D. Zhang, Y. Chee, A. Mondal, A. Tung, and M. Kitsuregawa. Keyword search in spatial databases: Towards searching by document. Proceedings of IEEE International Conference on Data Engineering, pages 688 - 699, Shanghai, China, 2009.

[10] Y. Zhou, X. Xie, C. Wang, Y. Gong, W.Y. Ma, "Hybrid Index Structures for Location-based Web Search," Proceedings of the 14th ACM international conference on information and knowledge management, Bremen 2005.

[11] G. K. Zipf: Human Behavior and the Principle of Least-Effort. Addison-Wesley, 1949

Common Interoperability Framework for Defense Architectures, A Web Semantic approach

Abdullah S. Alghamdi
Department of Software Engineering
College of Computer & Information Sciences
King Saud University, Riyadh, Saudi Arabia
Ghamdi@ccis.ksu.edu.sa

Zeeshan Siddiqui
Department of Software Engineering
College of Computer & Information Sciences
King Saud University, Riyadh, Saudi Arabia
Siddikui@gmail.com

Abstract— **Critical information exchange issues are alerting its users to divert their attention on this most vital subject. As for now the interoperability issues are not yet settled completely, many integration frameworks have been developed to overcome these issues. In this research we have developed an integration channel which uses W3C Semantic Web Specification techniques; we have named it *Common Interoperability Framework for Defense Architectures*. The model integrate the application architecture sensitive data in the form of RDF files on a secure and responsive security channel which is then picked up by the web crawlers, we have named as RXCrawl. The data resources are then gathered and collected in the form of metadata repository. We will only discuss the main functionality, usage, working and methods, RDF files description and functionality, RDF/XML engine functionality and the security channel description and web crawlers usage and methodology. This study is to overcome current interoperability issues between defense architectures, we will leave further details such as, the whole architecture framework's implementation, testing and performance measure, for our upcoming research papers.**

Keywords- Web Semantics, Semantic Interoperability, Information data-exchange, Enterprise Architecture (EA), Software Engineering (SE).

I. INTRODUCTION

Department of Defense (DOD) [1] is spotlighting and emphasizing more on interoperability issues than any other issue. While studying the Department of Defense Architecture Framework (DODAF) [2, 8 and 9] one can easily grasp the main point that is being covered up i.e. the interoperability issue. The focus of this research is on defense domain as compared to any other domain. Defense domains are considered the most crucial domains, therefore, DODAF covers interoperability issues in a more precise and discrete way than in any other architecture i.e. TOGAF [3], Zachman Framework [4], GERAM [5], CIMOSA [6], PERA [7] and more. Although, all these frameworks mentioned are also being utilized in the defense domain.

The data integration between defense domains is still a massive challenge as the whole architecture will affect the country-wide changes. The interoperability of different domains together cannot be resolved fully, many ideas were proposed and different architectures are developed, but none of them fully cover the system integration.

This paper covers the interoperability issue by following the techniques defined in W3C standards. In our view the data integration can be handled more practically and smoothly by following the standards and implementing web based interoperability architectures. Therefore, we have focused on the web semantic interoperability techniques to integrate sensitive defense applications to interchange mission critical data.

The rest of the paper is divided into following sections: Section II will cover the Background of this study and related work; Section III will discuss the Architecture and its working. Section IV will spread light on the Security Channel of the architecture,. Section V will discuss the conclusion, future aspects and acknowledgment part.

II. BACKGROUND & RELATED STUDY

To understand correctly that what is sent from a post to the execution post is the main key to achieve maximum interoperability. Semantic interoperability teaches us to exchange information robotically and to provide and develop such means to exchange that information correctly, securely and efficiently. Semantic interoperability provides preset transformation of the data from the source to the destination enclosed by the security to encrypt/decrypt data and to pass on the data to the requester as accurately as it was dispatched. Therefore, semantic interoperability can be defined as the robotic transformation of data on a wide and secure channel maintaining its original state throughout the process. In our approach, we have used Resource Description Framework (RDF) technique to collect the architecture data in a RDF file. [10] The W3C practice of RDF is a popular way of presenting semantically tagged data online. It is a framework for organizing metadata on the web. The RDF presents the metadata interoperability through common convention support to channel the data across different domains. The RDF document language is called RDF/XML and it is easy to exchange RDF based information on cross platforms. For more study about RDF see [10]

A. Related Study & Literature Review

Michael R. Hieb and Andreas Tolk build a component for integrating M&S with C4ISR for supporting future military operations. It was a position paper in the conference, authors have stressed upon the necessity for a

common architecture, common ontology, common set of algorithms and a superior technique to bring all this down together. [11]

Robert Suzic et al highlighted the Information exchange requirements (IER) and models (IEM). They have articulated the details of how to build information exchange model on the basis of information exchange requirements. [12]

J.H. ter Bekke in his paper Advantages of a Compact Semantic Meta Model, explains the XPLAIN data model and designed an effective conceptual framework for databases. [13]

Other then defense domains, SemanticWeb Meta-Data model was developed for e-learning by Mikael Nilsson, Matthias Palmer and Ambjorn Naeve. [14]

David Perme & Andreas Tolk, presented a Common Battle management language to integrate air & ground operations within. In this paper they have presented a BML language comprises of three views. [15]

Hans W. Polzer discussed the concept of net-centricity and service orientation from a system implementer view [16]

Paul Shaw and David J. Roberts presented a cross platform information exchange framework (CIEF), they discussed an operational design for the publication, location, and subscription to information in the correct mission context and monitor the operational use of information in the contest. [17].

III. Framework Architecture & its Working

In joint collaborations between different defenses nodes i.e. air, ground and sea with in a country or globally, the best way of information interchange is through web based architectures. In our framework model we have considered three different applications running under three different army domains i.e. army, air-force & navy. Now, considering these domains, we have following interoperability issues:

- Platform dependencies issues
- Architecture differences issues
- Software customization issues
- Technical terminologies difference issues
- Interface issues
- Database and repositories alignment issues, and
- Information security issues
- Many other interoperability issues

The framework and architecture should easily absorb the data transferring from different repositories and databases. Security measures must be taken to secure the propagation and transformation of the data between different architectures. We have designed this architecture to cover the information handling from all these aspects. In our research we have three systems having different operating systems and OS specific software application, either web based or desktop applications. As illustrated in below figure 1.0.

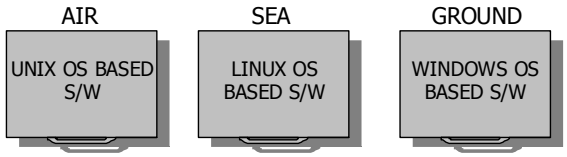


Figure 1.0, Different Defense Command Applications

These three defense commands mentioned and illustrated in figure 1.0 are using different military software application based on their operating system and requirements. There are many software applications for example CACI products [18] and these applications are following organization specific and customized military standards for example MIL-STAND [19].

These applications software are based on different enterprise architectures, i.e. Oracle, IBM and .Net architectures. The air command UNIX OS based app is using IBM enterprise architecture; sea command LINUX OS based software app is based on Oracle enterprise architecture and ground Command and Control (C2) Windows OS based software app is utilizing .Net framework architecture platform. As illustrated below in Figure 02.

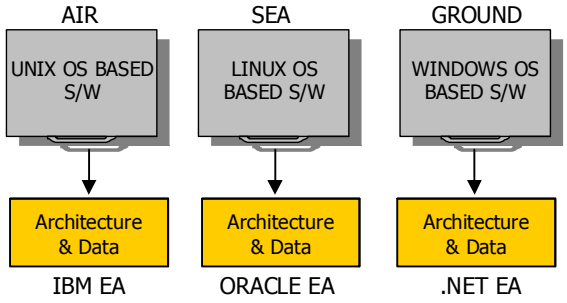


Figure 2.0, Air/Sea/Ground C2 Architecture & data based on different enterprise architecture

The architecture contains mission critical and mission crucial data that is highly sensitive and at the same time extremely desirable to interchange and to pass-on using an efficient and secure transformation layer or set of procedures in the form of an architecture that fetches the data and transform it in the original state as it was received.

Each architecture is transforming its information in the form of a document that is analyzing the data on the bases of common elements analysis and presentation. Beside what ever application is being used over what ever platform, the main source of data provision is the RDF file

that is created using RDF/XML language by analyzing common elements among different architectures.

A. *RDF/XML data plotting*

RDF uses tags that are XML tags, through which information is written in triples. There are two ways of RDF to streamline RDF data; Storage of data in the form of triples to query through a specific system using RDF ways, and, mapping of data with RDF from relational database.

The technique we have used in this research is the plotting of data already stored in a relational database with RDF files. In this plotting, the main role is the usage of *Ontologies* [20]. The terms which are used in RDF are described using Ontologies; describing them in-between relationships and rules to implicate. When the request is sent for the data from any RDBMS, it first passes to the RDF/SQL plotting engine and then this plotting will provide the requested data to the requester in response of the query.

B. *RDF/XML Plotting Function*

The RDF/SQL plotting engine comprises of two components; Ontology and Structure of SQL data. In our architecture the requester is a semantic web crawler, we have named it as RDF/XML crawler (*RXCrawl)*. Mainly these semantic crawlers are known as 'Scutter' [21] in the community. The simple web crawler only functions on HTML document and uses HTML specific anchors. Whereas, semantic web crawler differs in two ways, it also runs on RDF based documents other than only HTML and link condition in between information resources. The following figure 3.0 illustrates the above discussion.

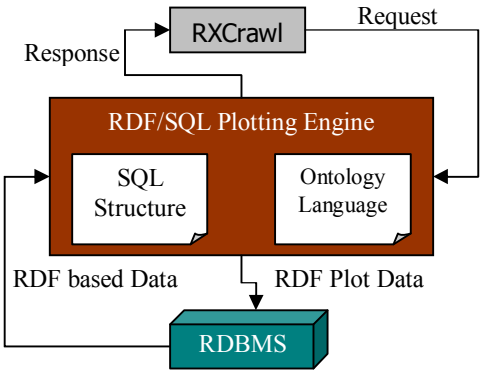


Figure 3.0, RDF/SQL Plotting Engine

C. *INSERT into RXCrawl*

The RXCrawl semantic crawler is responsible for querying the required crucial mission related data from different architectures. Following set of instructions is trailed to accomplish this.

Using API functions
Http POST (SPARQL query)
Http PUT (RDF+XML)
SPARQL *insert* using LOAD

IV. RDF/XML Security Architecture

Till now we have discussed the general working of our framework architecture, the vital perception of any architecture is its security policies and security framework. The security of any information exchange framework is the set of rules that should be followed strictly to design that transformation channel non fragile for the attackers or hackers. The information can be hacked or decrypted at many levels, i.e.

At the level of applications that are responsible for generating the data,
During the process of accessing the data,
During the transformation channel, and
During the process of writing the data to the recipient or at the recipient end.

Our RDF/XML security framework is based on the following aspects:

Security policies at the RDF/XML document level
Security Architecture of the RDF/XML document security
Object level security of the data nodes

Now, data association is covered using *DTD* and Tree representation. There are two types of security, policy based security and access level security. In our approach we have used the access level security

A. *Access level security architecture discussion*

Security violation is checked on the basis of the data request query and the result submitted to the XML engine. The tree contains expanded checks for violation. *RXCrawl* usage record from history file is used to combine with the result when there is no security policy violation. First the module is checked for violation and after that extended trees which are newly created. This history or Crawler usage record is updated and result is sent back to the crawler only in case of no security violation. Query request is rejected when security violation is observed at any level/phase. The request engine is accountable for engendering query responses. Admittance Control unit is involved in appraising consents of the crawler with support to security policy and data collection of the crawler.

B. *Algorithm test for the security architecture*

We have used several algorithms to test the policy based security architecture. Few tests were;

Security architecture request and response algorithm
Module security check

The security architecture has given the access level protection on the RDF/XML document generated as a result collected from different military architectures. Figure 4.0 will illustrate the RDF/XML security architecture.

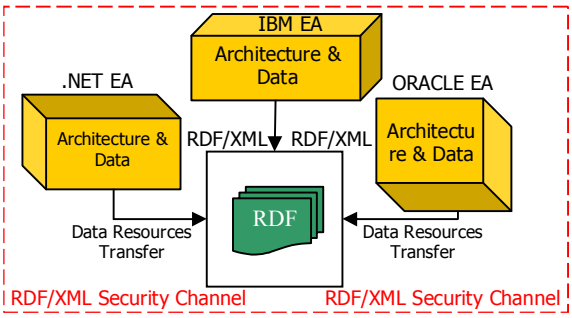


Figure 4.0, RDF/XML security area

The data which is accessed by *RXCrawl,* utilizing all web semantics techniques what we have discussed above and following the security rules which we have just covered, the *metadata* is collected in a *Metadata Repository* for further usage according to the mission needs and directions. Following figure 5.0 illustrates the whole sections discussion.

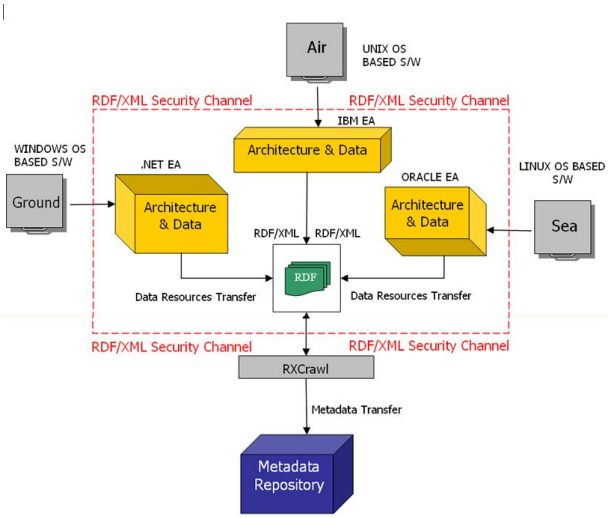


Figure 5.0, Common Interoperability Framework for Defense Architectures, CIFD

V. CONCLUSION

The platform dependency issue was covered when the multiple coating is performed by RDF/XML plotting engine. The architecture dependency is no more an issue if this measure can be exercised. Software customization or in our scenario defense software customization will have no affect on the architecture, as the applications are not directly sharing the data, the main communication is initiated using RDF files. Technical terminologies is handled by adopting the concepts of Ontologies, the main integrated interface is the RDF/XML plotting engine. The database repositories alignment is done by engine's SQL representations. The security aspects are widely handled by applying and implementing access control document level security on RDF/XML document.

REFERENCES

[1] Department of Defense united states, http://webaddress

[2] The Department of Defense Architecture Framework (DoDAF), version 2.0, *Volume 01: Introduction, Overview & Concepts, Manager's guide*, May 2009

[3] The Open Group Architecture Framework (TOGAF), TOGAF Version 8.1.1, online documentation, http://www.opengroup.org/architecture/togaf8-doc/arch/

[4] ZACHMAN International Enterprise Framework, John Zachman's concise definition of the Zachman Framework

[5] Generalised Enterprise Reference Architecture and Methodology (GERAM), International Conference on Enterprise Reference Architecture and Methodology

[6] Computer Integrated Manufacturing Open System Architecture (CIMOSA), A Primer on key concepts, purpose and business value, Online primer

[7] Williams, T., The Purdue Enterprise Reference Architecture, Computer & Industry archive, Volume 24, Issue 2-3, page 141-158

[8] The Department of Defense Architecture Framework (DoDAF), version 2.0, *Volume 02: Architectural Data and Model, Architect's guide*, May 2009

[9] M. Alessio, *presentation on DoD Architecture Framework Overview*, Operational Systems technical, May 2004

[10] Resource Description Framework RDF, *Model and Syntax Specification*, World Wide Web Consortium W3C

[11] Tork. A and Michael. R, *Building & Integrating M&S Components into C4ISR Systems for Supporting Future Military Operations*, International Conference on Grand challenges for Modeling and Simulation (ICGCMS '03), 2003

[12] Suzic. R and Choong, *Information Exchange Requirements and Information Exchange Model*, Proceedings of MDISIS 2008

[13] Bekke. J.H, *Advantages of a Compact Semantic Model,* 2nd IEEE Metadata Conference, Silver Spring

[14] Nilsson. M, Palmer. M, *SemanticWeb Metadata model for e-learning*, Proceedings of 11th WWW Conference, Hawaii, USA, 2002

[15] Perme. D and Tork. A, *Common Battle Management Language to integrate air & ground operations,* Spring Simulation Interoperability Workshop, San Diego, CA, April 2005

[16] Hans W. Polzer, *The Essence of Net-Centricity and Implications for C4I Services Interoperability,* AFCEA C4I Center Symposium Critical Issues in C4I, 20-21 May 2008, GMU, Fairfax Virinia Campus

[17] Shaw. P and David. J, *Cross Platform Information Exchange Framework (CIEF)*

[18] Consolidated Analysis Center Incorporation, http://www.caci.com, 1967

[19] United States Military standards, Mil-Stand, http:// http://dsp.dla.mil/APP_UIL/displayPage.aspx?action=content&contentid=66

[20] Collard. M, *Ontologies-based databases and information systems,* First & Second VLDB workshop, ODBIS/2005-2006, Trondheim, Norway, September 2-3, 2005, Seoul Korea, September 11, 2006, revised papers

[21] Biddulph, *Crawling the Semantic Web*, 20 04

Assessing Open Archive OAI-PMH implementations

Emanuele Bellini, Marcel Aime Deussom, Paolo Nesi
Distributed Systems and Internet Technology Laboratory
University of Florence, Department of Systems and Informatics
Http://www.disit.dsi.unifi.it
ebellini@dsi.unifi.it , nesi@dsi.unifi.it

Abstract

The number of works made freely accessible on the Web by institutions joined to Open Archive Initiatives such as libraries, research institutions, scientific and cultural archives, is constantly growing. Until now some international agreements among key players of Open Access approach have been established but the low level of standardization and the lack of a digital resource and repository certification authority prevents any quality control on the works and subsequently their trustable dissemination within the cultural and scientific community. In this scenario, the benefit of having freely accessible resources is lost. The study proposed in this paper aims at providing a statistical analysis of the state of art of OA implementations in order to outline weak practices, propose assessment metrics for self analysis and stimulate a general improvement of the quality of OA implementations.

1 Introduction

The Web has drastically changed the information environment where users of the humanities work and study and the information needs to be more accessible to become collective knowledge. To this end, following the initiatives as Berlin Declaration on Open Access, OA, to Knowledge in the Sciences and Humanities [12], a lot of institutions are implementing open access institutional archives. The main mission of the OA is the dissemination of the digital resources and their metadata. In order to address this aim, a number of software tools is available for implementing an open institutional repository such as dspace [8], Fedora [9], eprints [10], etc. To this end, the Open Archive Initiative has implemented the OAI-PMH protocol [1] for publishing and thus making possible the metadata harvesting among repositories. The OAI architecture identifies two logical roles: "Data Providers" and "Service Providers". Data Providers deal with both the deposit and the publication of resources in a repository they "expose" to provide the metadata about resources in the repository. They are the creators and keepers of the metadata and repositories of corresponding resources (digital items, digital essences, which are the effective files). Service Providers use the OAI-PMH interfaces of the Service Providers to collect and store their metadata as shown in [19] and [25]. They use the collected metadata for the purpose of providing one or more services across all the collected metadata like Pleiadi [20], Citeseer [21]. The types of services which may be offered include a query/search interface, peer-review system, cross linking, etc. Recently, an architectural shift was to move away from only human supporting end-user interfaces for each repository, in favour of both human end-user interfaces and machine interfaces for data collecting.

In this paper, the objective is highlight the features needed to evaluate the effectiveness of data provider implementations in terms of OAI-PMH. The quality of a Data Provider implementation and collected metadata can be evaluated in terms of: completeness and accuracy of metadata description, trust-ability of service, level of standard compliance and adoption, etc.

Therefore, the main goals of this paper are:

- To propose and provide a set of metrics to be used as statistical tool for assessing and monitoring the OAI-PMH service implementations in terms of quality, trustworthiness and standard compliance.
- To achieve a better comprehension on Open Access implementation weakness, stimulating and directing new efforts towards technology, policy, standardization levels since the usage of the current widespread solutions is too vague to be exploited at a reasonable cost in the open world.

In order to assess Open Archive implementations in terms of OAI-PMH services, a preliminary harvesting of a large number of OAs has been performed. This allowed the identification of a number of issues and common drawbacks about these services which could be praiseworthy to discuss about. They are summarised as:

- huge number of OA responding to OAI-PMH protocol, more than 2000. The experiment is based on crawling archives and repositories listed at: http://www.openarchives.org/Register/ListFriends [11];
- a number of different metadata sets for each items in each repository, not always the same, frequently not the same for the whole repository. The used metadata are not 100% compliant with those standards;
- a huge number of total records to be collected (the union of records of all the repositories for all the metadata and formats), more than 18 millions.
- a slow mechanism/protocol to access OAI-PMH repositories since servers and connections are slow, and the protocol itself is of high level.

To this end, in order to evaluate the typical Open Archive implementation with respect to the level of spread, metadata quality and service reliability, it has been necessary to perform a general harvesting and processing of the metadata collected from the data providers.

2 GRID Approach for Harvesting

The solution approach is based on OAI-PMH protocol, a REST-based full Web Service that exploits the HTTP protocol to communicate among computers, using either the GET or the POST methods for sending requests. It is well-known that web services are also a computing technique for systematically disseminating XML contents, but when the global amount of data increases, some problems come out. In the following paragraph, we outlined these issues requiring a parallelization approach to be tackled.

2.1 Harvesting from a huge number of repositories issues

According to OAI-PMH protocol, *Guidelines for Harvesting Implements* [24] and OA implementation tutorial [23], a client may put a request to OAI server to ask for the stored content descriptors. Answers are related to the accessible records, and adopted formats. The OAI-PMH protocol provides a list of discrete entities (metadata records) by XML stream. In many cases, these lists may be large and it may be practical to partition them among a series of requests and responses. In fact, the repository replies to a *list* request with an incomplete list and a resumption Token. In order to get responses as much as possible from the list of the OAs considered, the harvester has been performed more requests with resumption Token as arguments. The complete list then consists of the concatenation of the incomplete lists from the sequence of requests, known as a *list request sequence* [1].
Moreover, in the current version of the OAI-PMH protocol a 'verb' to obtain the number of the records that we are going to harvest is not defined. Thus it is impossible to estimate a priori the duration of the process in terms of counted metadata sets. It is clear that the number of records included in a incomplete list (or page) affects the harvesting performance. In some cases this number was only one, and yet the harvester had to perform requests as many as the records in the archive. The harvesting performance also depends on response delay that is related to the network bandwidth and machine performance used by the connected Open Archive. In some cases, this time was greater than 15s for each request. In order to cope with the complexity, a parallel solution has been set up and used as described in the next subsection.

2.2 GRID based metadata harvesting architecture

As it occurs with a web crawler, the harvester contacts and inspects the OA data providers automatically and it extracts metadata sets associated with digital objects via OAI-PMH protocol. Because of the computational weight of these processes, the harvester has been implemented by using the grid based parallel processing on DISIT cloud computing infrastructure. The grid solution has been realized by using AXMEDIS Content Processing (AXCP GRID) [3]. The computational solution has been implemented by realizing a parallel processing algorithm written in AXCP Extended JavaScript [17]. The algorithm has been allocated as a set of periodic processes replicated on a number of grid nodes, typically from 1 to 15 max. The process is managed by the AXCP Scheduler. It is possible to put in execution a number of rules that are distributed to the available grid nodes. Each rule can be periodically (or on demand) scheduled with an interval, for instance, of 1 minute from each running on a single node and the successive one. Each rule is a 'harvester' executor of an OAI-PMH request to obtain the metadata records, parsing the XML response and storing information in our local database called CHonline. In Figure 1, a schema of the architecture is shown.

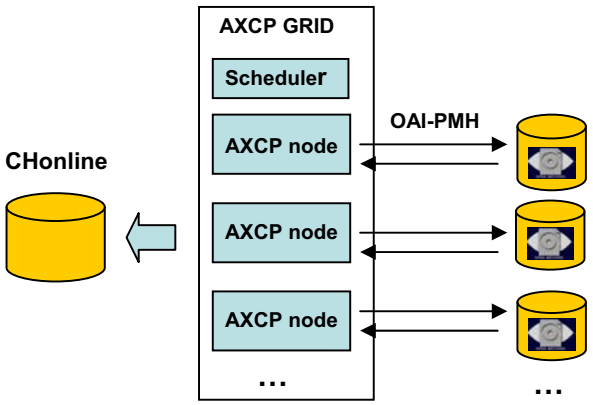


Figure 1 – Grid architecture for massive OA harvesting via OAI-PMH protocol on grid infrastructure of DISIT.

Each GRID node executes an identical autonomous harvesting rule that collects metadata from an Open Archive and populates the database according to the general status also collected into the database. This solution reduces the computational time up to a factor equal to the number of nodes used for completing the harvesting of repositories. In effect, the parallel solution is not only an advantage for the speed up, but also for the reduction of the time needed to get a new global version of the metadata collected in the OI repositories.

2.3 GRID-based harvesting workflow

This paragraph describes the grid base harvesting algorithm and workflow. Figure 2 shows a schema representing the consecutive steps performed by the harvesting rules on the grid.
Before performing the effective harvesting of the single records, two preparatory steps are needed: (i) to get the repositories information; (ii)_to get the metadata sets available for each repository. These two steps are performed into the grid with specific aperiodic/on-demand rules.
During the *first step* a rule for getting the repository list from http://www.openarchives.org/Register/ListFriends website is launched. This rule parses the XML list of OA repositories baseURLs and populates the repository table

of CHonline database. For example, a segment of the repository list is as follows:

[…]

<baseURL id="UOV.es">http://www.tdr.cesca.es/TDR_UOV/NDLTD-OAI/oai.pl</baseURL>

<baseURL>http://diglib.cib.unibo.it/oai/oai2.php</baseURL>

<baseURL>http://docinsa.insa-lyon.fr/oai/oai2.php</baseURL>

[…]

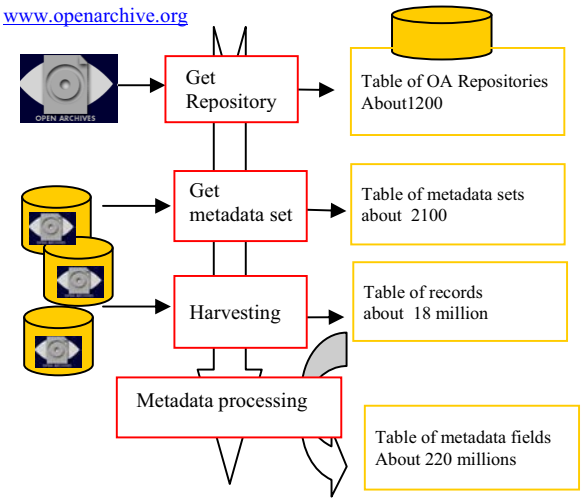


Figure 2 – Algorithm for harvesting

The repositories are identified with a <baseURL> field filled in with the URL of repository OAI-PMH interface and the repository ID. This rule may be also periodically scheduled for checking the availability of new repositories added to the list that have to be harvested by the system.

Once the repository list is obtained, the *second step* has to determine if the OA is active and which service may provide. To this end, a dedicated second rule is activated to both verify the activity of the OA and retrieve the metadata formats available by using the *ListMetadataFormats* verb of OAI-PMH. A repository is set 'not available' if it does not provide any response, so that at the next round it can be tried again. In fact, it may happen that a repository may be offline for some reasons. Therefore, each single OA provides the list of metadata according to the following example.

```
http://baseURL/request?verb=ListMetadataFormat
[
<ListMetadataFormats>
<metadataFormat>
<metadataPrefix>oai_dc</metadataPrefix>
<schema> http://www.openarchives.org/OAI/2.0/oai_dc.xsd
</schema>
<metadataNamespace>http://www.openarchives.org/OAI/2.0/o
ai_dc/</metadataNamespace>
</metadataFormat>
. . . . . .
</ListMetadataFormats>
```

The list of metadata sets of each repository is stored in the metadata formats table of CHonline database. It should be noted that if a metadata format is declared as being supported by an OA, this does not mean that it is available for all the items in the repository.

OA Harvesting. The harvesting rule gets access to the status table in CHonline database to obtain the first not processed archive/metadata-set and it starts with its crawling. Moreover, the harvesting rule parses the XML response, it extracts only the metadata information and it saves it in a single database field/chunk as a string. The harvesting rule is designed to harvest the records only from one repository managing the resumption Token. This approach is meant to reduce the rule time activity, but there are some cases where a rule could stay alive for hours (for instance if there are a lot of records to harvest and the OAI request has provided a short number of records)

Metadata processing. The metadata harvesting is the first step to collect data and per se it not sufficient to evaluate the quality of metadata implementation. In fact, it is not possible to extract specific metadata values that are related to a specific argument. Moreover the high number of implemented different metadata sets requires a tool for processing them in order to get the single metadata element. In the preliminary analysis, the harvester was tuned to collect at most 100 records for each metadata format managed by each OA. The preliminary process has been used to tune the processing according to the large number of different metadata sets.

Moreover, an additional grid rule got the XML of each non processed record stored in the database and it extracted the single fields. Therefore, each field of each specific record has been stored with its value, type, and additional information in the CHonline database. This poses the basis to perform a deeper analysis, as described in the following. This process led to a sort of an extended RDF [23] model and thus to a metadata normalization allowing queries on the single fields. This table turned out to be very huge (for each field of each metadata record a detailed field record is generated. For instance 15 new records are generated from a single DC based metadata record). The resulting table of single fields (of about 220 millions of records) has been mainly used as a qualitative metadata evaluation for the purpose of this paper. In particular, in this paper we are presenting a metric analysis on OA with some comments about the usage of some metadata fields, always aiming at improving the general understanding and quality of the OA.

3 OA Analysis via OAI-PMH

According to the above described solution for massive OA inspection and metadata harvesting, a set of metrics and considerations has been performed. They may be used to evaluate the implementation of OA as an effective tool for

disseminating scientific works via OAI-PMH service protocol. The performed analysis has been focused on understanding problems, practise and weakness at both quantitative and qualitative levels.
The proposed evaluation model has been developed and validated at three levels:

- **General Metrics**, a quantitative evaluation outlining a global view of the diffusion and effective available OA PMH implementation services;
- **Archive Metrics**, quality evaluation related to the usage and trustability of adoption of metadata sets;
- **Metadata Record Metrics** modeled and provided according to the adopted standard, assessing the usage of single metadata fields and their potential weakness.

In the following paragraphs, a selection of the adopted metrics for each level of analysis is provided to evaluate the current worldwide OA PMH services implementation. The proposed metrics are presented together with the values obtained on previously described tests and configurations.

3.1 General Metrics

At the general level a set of considerations has to be performed, which has been identified by the following proposed metrics.

OA available for harvesting. Sometimes the OA are not available for many different reasons but still remain registered on official/unofficial registers with their baseURL. This index is calculated on the basis of the number of answers by repository performing an OAI-PMH *listmetadataformat* request. In our experiment which was based on the general list of OA, only the 74% of them have an active OAI-PMH service. Generally, many institutions begin with a trial OA project within library or archive institutional department with no cost evaluation, usage training, institutional awareness, etc., and most of them run the risk of giving up soon, since the effective commitment was unexpected or underestimated.
OA with working metadata sets. This criterion has allowed evaluating the effective availability of a repository in providing metadata sets. In the general analysis we registered that about the 10% of the OAs have almost an error on a metadata set. The error can be caused by an XML not well-formed or by content not found. A more detailed statistical evaluation about the identified errors and problem needs should take more space to the provided in this short paper.
Metadata sets in the OA community
This criterion highlights the level of fragmentation of the present OA metadata implementations. There are several standards promoted by different communities to describe resources managed by OA, i.e., different metadata standards. The most common are: Dublin Core [6] (generally supported by default), METS [5], MPEG21 DIDL [4] (as a wrapper of other metadata models), etc. It is also well known that metadata sets may be different for different domains, cultural background, etc. For instance, metadata required to catalogue physics resources can be different with respect to those used for media or ICT works, and again different from those adopted for administrative institutional documents, etc. This lack of uniformity has generated several different standards and again for each of them, several different implementations and/or personalization of metadata sets. In our analysis, 153 different metadata schemas have been identified, over only 853 repositories; thus a high percentage. This count aggregates the records on the metadata Schema field (that is mandatory) from metadata format table.
Metadata sets managed by single OA
Typically, an OA provides its records via OAI-PMH service with more than one metadata set. The possibility for the user to select and harvest records with metadata set more structured and richer with respect to simple DC allows an easier metadata processing. For this reason, a systematic analysis to understand better how this possibility has been perceived and exploited, has been performed. In Figure 3, the distribution of the number of metadata set is reported. It is self-evident that most of the OAs prefer providing only one metadata sets, while the usage of two metadata sets is not unusual.

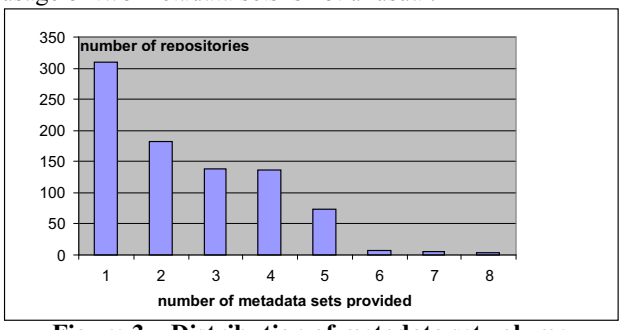


Figure 3 – Distribution of metadata set volume

Adoption of metadata sets
In Table 1, the percentages of the metadata sets' spread among the observed OAs are reported together with their schema and typical prefix. Table 1 reports the spread percentages of 16 most used metadata, with respect to the total number of different sets of 153. Noteworthy is that the DC is largely the most common, while after DC a number of metadata sets is in the range of 8-15% such as RDF, METS, MPEG-21, etc.

N	Prefix	Schema
100%	OAI_DC	http://www.openarchives.org/OAI/2.0/oai_dc.xsd
15%	MARCXML	http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd
15%	METS	http://www.loc.gov/standards/mets/mets.xsd
14%	Rfc1807	http://www.openarchives.org/OAI/1.1/rfc1807.xsd
14%	Oaimarc	http://www.openarchives.org/OAI/1.1/oai_marc.xsd
11%	MPEG21-DIDL	http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd
8%	RDF	http://www.openarchives.org/OAI/2.0/rdf.xsd
6%	Uketd_dc	http://naca.central.cranfield.ac.uk/ethos-oai/2.0/uketd_dc.xsd

6%	Junii2	http://ju.nii.ac.jp/oai/junii2.xsd
5%	Context_ob	http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:ctx
4,5%	Oai_etdms	http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd
4,5%	Xepicur	http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd
4%	junii	http://ju.nii.ac.jp/oai/junii.xsd
3%		http://epubs.cclrc.ac.uk/xsd/qdc.xsd
3%		http://www.proprint-service.de/xml/schemes/v1/PROPRINT_METADATA_SET.xsd
3%	xmetadiss	http://www.d-nb.de/standards/xmetadiss/xmetadiss.xsd

Table 1 – Diffusion of the most used metadata sets

4.2 Archive Metrics

At the level of a single archive, a number of issues has been addressed, namely: the mentioned schema, the distribution of metadata sets, the number of records, the presence of empty records, etc.

Reference Schema for the Metadata set. As to the schema definition, a clear problem has been identified for Dublin Core. In particular, the most common reference to DC schema is:

URI: http://www.openarchives.org/OAI/2.0/oai_dc.xsd

While, we found other four different aliases in the xml namespace definition that have problems:

- URI:http://www.openarchives.org/OAI/2.0/oai_dc/oai_dc.xsd (does not work)
- URI:http://www.openarchives.org/OAI/2.0/dc.xsd (does not work)
- URI: http://openarchives.org/OAI/2.0/oai_dc.xsd (work, may be not official or identical to the main)
- URI:http://www.openarchives.org/OAI/2.0/oai_dc/ (does not work)

Another example is for Xmetadiss metadata set schema. The common reference to Xmetadiss should be:

URI:http://www.d-nb.de/standards/xmetadiss/xmetadiss.xsd,

while some OAs use an alias which does not work:

http://www.ddb.de/standards/xmetadiss/xmetadiss.xsd

This scenario shows that an extended quality check is needed not only for the value of metadata field, but also for the schema reference to certify the metadata quality.

Use of Metadata set. As discussed in Section 4.1, the distribution of metadata set is quite spread. Moreover, there exist a 15% of institutions using metadata sets which are personal model (single instances in the distribution) or which do not have a significant number of institutions. The adoption of non-standard metadata set and schema affects the effectiveness of archive visibility and distribution. Here below some examples:

URI schema: http://libst1.nul.nagoya-u.ac.jp/akf/akf.xsd

URIschema: http://uhasselt.be/agris/1.0.xsd 1

URI schema: http://ubib.eur.nl/eur_qdc/1.0.xsd

Small OA Archives. Among the identified OA from the general list, a number of them are quite small. For example, 14% of the archives have less than 100 records. In some cases, they are experimental projects instead of well established institutional archives. The threshold at 100 records may be set to another value, while for this preliminary assessment we have considered smaller archives, namely those with less than 100 records.

Empty Metadata Records. In some cases, the OA presents some empty records. This problem is more common than expected. In particular, from the performed analysis we have identified that more than the 2% of archives have 80% empty records, the 84% of the archives have no empty records, while the 9% have less than the 5% of empty records. We can identify in 5% a threshold beyond which the quality of the archive falls over.

4.2 Metadata Record Metrics

At the level of the single metadata record some problems have been detected regarding the interpretation of the single metadata fields. In this case, a greater attention has been focussed on DC since it is the most common metadata model (as shown in the past). Moreover, it is well know that the use of simple Dublin Core [6] foresees a high level of flexibility for filling in the metadata field. The performed analysis has shown that a very few number of institutions did adopt a qualified DC model, as defined by standard recommended best practice with a controlled vocabulary such as RFC 4646[14] or ISO639-1 [15]. Moreover, the metadata multi-language system is managed in two modalities: using different instances of DC: language for each language or expressing different languages in the same field with a separator. The analysis has outlined that this separator can be arbitrary the sequent types: ‘ ,’ ‘ ;’ ‘ – ‘ ‘/’. Table 2 gives an overview of the different instances for coding languages as found in the harvested metadata , please note that some of them are even not correct.

language	instances	tot
English	en, eng, English, en_GB, en-GB, Englisch	6
Spanisch	es, spa, Espanol, Spanish; spa; , sp	6
French	fr,fre, French, French;, Francais, fra	6
Deutsch	ger ,de, German, Deutsch, ge	5
Greek	gr, gre, grc, ell	4
Italian	it,ita, Italian	3
Japan	jpn, ja, jp	3

Table 2 – Diffusion of most used metadata sets

DC:format. The DC:format field can be filled in with file format, or with the physical medium, or with the dimensions of the resource as described by standard definition [6]. The recommended best practice refers to using a controlled vocabulary such as the list of Internet Mime Types [13], [16]. According to the common

standards of IANA, each mimetype [13] can be coded with a limited number of instances, on the contrary, the collected mimetype contained even wrong coding in more than 10% of cases (depending on the specific type coded).

4. Conclusions

The adoption of Open Access model by the implementation of freely accessible institutional archive is growing more and more, particularly in scientific domains. On the other hand, the quality and trustability of the implementation and services provided still remains a requirement not well addressed and hard to assess formally and automatically.

The OA implementation landscape is very fragmented. Up to now, each institution tends to adopt its own policies for workflow, publication, access, descriptive metadata adoption, type and format of contents, distribution, etc. The scenario of the OA works by the scientific community is based on the assumption of interoperability; that is de facto hard to be massively viable. During the experiments, we realized that it would be difficult for an aggregator or OAI-PMH client to trust archives and get access to high quality records in every case. This paper has shown some of the most evident difficulties to maintain the archive services available on internet, so as to be able to answer to OAI-PMH requests. Moreover, the institutions present a certain delay in adopting and accepting standards. The number of different metadata schemas used to describe and classify OA resources, suggests that an international and community oriented policy for repository and related resource evaluation is missing. This can cause a reduction of credibility for the open access model as a viable channel to disseminate scientific knowledge, thus impacting on scientific community.

Presently, we are harvesting from repositories all the available records without limitation and by exploiting the GRID infrastructure. At the end of this process, a more exhaustive and detailed evaluation will be provided. If you are interested in inserting your OAI-PMH service in the analysis, please send an email to the authors of the paper at DSI/DISIT University of Florence. In this perspective, this paper's authors can perform on demand a direct specific assessment of your OAI-PMH service site and provide a corresponding record with suggestions to improve your service.

References

[1] C. Lagoze and H. V. de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html

[2] Herbert Van de Sompel et al. Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine December 2004 Volume 10 Number 12 ISSN 1082-9873

[3] AXMEDIS: Automating Production of Cross Media Content for Multichannel Distribution http://www.axmedis.org

[4] MPEG-21, Information Technology, Multimedia Framework, "Part 2: Digital Item Declaration," ISO/IEC 21000-2:2003, March 2003.

[5] METS, http://www.loc.gov/standards/mets/

[6] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/

[7] MARC ISO 2709 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41319

[8] DSpace http://www.dspace.org/

[9] Fedora http://www.fedora.info/

[10] EPrints for Digital Repositories http://www.eprints.org/

[11] Emanuele Bellini and Paolo Nesi - A Trust P2P network for the Access to Open Archive resources - IFLA2009 Milan

[12] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

[13] Freed, N. and N. Borenstein. 1996. "RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies," November 1996. http://www.ietf.org/rfc/rfc2045.txt?number=2045

[14] RFC4646- http://www.ietf.org/rfc/rfc4646.txt

[15] ISO639-1 http://www.infoterm.info/standardization/iso_639_1_2002.php.

[16] MIME-IANA http://www.iana.org/assignments/media-types/

[17] P. Bellini, I. Bruno, P. Nesi, "Visual Programming of Content Processing Grid", The 15th International Conference on Distributed Multimedia Systems, DMS2009, Hotel Sofitel, Redwood City, San Francisco Bay, USA, September 10 to September 12, 2009;

[18] P. Bellini, I. Bruno, P. Nesi, M. Spighi, "Taxonomy and review of Complex Content Models", in Proceedings of the 14th International Conference on Distributed Multimedia Systems (DMS'2008), Boston, USA, 4-6 September 2008, pp.71-76.

[19] Xiaoming Liu et al. « Arc - An OAI Service Provider for Digital Library Federation" D-Lib Magazine April 2001 Volume 7 Number 4 ISSN 1082-9873

[20] Pleiadi www.openarchives.it/pleiadi/

[21] Citeseer http://citeseer.ist.psu.edu/

[22] U. Muller, A. Powell, P. Cliff, H. Van de Sompel, C. Lagoze, Nelson M. and Warner S - OAI for Beginners – the OpneArchive Forum online tutorial http://www.oaforum.org/tutorial/english/intro.htm

[23] W3C- RDF/XML Syntax Specification http://www.w3.org/TR/REC-rdf-syntax/,

[24] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, . 2002. *Implementing Guidelines for the Open Archives Initiative for Metadata Harvesting: Guidelines for Harvesting Implementes* http://www.openarchives.org/OAI/2.0/guidelines-harvester.2002-06-10.htm

[25] OCLC – OAIHarvester, http://www.oclc.org/research/software/oai/harvester.htm

Advanced Pattern Match techniques for identifying Similar Transactions in a Database System

Thangaraj Karol Stuart
tkstuart@umd.umich.edu

Dr. Zhiwei (Tommy) Xu
zwxu@umich.edu

University of Michigan Dearborn, Dearborn, MI

Abstract

In a database system, it is difficult to identify and gather complete information about long running or frequently running transactions. When more number of long concurrent Transactions run in a production system, there is a greater chance that it can cause more blocking locks / deadlocks, causing 100% CPU, 100% Disk I/O and the system will be literally unusable. Currently, there is no way to extract all SQL statements in a transaction or identify similar SQL statements of a Transaction, running in a database systems. This paper provides a method to extract all SQL statements in a Database Transaction, including the time taken to execute the SQL statement and the SQL explain plan. Three String Pattern Matching Algorithms (1) Levenshtein Distance (Edit Distance) Algorithm (2) Jaro Winkler Algorithm and (3) Boyer-Moore-Horspool Algorithm are used to identify unique transactions. These three algorithms are compared and the faster String Pattern Matching algorithm is identified.

Keywords: Blocking Locks, Concurrency, Database Performance, PatternMatch, Concurrent Transactions

1. Introduction

The complexity and size of database systems are steadily growing. At the same time, reliability is becoming a much more vital concern for economy as database system pervades our societies and lives, playing important roles in banking, automobile manufacturing, running financial markets and operating defense systems. The cost of slow performance produces heavy loss to a company. When a performance issue arises, it is difficult for the developers to identify the problematic transactions.

In a high concurrency system, the technique to extract SQL Transactions was never performed. The Pattern Matching technique to identify the Similar SQL Patterns in a Transaction was never attempted.

To analyze the complexity of a database system, the SQL transactions should be studied and analyzed. This will help the Developers to concentrate on improving the performance of problematic Transactions like

(1) Long running Transactions

(2) Frequently running Transactions

(3) Transactions causing Blocking Locks/Deadlocks

By knowing these information, a Product Data Management Company can know what Engineering process, users are using the most, a retail company can know what products the customers are buying more, a banking company will know what features like Pay Bills, View Bill, etc, are used by users the most.

A different variety of String Pattern mining capabilities has been developed over the years. They cover a broad spectrum of topics including: (i) Fast String Pattern Matching [3, 4, 5, and 6]; (ii) Levenshtein Distance Algorithm (iii) Boyer –Moore Algorithm (iv) Jaro Winkler Algorithm (v) Wu-Manber Algorithm, etc...

An increasing amount of work is published on mining software “development” data. [1] [2]. A popular area of applications for mining data is the market basket analysis, which studies the buying behaviors of customers by searching for sets of items that are frequently purchased either together or in sequence. To mine customer behavior data, code needs to be modified with efficient logging to identify the different modules in a system, so that the customer behavior can be identified. Oracle Database system provides a tool called OEM [10] (Oracle Enterprise Manager), which provides information about all SQL statements happening in the system. But it does not include the details of Transactions.

Research has not been done so far to mine SQL statements in a Database Transaction and to dynamically use the runtime data to improve the performance of the system or to get statistical information about the usage of certain modules. Runtime Usage data is rich in information and gives the type of Transactions being used in a system. This paper discusses about mining database SQL data to provide better application transaction understanding.

Three different algorithms are used to compare the string pattern matching in this paper. They are (1) Edit Distance Algorithm (2) Jaro Winkler Algorithm and (3) BoyerMooreHorspool Algorithm

2. Data Extraction & Transaction Patterns

2.1 Preprocessing – Extraction of SQL Transactions

Oracle database application is selected for getting the transaction data. There is no direct/easy way to extract the Transaction data from the Oracle database. A cronjob is written to extract the Transaction data.

When a Transaction is identified, the trace for that Transaction is enabled by the Cronjob. The Trace will create a file for each Transaction. When that Transaction is completed, the Cronjob will disable Trace for that Transaction.

So, one file will be created for each Transaction.

In the Preprocessing stage, these trace files are read one by one and one line of data for each Transaction Trace file will be generated. The data extraction from Oracle will be sent to the three algorithms with the following format.

p_1|s_1, t_1, e_1| s_2, t_2, e_2| s_3, t_3, e_3| s_4, t_4, e_4| s_5, t_5, e_5| s_6, t_6, e_6| s_7, t_7, e_7| s_8, t_8, e_8| s_9, t_9, e_9| s_{10}, t_{10}, e_{10}|
p_2| s_3, t_3, e_3| s_4, t_4, e_4| s_5, t_5, e_5| s_6, t_6, e_6| s_7, t_7, e_7| s_8, t_8, e_8| s_9, t_9, e_9| s_{10}, t_{10}, e_{10}|
P_3| s_3, t_3, e_3| s_4, t_4, e_4| s_5, t_5, e_5| s_5, t_5, e_5| s_5, t_5, e_5| s_5, t_5, e_5| s_5, t_5, e_5| s_6, t_6, e_6| s_7, t_7, e_7| s_8, t_8, e_8| s_9, t_9, e_9| s_{10}, t_{10}, e_{10}|
P_4| s_3, t_3, e_3| s_4, t_4, e_4|
…
…

Where:

- p_1, p_2, p_3, p_4 are Pattern names.
- s_1, s_2, s_3… are SQL statements
- t_1, t_2, t_3… are time taken for execution of SQL statements $s_1.s_2,s_3$…
- e_1,e_2,e_3,… are SQL explain plan of SQL statements $s_1.s_2,s_3$…

2.2 Method to identify Similar Transactions

To formally define transactional pattern mining, Let X is the Transaction Set in a database System and S = $\{s_1, s_2, s_3, ...s_n\}$ be a set of statements in a Transaction X, t= $\{t_1, t_2, t_3, ...t_n\}$ be the corresponding time taken for the statements. And e = $\{e_1, e_2, e_3, ...e_n\}$ be the explainplan for these sql statements (e.g.)

where s_1, s_2, s_3 are individual SQL statements in the transaction.
(e.g.)

s_1 = select LXNAME from LXBusinessObject table for update.

s_2 = select LXSName from LXState where statename =:BINDVAR1.

s_3 = select LXFName from LXFile where filename = :BINDVAR1

t_1 = timing for execution of s_1

t_2 = timing for execution of s_2

t_3 = timing for execution of s_3

e_1 = plan for execution of s_1

e_2 = timing for execution of s_2

e_3 = timing for execution of s_3

Let P = $\{p_1, p_2, p_3, ...p_n\}$ be a set of patterns in a Transaction X, p_i is an ordered set of statements where p_i can be like $\{s_1, s_2\}$ or $\{s_1, s_2, s_3\}$

(e.g.)

p_1 = release parts in the Engineering change process. (p_1 € S)

p_2 = release drawings in the Engineering change process (p_2 € S)

p_3 = release drawings in the Engineering change Request (ECR) (p_3 € S)

p_4 = release Engineering Change Order (ECO) (p_4 € S)

For the Given: Transaction Set X,

Find all the qualified Similar Transactions (sequence of statements)

If the given input is as follows:

p_1\|s_1, t_1, e_1\| s_2, t_2, e_2\| s_3, t_3, e_3\| s_4, t_4, e_4\| s_5, t_5, e_5\| s_6, t_6, e_6\| s_7, t_7, e_7\| s_8, t_8, e_8\| s_9, t_9, e_9\| s_{10}, t_{10}, e_{10}\|
p_2\| s_3, t_3, e_3\| s_4, t_4, e_4\| s_5, t_5, e_5\| s_6, t_6, e_6\| s_7, t_7, e_7\| s_8, t_8, e_8\| s_9, t_9, e_9\| s_{10}, t_{10}, e_{10}\|

Say, s_1, s_2, s_3....s_{10} be tokens and $s_1|s_2$, $s_2|$ s_3, $s_3|$ s_4,...be SequenceTokens

(1) Find if $p_1 = p_2$, if not, go to Step (2)
(2) Find number of exact tokens in p_1 and p_2 . In the above given example p_1 has 10 tokens and p_2 has 8 tokens
(3) Find total number of sequential tokens (i.e.) $s_1|s_2$= SeqToken1, $s_2|$ s_3= SeqToken2,... $s_3|$ s_4= SeqToken3 in p_1 and p_2. In the above example, p_1 has 9 SequenceTokens and p_2 has 7 SequenceTokens.
(4) Find the similarity of tokens using the three Algorithms. (i.e.)
    - Using Edit Distance, find the number of insertions or deletions of characters, which can make both the Patterns match.
    - Using Jaro Winkler, find the Jaro Winkler value and
    - Using the Boyer Moore Horspool, find the mismatched characters in the given token.
(5) Find the similarity of SequenceTokens using the three Algorithms. (i.e.)
    - Using Edit Distance, find the number of insertions or deletions of characters, which can make both the Patterns match.
    - Using Jaro Winkler, find the Jaro Winkler value and
    - Using the Boyer Moore Horspool, find the mismatched characters in the given token.
(6) If the average value of the number of matching regular and sequence tokens is greater than 0.7 value, then we can assume both patterns match. This 0.7 value is determined from the Training Data.

3 Case Studies and Experiment Results

3.1 Problem Description

To identify similar Transactions, Transactional data is required. There are no direct methods to capture the Transactions in a SQL statement, without affecting the System Performance.

Tracing features are available for capturing all SQL statements occurring in a Database. But enabling this Tracing feature will capture all SQL statements and also will hinder the overall Performance of the System. Also capturing all the SQL statements in a system will not help in identifying the high concurrent transactions, long running transactions, etc…

Even, if the Transactions and its SQL statement are captured, there is no direct algorithm that can be used to identify similar transactions.

3.2 Case Studies

A PLSQL Program is written to extract the Transactional SQL statements including the SQLID, TimeTaken and Explain Plan.

The data extracted from the database system will be like

PatternName|No.OfSQLStatements |"SQL1HashValue","SQL1-TimeTaken","SQL1-Plan"| "SQL2HashValue","SQL2-TimeTaken","SQL2-Plan"| "SQL3HashValue","SQL3-TimeTaken","SQL3Plan"| …..

Table 1 Transaction Data Input

From this data, only the SQL Hashvalue will be extracted to feed as an input to the three algorithms.

PatternName| SQL1HashValue| SQL2HashValue| SQL3HashValue|….

Table 2 Processed Transaction Data Input

Similar transactions are identified using the method given in Section 3.2 using the three algorithms defined in Section 2.1, 2.2 and 2.3

3.2.1 Case Study - Similar Transactions with multiple two pairs of SQL statements.

Let's take two transactions T_1 and T_2 . Let's say the SQL queries of T_1 are $s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8$ and SQL queries of T_2 are

$s_1, s_2, s_3, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5,$

$s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_6, s_7, s_8$.

The SQL statements in T_1 can be like

4275138295|1272033463|654733891|xxxxxxxxxx|yyyyyyyyy| 745409377|1688851807|2635386423

and SQL statements in T_2 will be like

4275138295|1272033463|654733891|
xxxxxxxxxx|yyyyyyyyyy|xxxxxxxxxx|yyyyyyyyyy|
xxxxxxxxxx|yyyyyyyyyy|xxxxxxxxxx|yyyyyyyyyy|
xxxxxxxxxx|yyyyyyyyyy|xxxxxxxxxx|yyyyyyyyyy|
xxxxxxxxxx|yyyyyyyyyy|xxxxxxxxxx|yyyyyyyyyy|
xxxxxxxxxx|yyyyyyyyyy|745409377|1688851807|
2635386423

The second Transaction T_2 has executed more than two pairs of additional SQL statements xxxxxxxxxx|yyyyyyyyyy| when compared with T_1 .

For testing this scenario, 90 different Transactions were captured. Out of which, 12 transactions had incorrect data and was removed. The remaining 78 data were used for testing. This data is split into three sets for testing. The output file of the Transaction Trace, which is the input file for the three algorithms is generated and is of the format given in Table 1and will be processed to give output as in Table 2.
(i.e.) the SQLtimetaken and SQLhashvalue is removed before being processed by the Algorithms. However they will be needed for performance analysis.

The three algorithms used to identify the similar transactions are implemented using MATLAB.

As shown in Figure 2, on Comparison of the given SQL statements, Jaro Winkler algorithm worked faster than EditDistance or BMH.

The accuracy of the Pattern Classification was same for all three algorithms.

3.2.2 Case Study Conclusion:

In this Case study, we have taken Transactions with multiple SQL statements of same type. This case study can be treated as user performing a search and then clicking Page 1, Page 2, etc… to view its contents or Human Resources Manager searching for salary for each employee and clicking each employee link to see his/her salary details..

In this Case, we have taken two SQL statements which get repeated multiple times.

Pattern1|00000000000|999999999||00000000000|999999999||00000000000|999999999||00000000000|999999999|2208092950|3705610144|745409377|1688851807|2635386423

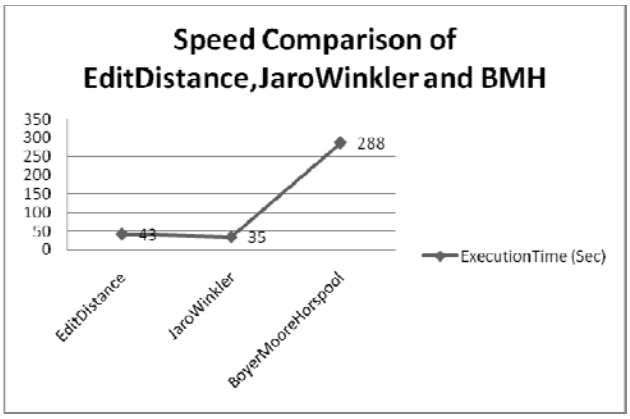


Figure 2 Speed Comparison of EditDistance, JaroWinkler and BMH

Pattern2|00000000000|999999999||00000000000|999999999||00000000000|999999999||00000000000|999999999|2208092950|3705610144|745409377|1688851807|2635386423
Pattern3|2208092950|3705610144|745409377|1688851807|2635386423|00000000000|999999999|00000000000|999999999|00000000000|999999999|00000000000|999999999
Pattern4|00000000000|999999999|00000000000|999999999|00000000000|999999999|00000000000|999999999|00000000000|999999999|00000000000|999999999|2208092950|3705610144|745409377|1688851807|2635386423
Pattern5|00000000000|999999999|00000000000|999999999|00000000000|999999999|00000000000|999999999|2208092950|3705610144|745409377|1688851807|2635386423
……
Here 00000000 and 999999999 are repeated most.

Say 00000000 = SELECT EMPID,EMPNAME FROM EMP WHERE EMPNAME = :BINDVAR

And Say 999999999 = SELECT SAL FROM SALARY WHERE EMPID = :BINDVAR

Then this shows that a HR Representative is running a Transaction to pull the salary of employees.

When these kind of similar Transactions are performed and extracted, then Jaro Winkler Algorithm is classifying the data at a faster rate. All three algorithms classified the data correctly.

BoyerMooreHorspool algorithm scans the characters of the pattern from right to left beginning with the rightmost one. In case of a mismatch (or a complete match of the whole pattern) it uses two pre-computed functions to shift the window to the right. It has to compare a set of character and move bit by bit from left to right. Two while loops are created to compare each character in s_i with each

character in s_j, where i and j are loop iteration of SQL1 and SQL2. Since while loops are used, this process is costly. But, If the comparisons of strings are lengthy, then BoyerMooreHorspool Algorithm might work even more efficiently, as it skips a set of characters in case of mismatch. But in our example, there are no lengthy SQL tokens.

Edit Distance Algorithm uses a matrix array. For given two strings, a and b:

```
if a == b, distance = 0
 m = length of a and n = length of b
 if m == 0, distance = n
 if n == 0, distance = m
```

Edit Distance constructs a matrix w, containing $m+1 * n+1$ elements. It computes for each letter in m and each letter in n. But the calculation involved is simpler than BoyerMooreHorspool Algorithm.(i.e.)

w[r,s] = Min (left diagonal + substitution cost, above + delete cost, left + insert cost)

Levenshtein (Edit) distance is calculated as the value in last row, last column in the matrix array.

JaroWinkler Algorithm doesn't use complex loops. It uses the Jaro formula and Winkler Enhancement formula to compute the distance. So, it is faster than Levenshtein and BoyerMooreHorspool.

So, for transaction comparison with SQL statements like $s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8$ and $s_1, s_2, s_3, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_4, s_5, s_6, s_7, s_8$, the Jaro Winkler is better than BoyerMooreHorspool Algorithm and Levenshtein Distance (Edit Distance) Algorithm in terms of execution speed.

4. Conclusion

As explored in this paper, String Pattern Matching of SQL transactions provide a way to discover useful hidden /long running transaction patterns and frequently running transaction patterns in a transaction database. To count the Transaction occurrence in the database, each transaction event should be captured. This procedure is not so costly when compared to capturing all SQL statements in a schema. Once captured, already existing three algorithms (Edit Distance, Jaro Winkler and BoyerMooreHorspool algorithm) are used to identify similar transactions. Experimental studies to evaluate the performance of algorithms using SQL String Tokens are performed. The Jaro Winkler algorithm classified the data at a faster rate. As similar transactions are identified in this paper, the problem of automatic healing of Blocking locks and deadlocks can be performed as future enhancement.

References

[1] R. Agrawal and R. Srikant, Mining Sequential Patterns:Generalizations and Performance Improvements. In Proc. of the 5th Int. Conf. on Extending Database Technology (EDBT), 1996.

[2] T. Ying, Predicting Source Code Changes by Mining Revision History. M.Sc. Thesis, Dept. of Computer Science, University of British Columbia, 2003.

[3]. Aho, A.V . , and Corasick, M.J. Fast pattern matching: An aid to bibliographic search. Comm. ACM 18, 6 (June, 1975), 333-340.

[4]. Beeler, M., Gosper, R.W., and Schroeppel, R. Hakmem. Memo No. 239, M.I.T. Artificial Intelligence L a b.,M.I.T., Cambridge, Mass., Feb. 29, 1972.

[5]. Dewey, G. Relativ Frequency of English Speech Sounds. Harvard U. Press, Cambridge, Mass., 1923, p. 185.

[6]. Knuth, D . E . , Morris, J . H . , and Pratt, V.R. Fast pat tern matching in strings. TR CS-74-440, StanfordU., Stanford, Calif., 1974.

[7] M.-S. Chen, J. Han, and P. S.Yu. Data Mining: An Overview from Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6):866—883, December 1996.

[8] J. Han and J. Pei. Mining Frequent Patterns by Pattern-Growth: Methodology and Implications. ACM SIGKDD Explorations (Special Issue on Scaleble Data Mining Algorithms), December 2000.

[9] H. Mannila and D. Rusakov. Decomposition of event sequences into independent components. Proc. of the First SIAM Conference on Data Mining, 2001.

[10] Bert Rich, Oracle Enterprise Manager - Getting Started with Oracle Change Management Pack, March 2002

[11] Stella Budrean Yanhong Li Bipin C. Desai , High Availability Solutions for Transactional Database Systems 2003

[12] John F. Elder IV & Dean W. Abbott A Comparison of Leading Data Mining Tools August 28, 1998

[13] Jia-Yu Dai, Don-Lin Yang, Jungpin Wu, and Ming-Chuan Hung An Efficient Data Mining Approach on Compressed Transactions 2008

A Display Adaptation Model for Nomadic Users

Luis Carlos Niño-Tavera
Computer Science Department
Pontificia Universidad Javeriana
Bogotá, Colombia
l-nino@javeriana.edu.co

Angela Carrillo-Ramos
Computer Science Department
Pontificia Universidad Javeriana
Bogotá, Colombia
angela.carrillo@javeriana.edu.co

Edgar Enrique Ruiz
Computer Science Department
Pontificia Universidad Javeriana
Bogotá, Colombia
eruiz@javeriana.edu.co

***Abstract*— This paper describes an application intended for a hospital which validates and presents the benefits of establishing *NOMAD* as a model for display adaptation. *NOMAD is* a model oriented to nomadic final users that improves adaptation of information on any device that can access several different Information Systems (*IS*). *NOMAD* includes Ubiquitous Computing characteristics that allow the interaction process between a user and his/her environment to be a natural and intuitive process. *NOMAD* is based on personalization, taking into account a user's preferences with regard to the display of information and the restrictions imposed by the physical characteristics of his/her *HAD* (acronym of Heterogeneous Access Device, for example, laptop, desktop, PDA and cellular phone).**

Keywords- Display Adaptation, Ubiquitous Computing, Device Features, Personalization and Session

I. INTRODUCTION

Currently, the ability of a nomadic user to access different Information Systems (*IS*) is evident. After receiving a great amount of information and assuming it is adapted according to user preferences, the next step is to identify methods that could be used to display this information to the user. Due to the progressive decentralization of computing an increase in what is known today as mobile computing, has been generated. According to the *W3C* [18], mobile/ubiquitous computing is a new personal computer paradigm, which allows a nomadic user to obtain required information regardless of where, when or the device by which a user is connected. In concordance, it is worth mentioning that a nomadic user can build a query with great amounts of information that is beyond the display capacity of the user's *HAD* (acronym of ***H***eterogeneous ***A***ccess ***D***evice, such as desktop, laptop, *PDA* and cellular phones) [1]. During the last decade there has been an increase in the use of embedded devices such as the *PDA* [20] with the objective of providing access to different *IS*. Such *IS* contain a lot of information, including different kinds of multimedia data (*e.g.*, images, audio and video) which may not be displayed on the user's access device due to its characteristics and physical restrictions [20]. This way, it is clear there is a need to take into account the access device (mobile or not) through which a user executes a query and receives information.

Searching in different *IS*, a user might feel overwhelmed by the amounts of information he/she can retrieve, as it may come from multiple systems such as the Internet or a simple organization database. The outcome of such a search can generate great amounts of information; in some cases, the information can be unnecessary according to the user's needs and situation [9]. From the need to improve the quality of search results, the need to consider context is born, with the objective of adapting information [13].

The objective of display adaptation is to offer information in a way the user wants it (especially if he/she is a nomadic user), considering three main aspects: *i)* the user profile, *ii)* the device profile through which the user invokes services and receives information, *iii)* localization and the moment in which a query is executed. A *user profile* has a set of characteristics such as tastes, interests, interactions and activities of a specific user, with the purpose of representing him/her in the system. According to Bouzeghoub *et al.* [5], a profile can be defined as a personalized model considered in an information access process, which allows it to be received and displayed according to a user's needs, his/her interests, history in the system and preferences. According to Cannataro *et al.*, [6], besides the different kinds of users, there are diverse types of access devices with different physical (*e.g.*, hardware) and logical (*e.g.*, software) aspects. This difference between *HAD*, indicates the need of considering their capabilities and characteristics with the purpose of adjusting the presentation of the information in each device. Belotti *et al.*, [4] considers that it is necessary that user preferences are taken into account, since they can affect directly the interactions between the user and the system.

In order to consider adaptation features with the objective of improving the display of information on different kinds of *HAD*, we have created *NOMAD*, a display model which considers aspects such as context, a user profile and the characteristics of the *HAD* in order to adapt the presentation of information on a user's *HAD*.

The structure of this paper is the following: Section II presents related works. Section III defines the *NOMAD* architecture. Section IV exemplifies, through a practical application, the validity of the model and the benefits of establishing *NOMAD* as a model to display adaptation. Finally, Section V enumerates conclusions and future work.

II. RELATED WORKS

The *NOMAD* architecture adds value to information display by considering adaptation features. Table I presents a comparison of related works with regard to the user profile representation. The used annotation for all the tables in this paper is: '+' contemplates the aspect, '-' does not contemplate the aspect and '?' indicates there is not enough information for this aspect.

TABLE I. USER PROFILE.

	[5]	[7]	[13]	[15]
Basic Data	+	+	+	?
Performance preferences	-	+	-	-
Time-Space	-	+	+	+
History	-	?	-	-

In Table I, it can be viewed that most of the mentioned works consider the basic information of the user; with exception of [7], [13] and [15]. None of them consider performance preferences or time-space features at the same time. It is important to clarify that not one of the mentioned works considers the user's history to enrich his/her profile.

TABLE II. HETEROGENEOUS ACCESS DEVICES

		[17]	[19]	[18]	[11]	[2]
Hardware	**Memory**	+	-	-	-	-
	Battery	-	-	-	-	-
	Screen	-	+	+	+	+
	Proc. Speed	+	-	+	-	-
	Network	+	-	+	-	?
Software	**OS**	+	+	+	+	-
	Browser	-	+	+	+	+
	Application	-	-	?	-	-

Table II presents a comparison of related works with regard to the *HAD* features. It can be observed that it becomes necessary to consider the characteristics of the *HAD* so the best adaptation of display can be achieved, considering the *HAD's* restrictions. For example, an aspect that may change performance in a considerable way are the available applications in the *HAD*, because the display of certain information may depend on the adequate execution and availability of a program or an application.

TABLE III. CONTEXT

	[3]	[8]	[16]	[13]	[10]	[15]
Location	+	+	+	+	+	+
Time	?	-	?	+	?	+
Sensors	+	+	+	?	-	-

From Table III and the aspects taken into account in the context, we can see that time is not considered in most of the mentioned works, being this an important aspect to keep in mind when receiving information. An example to highlight the importance of this aspect is to consider the waiting time for a clinical history, from the point of view of a physician, or a fireman in the case of a fire alarm message because in these two contexts delivery is different. In the case of the physician, the elapsed time to download the clinical history is not considered vital. In the second case, the delay of the message could be considered serious.

III. *NOMAD*

This section describes the components of *NOMAD*, a model which personalizes the display of information on the user's *HAD*. The *NOMAD* architecture is composed of seven layers, represented in the Figure 1. A detailed description of *NOMAD* can be found in [14].

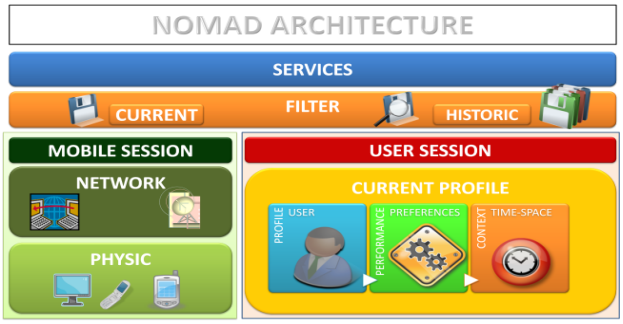


Figure 1. *NOMAD* Architecture.

The first layer corresponds to the *Physical Layer*. In this layer, *NOMAD* proposes to generate a *CC/PP* [19] extension, considering both physical and logical aspects of the *HAD*. Certain characteristics are included like the capacity to execute or not certain applications (*e.g.*, JVM, video viewer), the amount of applications that are being executed, the file formats that *HAD* can display (*e.g.*, text, image, video) and the features of that format (*e.g.*, size, file type).

The second layer (*Network Layer*) determines the capacity of the device from the network connection's point of view, considering features such as connection type (wired or wireless), connection protocol (*e.g.*, IrDa, Bluetooth, Wi-Fi), communication protocol, bandwidth, traffic type and transmission media. This layer is responsible for detecting changes that may occur in a connection, and of informing to the *Mobile Session Layer* any change that has occurred in the *Physical Layer*.

The third layer corresponds to the *Current Profile Layer*. In this layer, a user profile is defined, composed basically of three types of preferences: display, performance and time-space context. These preferences allow the encapsulation of a user's needs aiming at displaying on his/her *HAD,* the information as he/she wants. Each kind of preferences is exemplified in [14]. Communication between this layer and the *User Session Layer* is done through a protocol where the *Current Profile Layer* sends information about the user's preferences. When it receives information, the *User Session Layer* can understand it and then process it considering the user profiles registered in a history.

The fourth layer is the *Filtering Layer*. In this layer, there are two important aspects to consider: the first one corresponds to the current session (generated by the service provided by the *Current Profile Layer*) and the second one concerns a history where the previous sessions are saved.

The objective of the *Mobile Session Layer* (fifth layer) is to capture information provided by the *Physical Layer* and the *Network Layer* respectively, joining the different characteristics and restrictions of the device and network, all of this in real time.

The *User Session Layer's* (sixth layer) main objective is to capture user display preferences taking into account the characteristics of the current session (*e.g.*, user, formats, response times, activities), considering its relevance and the

characteristics and preferences from the historical sessions, which would avoid doing several iterations of user queries.

The *Services Layer* (seventh layer) has two general objectives: *i)* the first one, is to capture information from the entire current session in order to obtain both characteristics of the device and the user, all of them from the services provided by the 5th and 6th layers respectively. *ii)* The second objective is to capture the needs/requirements demanded by the requested service from the application executed on the *HAD*, dynamically reviewing the characteristics of the device and the user profile provided by *NOMAD*, in order to personalize the display of information obtained as a result of the requested service.

IV. An Application for a Hospital

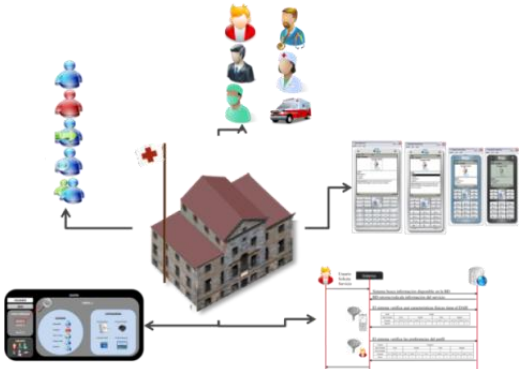


Figure 2. Practical Application Environment: Hospital.

We have developed an application related to a physician who at his work site depends on a direct way on his *HAD* for the development of his daily activities. We have used Java Micro Edition Technology (*JME*) in order to develop such an application for *PDA* and cellular phones. We tested the application by means of the emulation environment included on the Sun Java Wireless Toolkit (*WTK*). All concepts related to the application are represented in *Figure* 2. *NOMAD* defines six groups of users: *i)* A default user; *ii)* the administrative personal; *iii)* the group of surgeons; *iv)* the doctors; *v)* the nurses group; *vi)* the ambulance drivers.

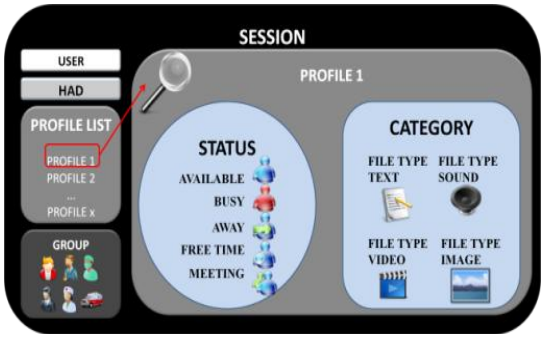


Figure 3. User Session.

A user must define his/her status so the other users know a little bit more about his/her session besides general or personal information. Status can be: Available, Busy, Away, Free time and Meeting. The user session consists of one or more profiles. During the entire session, there can be moments where the session loses its profile attribute. However, in order to get access to the *IS*, the user must select a profile. *Figure* 3 shows the basic components of a session on the system: *i)* the user's characteristics such as name, age and gender, *ii)* the physical *HAD* features associated with the current session user, *iii)* the group assigned to the user *iv)* the selected profile.

We can define different *HAD* features maintaining the same user preferences; in this way, the filter at the *Physical Layer* is contemplated, and the display of the same information is personalized according to the *HAD*. As can be appreciated in *Figure* 4, depending on the features of the *HAD* the display varies, on the left *HAD*, the display corresponds to a video that represents the course the ambulance driver must follow; however, on the right side of the image, this *HAD* does not support video format, because of that the displayed information is in text format.



Figure 4. Display Example with different *HAD* features.

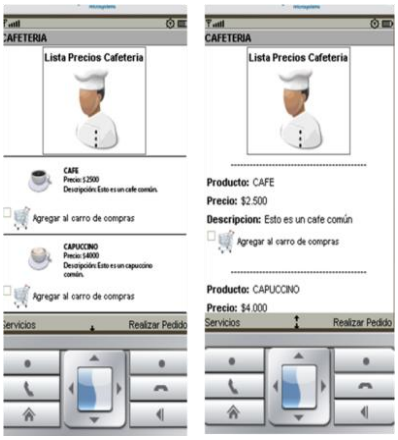


Figure 5. Information Display: same *HAD*, different information about display preferences.

In *Figure* 5, two *HAD* with the same physical features are showed, however the queries of the different *IS* associated to the system, provide information of a cafeteria service through images or text. Once the *HAD* features have been defined, the filtering system verifies the user's preferences, the user on the left in *Figure* 5 prefers images while the user on the right prefers text. It is worth mentioning that the features showed could be from the same user that defines two profiles with different preferences and the access to information will be

affected depending on the chosen profile for the current session.

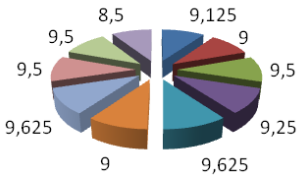


Figure 6. Survey Results.

We applied a survey to a users group in order to evaluate aspects such as: *Navigability of the application* (9.125), *Menu Coherence* (9), *Help screens* (9.5), *Undo Action* (9.25), *Consistent Error Message* (9.625), *Data Distribution on the Screen* (9), *Data Coherence* (9.625), *Information Display* (9.5), *Number of screenshots to find a service or the requested information* (9.5), *Response Time* (8.5). This survey was applied after executing tests with help from the users. This survey showed a high acceptation level of the application (see *Figure* 6). The total average (over 10) was 9.2625.

TABLE IV. CLINICAL HISTORY TEST RESULTS

Num. of Users	*HAD* constraints	User Pref. Order	Data Coherence	Response Time	Data Lost
3	None	I, V, T	Totally Coherent	7-10 s.	None
5	No Sound	I, S, T	Totally Coherent	8-10 s.	None
5	Only Text	T	Totally Coherent	0-2 s	None
5	None	V	Totally Coherent	8-10 s	None
8	Text	I, T	Totally Coherent	3-6 s	None
8	None	V,I	Totally Coherent	5-8 s	None
8	None	I, V, T, S.	Totally Coherent	9-12 s	None

We executed different tests for the application. The tests applied to the clinical history service are presented in Table IV. The conventions are: I: Image, S: Sound, T: Text, V: Video. All the tests were successful with regard to user's needs and preferences. However, some response times where affected because of latency and other problems concerning the web server.

V. CONCLUSIONS AND FUTURE WORK

It is expected that the *NOMAD* architecture will be adaptable to any architecture that requests services related to the display of information, as a portable and robust component, useful for any kind of application or system. However, it is intended that the components of the architecture contemplate a wide amount of aspects related to the characteristics of the newer generations of *HAD*. It is also important to work on modeling session characteristics that allow the personalization of display in a particular session.

ACKNOWLEDGMENT

The authors want to thank Janet Arciniegas and Andrea Barraza for their comments on our work.

REFERENCES

[1] T. Agoston, T. Ueda and Y. Nishimura, "Pervasive Computing in a Networked World", INET 2000, 10th Internet Society Conference. http://www.isoc.org/inet2000/cdproceedings/3a/3a_1.htm#s1. (2000)

[2] I. Amendola, F. Cena, L. Console, A. Crevola, C. Gena, A. Goy, S. Modeo, M. Perrero, I. Torre and A. Toso, "UbiquiTO: A Multi-device Adaptive Guide", in Mobile HCI 2004, LNCS, vol. 3160, pp. 409-414, Springer, 2004.

[3] J. Bae, J. Lee, B. Kim and S. Ryu, "Next Generation Mobile Service Environment and Evolution of Context Aware Services", in EUC 2006, LNCS, vol. 3160, pp. 459-468, Springer, 2006.

[4] R. Belotti, C. Decurtins, M. Grossniklaus, M. Norrie and A. Palinginis, "Interplay of Content and Context", in ICWE 2004, LNCS, vol. 3140, pp. 187-200, Springer, 2004.

[5] M. Bouzeghoub and D. Kostadinov, "Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils", in CORIA 2005, pp. 201-218, 2005.

[6] M. Cannataro, A. Cuzzocrea and A. Pugliese, "A Multidimensional Approach for Modelling and Supporting Adaptive Hypermedia Systems", in EC-Web 2001, LNCS, vol. 2115, pp. 132-141, Springer, 2001.

[7] A. Carrillo-Ramos, J. Gensel, M. Villanova and H. Martin, "PUMAS: a Framework based on Ubiquitous Agents for Accessing Web Information Systems through Mobile Devices", in SAC2005, 20th Symposium on Applied Computing, pp. 1003-1008. ACM, 2005.

[8] D. Chalmers, M. Sloman and N. Dulay, "Map adaptation for users of mobile systems", in WWW 10, 10th Int. Conference on WWW, pp. 735 – 744, ACM.

[9] B. Corrie, A. Patrick, J. Singer, S. Noël, K. El-Khatib, B. Emond, T. Zimmerman and S. Marsh, "A QoE Sensitive Architecture for Advanced Collaborative Environments", in 3rd Workshop on Advanced Collaborative Environments, pp. 319-322, ACM.

[10] F. Gandon and N. Sadeh, "A Semantic E-Wallet to Reconcile Privacy and Context Awareness", in ISWC2003D, LNCS, vol. 2870, pp. 385-401, Springer, 2003.

[11] J. Indulska, R. Robinson, A. Rakotonirainy and K. Henricksen, "Experiences in Using CC/PP in Context-Aware Systems", in MDM 2003, LNCS, vol. 2574, pp. 247-261, Springer, 2003.

[12] J. Kim, U. Kang, and J. Lee, "Content-Based Filtering for Music Recommendation Based on Ubiquitous Computing", in Intelligent Information Processing III, Int. Conf. on Intelligent Information, LNCS, vol. 228, pp. 463-472, Springer.

[13] T. Lech and L. Wienhofen, "AmbieAgents: A Scalable Infrastructure for Mobile and Context-Aware Information Services", in AAMAS 2005, 4th Int. Conf. on Autonomous Agent and MAS, pp. 625-631, ACM, 2005.

[14] L. C. Niño-Tavera, A. Carrillo-Ramos, E. Ruiz, "Display Adaptation in *NOMAD*ic Environments", in MoMIR2009, 3rd International Workshop on Mobile Multimedia Information Retrieval, pp. 548-552. ACM, 2009.

[15] C. Panayiotou and G. Samaras, "Mobile User Personalization with Dynamic Profiles: Time and Activity", in OTM Workshops and Posters, LNCS, vol. 4278, pp. 1295-1304, Springer.

[16] M. Park, J. Hong and S. Cho, "Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices", in UIC 2007, LNCS, vol. 4611, pp. 1130-1139, Springer, 2007.

[17] C. Velasco, Y. Mohamad, A. Gilman, N. Viorres, E. Vlachogiannis, A. Arnellos and J. Darzentas, "Universal access to information services-the need for user information and its relationship to device profiles", in Universal Access in the Information Society. LNCS, vol. 3, no. 1, pp. 88-95, Springer.

[18] W3 development: User Agent Profile: http://w3development.de/rdf/uaprof_repository/

[19] Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0. http://www.w3.org/TR/2007/WD-CCPP-struct-vocab2-20070430/.

[20] Y. Zhang, S. Zhang and S. Han, "A Pervasive Service Framework for Pervasive Computing Applications", in UIC 2007, 4th Int. Conf. Ubiquitous Intelligence and Computing. LNCS, vol. 4611, pp. 1024-1033, Springer, 2007.

A Hybrid Approach for Adapting Web Graphical User Interfaces to Multiple Devices using Information Retrieved from Context

Carlos E. Cirilo[§], Antonio F. do Prado[§], Wanderley L. de Souza[§], Luciana A. M. Zaina[*]
Federal University of São Carlos (UFSCar), [§]São Carlos/[*]Sorocaba, SP, Brazil
{carlos_cirilo, prado, desouza}@dc.ufscar.br, lzaina@ufscar.br

Abstract

Ubiquitous Computing has enabled users the easy access to services and applications from anywhere, at anytime, and using any device. In this scenario, software development presents a number of challenges, among them the need for adapting applications to different access devices. Each device has a set of specific capabilities, which difficult to build and maintain specialized versions of the application to run in each type of device (static adaptation). Despite the heterogeneity, these applications should operate in a compatible manner to the access device profile, preserving its characteristics so that user interaction is not compromised. From information taken out of context it is possible to adapt applications at execution time according to characteristics of the environment in which they operate (dynamic adaptation). In this sense, this paper presents an approach that combines static and dynamic code generation for adapting Web graphical user interfaces using information retrieved from context. A case study was developed to investigate the feasibility of the proposal.

1. Introduction

One of the purposes of Ubiquitous Computing is to expand human activities with new services and applications that can adapt to circumstances in which they are used. Among the challenges that arise from this scope is the need of applications that offer support for different devices in distinct contexts. A critical issue is how to develop this kind of application. Each device has a specific profile, which comprises its particular features like processing power, graphics capabilities, screen size, support for a certain markup language, among others. Besides, new platforms and new devices are frequently released. Due to the wide range of available devices and the discrepancies among their profiles, it is difficult to build and maintain specific versions of the application for each device and every platform [1], [2].

Web applications are examples of programs which can be accessed from different types of devices. Today, for example, it is possible to check emails, read news, shop, view the weather, browse content, etc., using the Web browser of a cell phone or smartphone. With the advent of Web 2.0, Web applications have become more interactive, providing rich interfaces that resemble desktop applications, allowing users a more meaningful interaction with the application [3]. However, adapting these applications to the heterogeneity of devices becomes a task even more complex, due to the need to preserve their peculiar interaction characteristics that allow users to get a richer experience with the application.

Thus, it is necessary that applications are aware of available resources to adjust themselves appropriately. Therefore, it is necessary to these applications to be sensitive to the context in which they operate. *Context-Sensitive Systems (CSS)* are able to automatically adapt their behavior and content in order to fit the applications features to the requirements of the context by using information extracted from it, such as the device profile, the user preferences, the characteristics of the access network, the user location, and so on [4]. In this light, context sensitivity is extremely useful, for instance, to allow a Web application adapts its graphical user interface according to the device profile retrieved from the interaction context.

Two strategies of code generation can be employed to carry out the interface adaptation: *at development time (static adaptation)*, which means that a specific interface version is developed separately for each target device; and *at execution time (dynamic adaptation)*, where the interface code is automatically generated from abstract descriptions when user accesses the application [2], [5]. This paper presents a hybrid approach for adapting rich interfaces of Web 2.0 applications. The approach combines static adaptation, by building a few generic versions of the interface for some device groups, with dynamic adaptation, through the adaptation of snippets of the code already built using context information at runtime.

The remainder of this paper is organized as follows: Section 2 introduces the main concepts and techniques related to developed work; Section 3 presents the proposed approach showing the UbiCon framework built for allowing hybrid interface adaptation, as well as its reuse in an application as case study; Section 4 discusses some related works; and Section 5 presents concluding remarks and further work.

2. Concepts and Techniques

The main concepts related to the proposed approach include Web 2.0, context-sensitive systems, and interface adaptation strategies.

2.1 Web 2.0

The term Web 2.0 has been adopted to refer to a new wave of Web applications designed specially to support collaboration and the sharing of user-generated content [6]. Besides harnessing collective intelligence, Web 2.0 encompasses a number of other principles [3]. Among them, stands out the use of rich interfaces for affording users a more meaningful experience with the applications. The so-called *Rich Internet Applications (RIAs)* have adopted technologies that enable the creation of more attractive user interfaces, providing

the sensitivity, features and functionalities that resemble desktop applications. Features such as asynchronous communication with the server using *Assynchronous JavaScript and XML (AJAX)*, drag-and-drop components, sliding panels, components for capture and display videos, maps, online spreadsheets and text editors, are examples of rich interface components which enable greater interactivity and improve the overall users' experience [7].

2.2 Context-Sensitive Systems (CSS)

Context sensitivity is related to the adaptation of application according to its location of use, the nearby people or objects, as well as the changes occurring in the environment over time. A *Context-Sensitive System (CSS)* is able to adapt its operations without explicit intervention of the users, providing information and services that are relevant for users perform their tasks using information taken out of the interaction context [4], [9].

Context guides the variations in CSS behavior, enriching the user interaction either by influencing recommendations or by enabling adaptations of any kind [10]. For example, among the actions performed by a CSS one can mention the content adaptation, in which an original content is converted into a series of new formats compatible with the delivery context. This adaptation may be, for instance, a transformation, where the markup language of a Web interface is modified to fit more appropriately to the device profile; and/or a transcoding, where images are converted or resized [11].

Many definitions for context have been proposed to make it an operational concept in computational terms. A recent definition suggests a distinction between *context* and *contextual element*. This definition states that a *contextual element (CE)* is any piece of data or information that characterizes an entity in a domain; and the context of an interaction between an agent (human or software) and an application, with *focus* on a task, is the set of instantiated CEs that are necessary to support the task to be performed [10].

2.3 Interface Adaptation Strategies

One of the critical aspects in the development of ubiquitous applications is the premise that they should be able to execute and adapt themselves to the heterogeneity of user's computing devices and to the environment in which he is immersed. In general, two strategies of code generation for interface adaptation can be employed [2]:

- *Generation at development time (static adaptation)*: Different versions of the interface, each one with appropriate contents to a particular device, are built during the development process (i.e., *before* the execution of the application). Thus, it is possible to drive implementations that take maximum advantage of specific features of a given device. However, this approach raises several problems, such as: high costs and longer development time (due to the need to build a specialized version of the interface for each type of device); high maintenance costs (once a slight change in interface requirements demands isolated changes in several versions); endless development process (since as new devices are released other versions should be built); among others [8].
- *Generation at execution time (dynamic adaptation)*: The interface code is generated during the execution of the application. Generally, this approach is used for adaptation of Web applications based on request/response. The adaptation process can be a transformation between two markup languages (e.g. XHTML into WML), or a code generation from definitions in abstract interface models. During the adaptation process, CEs such as the devices capabilities, the user profile, etc., can be used in order to adapt the final presentation of the interface. Thus, the development process is simplified because there is no need to build and manage numerous interface versions. However, once the interface definition should be generic enough to attend the features of a variety of devices, this strategy restricts access to specific functionalities of devices [2]. Furthermore, since all interface code must be generated at execution time, the code processing time can impact on application performance.

3. Hybrid Approach for Adapting Web Graphical User Interfaces

To overcome the shortcomings of each interface adaptation strategy, the approach combines static and dynamic code generation for adapting rich interfaces using context. In this case, a few generic versions of the interfaces are built, each one suitable for a certain group of devices (static adaptation). When accessing the application, the version that best fits to the type of access device is selected; the code snippets that need to be refined, such as images, size of interface components, etc., are dynamically adapted according to the device profile retrieved from the interaction context (dynamic adaptation); and finally the interface with modified content is delivered to the user's device.

This way, there is a reduction in the complexity of development, since the number of versions is reduced. Moreover, as only code snippets are adapted instead of the entire code, the impact on application performance is lower. Finally, since the versions are further refined during the dynamic adaptation part to fit to the device profile retrieved from context, there is greater use of available resources on the device and a better interface adaptation.

To support the approach, it was developed a framework called *Ubiquitous Context Framework (UbiCon)* which provides services to manipulate context and adapt user interfaces in a hybrid fashion according to the device profile retrieved from the interaction context. Following UbiCon framework is presented.

3.1 Ubiquitous Context Framework (UbiCon)

UbiCon is a framework that aims at facilitating the development of Web-based CSS. UbiCon abstracts the functionalities related to context manipulation and provides services for adapting user interfaces to different devices that accesses applications. UbiCon was implemented in Java. Figure 1 shows the component model of the UbiCon, which was structured in four modules: *Acquisiton*, *Processing*, *Dissemination*, and *Content Adaptation*, based on the *Context Architecture* proposed in [10].

3.1.1 Acquisition Module

This module encompasses the components that access directly the context sources (e.g. stored profiles, databases, camera and sensor drivers) for acquiring the raw CEs, i.e., in the format delivered by such sources. These components are called *adapters* [10], because they encapsulate the details of access to context sources, connecting the framework to several context sources and providing generic

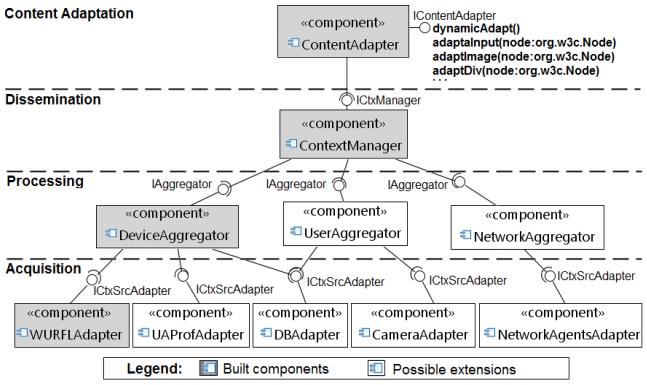


Figure 1. UbiCon component model.

methods for recovery the CEs through the *ICtxSrcAdapter* interface.

The context sources are, by nature, heterogeneous, autonomous and dynamic, due to the fact that they exist independently of the CSS [10]. Thus, in UbiCon there is an appropriate adapter component for each context source. The adapters use specific drivers and *Application Programming Interfaces (APIs)* of the corresponding context source for access its internal features in order to retrieve the CEs available. This granularity makes UbiCon more extensible, since it facilitates the inclusion of new context sources or the removal of those that no longer exist or have become deprecated.

In Figure 1, the adapter component named *WURFLAdapter* accesses the XML database called Wireless *Universal Resource File (WURFL)* [12], which stores the profiles of thousands[1] of devices from different brands and models. Being a public database, WURFL receives updates on a daily basis by developers spread around the world interested in contributing to the completeness and correctness of the information contained therein. Therefore, WURFL was adopted as main context source in UbiCon to acquire the devices profiles.

Figure 2 shows a code snippet of *WURFLAdapter*. For retrieving device profiles stored into WURFL database it is used the WURFL Java API. Context capture is based on *User-Agent* field of the *Hypertext Transfer Protocol (HTTP)* request originated by the current user's access device. Through this contextual information, it is possible to retrieve the other contextual elements related to device profile stored into WURFL.

Due to the extensible nature by which UbiCon was designed, other context sources such as *User Agent Profile (UAProf)* [13] and profile databases can be added to the framework to complement the acquisition of CEs associated to the device profile, which implies the implementation of specific adapters for access these context sources, as illustrated in Figure 1. In addition, context sources that provide CEs related to the *user* (e.g. databases that stores the user profiles, video cameras that capture the user current location), the *access network* (e.g. software agents that monitor the network traffic), among others, can also be incorporated into UbiCon to make it more comprehensive.

3.1.2 Processing Module

This module includes the *aggregators* [9] components that process and group the CEs according to the domain entities that they characterize. The processing involves the transformation, if necessary, of raw CEs to a format or a level of abstraction most

[1] In the last update of WURFL available in July, 26th, 2010, around 13,452 device profiles were accounted.

```
package br.ufscar.dc.ubicon.ctxsources.adapters;
import net.sourceforge.wurfl.core.DefaultDeviceProvider;

public clas  WURFLAdapter implements ICtxSrcAdapter {
 private String wurflPath = "wurfl.zip";
 private String patchPath = "web_browsers_patch.xml";
 private WURFLManager manager;
 private Device deviceProfile;
 ...
 public WURFLAdapter (HttpServletRequest request) {
   ...
   deviceProfile = manager.getDeviceForRequest(
                              request.getHeader("User-Agent"));
 } //end of constructor
 ...
 @Override
 public String getContextualElement(ContextualElement ce) {
  switch(ce) {
   case DEVICE_DISPLAY_COLUMNS_NUMBER:
      return deviceProfile.getCapability("columns");
   case DEVICE_DISPLAY_RESOLUTION_WIDTH:
      return deviceProfile.getCapability("resolution_width");
   case DEVICE_PRODUCT_INFO_IS_MOBILE_DEVICE:
      return deviceProfile.getCapability("is_wireless_device");
   ...
  }
 } // end of getContextualElement method
 ...
} //end of WURFLAdapter class
```

Figure 2. Code snippet of *WURFLAdapter*.

suitable for use in the systems. Furthermore, possible conflicts among several values for the same CE obtained from different context sources also is solved in this module.

Each aggregator hides the complexity inherent in handling the various adapters of Acquisition Module and provides a single interface, called *IAggregator*, for obtaining the CEs of a given domain entity. For example, in Figure 1 the component *DeviceAggregator* represents the access device and interacts directly with all adapters that access the context sources which provide the CEs related to the device profile.

As depicted in Figure 1, besides *DeviceAggregator*, aggregators that represent other possible domain entities, such as the user, the access network, etc., can be built to extend the UbiCon as new context sources which provide CEs associated to these entities are added to the framework.

3.1.3 Dissemination Module

This module is composed of a single component called *ContextManager* that provides the CEs manipulated in Processing Module to interested context consumers through the *ICtxManager* interface. The task for determining from which aggregator to request a given CE is encapsulated by *ContextManager*. Figure 3 shows a code snippet of *ContextManager*. The *getContextualElement* method selects the most appropriate aggregator for obtaining the requested CE, which is identified through the item of enumeration called *ContextualElement* passed as parameter to that method. In this case, if the label of the CE starts with "DEVICE_", then *DeviceAggregator* is selected.

```
package br.ufscar.dc.ubicon.manager;
import br.ufscar.dc.ctx.aggregators.managers.DeviceAggregator;

public class ContextManager implements ICtxManager {
  private DeviceAggregator deviceAggregatorManager;

  public ContextManager(HttpServletRequest request) {
     if (deviceAggregatorManager == null)
         deviceAggregatorManager = new DeviceAggregator(request);
  }
  @Override
  public String getContextualElement(ContextualElement ce) {
     if (ce.toString().startsWith("DEVICE_")) {
         return deviceAggregatorManager.getContextualElement(ce);
     }
     return null;
  }
  ...
}//end of class ContextManager
```

Figure 3. Code snippet of *ContextManager*.

3.1.4 Content Adaptation Module

This module is responsible for user interface adaptation according to access device, user and network profiles retrieved from the interaction context. As depicted in Figure 1, it consists of a component called *ContentAdapter*, whose content adaptation services are provided through *IContentAdapter* interface. The *dynamicAdapt* method performs the reading of the requested X/HTML document and invokes appropriate methods for adapting the various interface content such as images, input text fields, tables, among others.

To guide user interface adaptation, *ContentAdapter* component consumes the CEs associated to the adopted profiles, which are obtained through the services provided by *ICtxManager* interface of Dissemination Module. To implement *ContentAdapter* it was used the *Document Object Model (DOM)* API [14], which allows the manipulation and transformation of XHTML and HTML documents. DOM API allows for reading and writing these documents into memory by structuring them in a DOM objects tree (e.g. *Node*, *Document*). Thus, it is possible analyze and modify the Web pages dynamically as needed. In order to work properly *ContentAdapter* requires well-formed X/HTML files as input. However, since many documents available on the Web do not satisfy this requirement, it is used the TagSoup API [15] to correct missing and mismatching tags before processing them using DOM API.

Figure 4, for instance, shows the code of *adaptInput* method which adapts text input fields. This method receives as parameter a DOM object of Node type that represents the <*input*> node of the document to be adapted, which was pre-processed through TagSoup API to make it a well-formed node. During adaptation it is retrieved the columns number of the device screen via *ICtxManager* interface. Then the "size" attribute value of the input node is obtained. Finally, such attribute value is reduced if it is greater than the columns number of the access device screen.

Once built, UbiCon is reused in the development of applications, as shown in Figure 5. The content adaptation is done on a hybrid manner. Thus, when a client accesses an application by sending an HTTP request (1), the application *Servlet* invokes the *dynamicAdapt* method of UbiCon (2). UbiCon retrieves the access device profile from

```
package br.ufscar.dc.ubicon.contentadaptation;
import br.ufscar.dc.ctx.manager.ContextManager;
public class ContentAdapter implements IContentAdapter{
  private ContextManager cm;
  private HttpServletRequest request;
  private HttpServletResponse response;
  private String interfaceID;
  ...
  private void adaptInput( Node inputNode) {
   int displayColumnsNumber = Integer.parseInt(getContextManager()
           .getContextualElement(
                   ContextualElement.DEVICE_DISPLAY_COLUMNS_NUMBER));
   Node inputType = inputNode.getAttributes().getNamedItem("type");
   if (inputType.getTextContent().equals("text")
         || inputType.getTextContent().equals("password")) {
     String inputSize = inputNode.getAttributes().getNamedItem("size")
             .getTextContent();
     if (inputSize != null && inputSize.length() > 0) {
         if (Integer.parseInt(inputSize) > displayColumnsNumber) {
             replaceAttValue(inputNode, "size", Integer
                     .toString(displayColumnsNumber));
         }
     }
   }
  } //end of adaptInput method
  ...
} //end of ContentAdapter class
```

Figure 4. Code of *adaptInput* method.

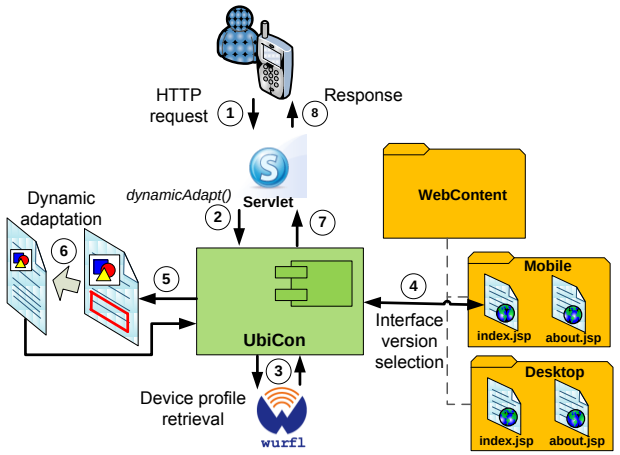


Figure 5. Reuse of UbiCon.

WURFL (3) and selects the requested Web page from the most appropriate interface version according to the device profile obtained (4). Then a copy of the chosen page is created into memory, its ill-written excerpts are corrected, and the code snippets of interface content that need to be refined to meet device profile are identified (5). The necessary adjustments are applied (6), the copy of the adapted page is written in the output stream of HTTP response (7) and finally sent to the user's device through the application *Servlet* (8).

To determine the most appropriate interface version, UbiCon reads a XML file, called *context-rules.xml*, which contains the definitions of rules that determine the choice of a suitable generic version to current access device. This file is prepared by the application developer, who decides, at development time, the best combination of access device characteristics (contextual elements) that determines the choice of each interface version. As shown in *context-rules.xml* file illustrated in Figure 6, the rules for choosing an interface version are expressed in *contextRule nodes*, which specify the CEs and their matching values that determine the choice of the interface version related to *context node* in which they are contained. For example, in Figure 6 the interface version for desktops is associated with the *context node* whose *interfaceID* attribute contains the value "Desktop". This node indicates that such version will be selected only if the contextual element called DEVICE_PRODUCT_INFO_IS_MOBILE_DEVICE retrieved from interaction context matches the value "false".

It is also possible to observe in *context-rules.xml* file in Figure 6 the *dynamicAdaptation* attribute of *context node*, which indicates whether a certain interface version should be dynamically adapted by UbiCon after its choice. The presence

contextSpecification	
webApp	ASPS
context	
interfaceID	Desktop
dynamicAdaptation	true
contextRule	
contextualElement	DEVICE_PRODUCT_INFO_IS_MOBILE_DEVICE
matchOperator	eq
matchValue	false
CERelevance	3
context	
interfaceID	Mobile
dynamicAdaptation	true
contextRule	
contextualElement	DEVICE_PRODUCT_INFO_IS_MOBILE_DEVICE
matchOperator	eq
matchValue	true
CERelevance	3

Figure 6. Reuse of UbiCon.

of this attribute is justified by the possibility of building up highly specialized interface versions for specific devices and therefore does not require additional refinements.

3.2 Case Study

In order to investigate the feasibility of the proposal, an application in the Emergency Healthcare domain was developed. The UbiCon framework has been used to develop the Web module of the *Ambulance Space Positioning System (ASPS)* [16] so that it could be accessed from a desktop or smartphone. ASPS emerged from an experimental study which aimed at investigating the use of the signals from *Global System for Mobile Communication (GSM)* antennas for the location of people or objects, focusing on the location of emergency service vehicles like ambulances. With this functionality the ASPS allows the fleet management team to monitor the mobility of the ambulances and to direct a given ambulance to an emergency care in a location near the region in which the ambulance currently is.

Due to its ubiquitous nature, ASPS is composed of three parts. The first one, which runs on the GSM terminal installed in the ambulance, performs the calculation of the position through the triangulation of the nearest cell phone antennas and transmits the calculated position to a server via HTTP. The second one, which runs on the server, processes the data transmitted by the GSM terminal and stores them in a spatial database. The last one is the Web module that includes the ASPS administrative page. This module has a graphical interface to visualize the position of ambulances on the city map. In short intervals, asynchronous requests are sent to the server by using AJAX for retrieving and updating the most recent position of ambulances on the map.

ASPS Web module was developed using *Java Server Faces (JSF)* framework. Figure 7 shows the code of ASPS application *Servlet* called *ContextServlet*, where the reuse of UbiCon is accomplished. The *doGet* method, which intercepts HTTP requests, invocates the *dynamicAdapt* method of UbiCon. From there, UbiCon retrieves the device profile from the interaction context, selects the most appropriate interface version according to the recovered device profile (using the *context-rules.xml* file presented in Figure 6) and adapts the application interface to fit it to the peculiarities of the current access device.

To test the execution of the interfaces, the Web browsers of HTC G1 and iPhone smartphones, as well as the Web browser of a personal computer have been used. For test proposals, the position data have been obtained by the simulation of the

```
package br.ufscar.dc.asps.servlets;
import br.ufscar.dc.ubicon.contentadaptation.ContentAdapter;
public class ContextServlet extends HttpServlet {
  private ContentAdapter ca;

  public ContextServlet() {
      super();
  }

  protected void doGet(HttpServletRequest request,
          HttpServletResponse response)
              throws ServletException, IOException {
      if (ca == null) {
          ca = new ContentAdapter(request, response);
      } else {
          ca.reload(request, response);
      }
      ca.dynamicAdapt();
  }
}
```

Figure 7. ASPS application *Servlet*.

triangulation execution by using the GSM terminal emulator available with Android SDK. The signal intensities of the antennas for the calculation of positioning have also been obtained through simulation [16]. Figure 8 shows the results of the ASPS Web module execution in three different devices. Figure 8(a) and (b) show the ASPS administrative page viewed, respectively, on the HTC G1 smartphone screen and on the iPhone screen. In both cases, the interface version delivered was the version built for mobile devices. To point out the difference obtained with the dynamic content adaptation performed by UbiCon, the adapted page and its corresponding page without adaptation are shown. Figure 8(a) shows the top of the Web page that contains the logo and the system name. It is possible to note that, with the interface adaptation, the logo was resized to fit the HTC G1 screen. Figure 8(b) shows the bottom of the same page, which has a tabbed panel with a Google Maps mashup. It can be noted that, with the interface adaptation, the tabbed panel component has been adjusted, not to extrapolate the width of the iPhone screen. Figure 8(c), in turn, shows the interface viewed on a personal computer screen. In this case, the version of the interface delivered was the version built for desktops.

4. Related Works

Several works related to context-sensitive applications development have been proposed by the academic community, including processes (e.g. [10]), tools (e.g. [2], [5]) and frameworks (e.g. [9], [11], [17]).

Extended Internet Content Adaptation Framework (EICAF) [11] is a framework for content adaptation of Web applications. EICAF uses ontologies for describing the profiles of the devices, users, and other entities, and employs Web services for performing content adaptation by combining the contextual information expressed in profiles. Although content adaptation services provided by EICAF are similar to those of UbiCon, in the latter the device profiles are retrieved from WURFL, which is constantly updated by development community as new devices are released in market, so that content adaptation can be performed for a wider range of devices. Besides, UbiCon focuses on rich interface adaptation.

Semantic Context-aware Ubiquitous Scout (SCOUT) [17] is a framework for building context-sensitive applications for mobile devices. SCOUT allows the mapping of real world entities (e.g. people, places, objects) to virtual entities on the Web ("Web presence"), so that resources/services provided by these entities are specific of their location and are accessible when users are close to them. UbiCon differs from SCOUT by providing content adaptation services instead of location-based services. Moreover, besides content adaptation, UbiCon allows for adding other services types upon its generic *Acquisition*, *Processing* and *Dissemination* modules, making it more comprehensive.

XMobile [2] is an environment for adaptive interfaces generation of form-based mobile applications. It consists of a platform-independent user interface components framework, and a code generation tool which maps models that describe the application into executable code. Unlike UbiCon, the content adaptation in *XMobile* is done at development time (static adaptation) by constructing specific interface versions for each target platform with aid of the



Figure 8. Execution of ASPS Web application in three distinct devices.

code generation tool. Moreover, *XMobile* does not use context at runtime for performing adaptations.

Semantic Transformer [5] is a tool used for automatic transformation of Web pages designed originally for desktop platform into Web pages suitable for mobile devices. This tool acts as a Proxy that detects HTTP requests originated from mobile devices and processes the requested Web page by placing it in a proper format for viewing on the mobile device. Differently of UbiCon, which employs a hybrid approach for content adaptation, the code generation for content adaptation in *Semantic Transformer* is completely done at runtime.

5. Conclusions and Further Work

This paper presented a hybrid approach for adapting Web interfaces according to access device profile retrieved from context, by combining code generation at development time and dynamic optimization of this code at execution time. To this end, it was developed a framework called UbiCon, which provides services to context manipulation and content adaptation.

In order to investigate the applicability of the proposal, a case study in Emergency Healthcare domain was developed by reusing UbiCon. In spite of a purely static or dynamic interface adaptation, the proposed approach contributed for a reduction in the complexity of development, because there was no need to build a specific interface version for each existing device type. Instead, a few generic versions for some devices groups were developed. Moreover, UbiCon enabled the rapid development of the application, hiding the issues related to context acquisition and content adaptation, allowing to keep focus on the functionalities related to the business requirements of the developed application.

As future work, one intends to extend UbiCon by including context sources that provide contextual elements which characterize other entities besides the access device, making it more comprehensive and allowing adaptations based on other profiles, such as user and access network profiles. Another future work is the extension of UbiCon with the inclusion of modules that provide other context-sensitive services, like recommendation and location-based services, enabling a greater reuse of the UbiCon framework.

Acknowledgment

Thanks to CAPES – the Brazilian Coordination for the Improvement of Higher Level Personnel – for its financial support.

References

[1] K. Gajos, and D. S. Weld, "SUPPLE: automatically generating user interfaces", Proc. Int. Conf. Intelligent User interfaces, Jan. 2004, pp. 93-100.

[2] W. Viana, and R. M. C. Andrade, "XMobile: a MB-UID environment for semi-automatic generation of adaptive applications for mobile devices", Journal of Systems and Software, vol. 81, issue 3, Mar. 2008, pp. 382-394.

[3] T. O'Reilly, "What is Web 2.0: design patterns and business models for the next generation of software", 2005. Available: http://oreilly.com/web2/archive/what-is-web-20.html. Accessed: September, 2009.

[4] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems", Int. J. Ad Hoc and Ubiquitous Computing, Jun. 2007, pp. 263-277.

[5] F. Paternò, C. Santoro, and A. Scorcia, "Automatically adapting web sites for mobile access through logical descriptions and dynamic analysis of interaction resources", Proc. W. Conf. Advanced Visual interfaces, May 2008, pp. 260-267.

[6] M. C. Norrie, "PIM Meets Web 2.0, " Proc. Int. Conf. on Conceptual Modeling, Oct. 2008, pp. 15-25.

[7] P. J. Deitel, and H. M. Deitel. AJAX, Rich Internet Applications, and Web Development for Programmers, 1st ed., Prentice Hall PTR, 2008

[8] J. Eisenstein, J. Vanderdonckt, and A. Puerta, "Adapting to mobile contexts with user-interface modeling," Proc. IEEE Workshop on Mobile Computing Systems and Applications, Dec. 2000, pp. 83-92.

[9] A. K. Dey, "Understanding and using context", Proc. Personal Ubiquitous Computing, Jan. 2001, pp. 4-7.

[10] V. Vieira, P. Tedesco, A. C. Salgado, "A process for the design of context-sensitive systems", Proc. Int. Conf. on Computer Supported Cooperative Work in Design, Apr. 2009, pp. 143-148.

[11] M. Forte, W. L. de Souza, A. F. do Prado, "Using ontologies and Web services for content adaptation in Ubiquitous Computing, ". J. Syst. Softw., vol. 81, issue 3, Mar. 2008, pp. 368-381.

[12] WURFL. Wireless Universal Resource File. Available: http://wurfl.sourceforge.net. Accessed: February, 2010.

[13] OMA. 2001. WAG UAProf. Available: http://www.openmobilealliance.org. Accessed: February, 2010.

[14] SUN. Common DOM API. Available: http://java.sun.com. Accessed: February, 2010.

[15] TAGSOUP. TagSoup - Just Keep On Truckin. Available: http://home.ccil.org/~cowan/XML/tagsoup. Accessed: February, 2010.

[16] A. Bellini, C. E. Cirilo, et al., "A low cost positioning and visualization system using smartphones for emergency ambulance service", Proc. ICSE Workshop on Software Engineering in Health Care, May 2010, pp. 12-18.

[17] W. Woensel, S. Casteleyn, O. de Troyer, "A framework for decentralized, context-aware mobile applications using semantic web technology, " Proc. of the Confederated Int. Workshops and Posters on on the Move To Meaningful internet Systems, Nov. 2009, pp. 88-97.

Agile User Interfaces in Disaster Management

Veselina Milanova Kay-Uwe Schmidt Jörg Dörflinger Axel Spriestersbach Susan Thomas
SAP Research
Vincenz-Prienitz-Strasse 1
76131 Karlsruhe, Germany
{veselina.milanova, kay-uwe.schmidt, joerg.doerflinger, axel.spriestelbach, susan.marie.thomas}@sap.com

Abstract: *In disaster situations there are a lot of experts and decision makers involved and all of them need to share information. It is strongly desirable that everybody engaged is able to exchange applications with other experts in order to gain an overview of the current situation. That is where migratory applications would be very helpful. In this paper we present a migration tool for the disaster management domain that enables the easy exchange of applications and provides task continuity when changing devices. Decision makers can use the tool to analyze and overlay simulation data and thus elaborate response activities in a faster and more reliable manner.*

Keywords: *migratory applications, disaster management systems, agile user interfaces*

1 Introduction

Typically, there are many different end terminals available in a user's environment, but it is not yet possible to seamlessly change devices and to continue performing already begun tasks. Migratory services technology can change this situation.

Especially in the domain of disaster management such services would significantly improve the exchange of information and the rapid and reliable elaboration of response actions. When collaborating, experts from different organizations need to compare and overlay their results. The integration of these results is very important for getting an overall idea of how the whole emergency situation is eventually going to change. On the basis of the aggregated picture important response activity decisions can be taken, reducing the number of mistakes due to oversight.

In order to effectively support the decision making process in disaster situations, a migratory service that provides enough flexibility for the end user is needed. Furthermore, it needs to have reasonable response times, since they are vital for seamless and smooth migration.

Novel methods are necessary in order to meet all of these challenges. In recent years, research on migratory interfaces has started. An architectural framework developed at the Carnegie Mellon University enables device change, and looks for similar applications on the new device so the user can continue the begun tasks [1]. A programming model which supports developers in building migratory applications has also been proposed [2]. Other solutions have been described in [3], [4] but none of them fully suits the needs of the disaster management domain.

In this paper we present our general purpose migration approach to disaster management that enables the migration of application components or whole applications. The tool also builds aggregated views out of different geo simulations migrated to the same target device. Furthermore, it supports on demand the synchronization of source and target devices after completion of the migration process.

The paper is structured as follows. Section 2 describes a motivating example with the help of which we elicit requirements for a migratory application as presented in Section 3. The architecture of the tool is explained in Section 4. Section 5 gives insights into the implementation and Section 6 summarizes the results from the performed evaluation. Related work is discussed in Section 7, while Section 8 gives a summary of the main results.

2 Motivating Example

The Emergency Operations Center (EOC), situated in Portland, is responsible for the coordination of the joint civil protection forces in disaster situations. Recently, the center has been busy because of the rainy weather and the anticipated flooding of the Willamette River. The managing director of the EOC, Mr. Smith, is coordinating the response planning activities. He has contacted Professor Brown from the Center for Disaster Management in Seattle and asked him to forecast the potential impact of the rising river. Mr. Smith also needs information about the daily traffic situation in Portland. In order to get it he has contacted Mr. Thompson from the Urban Development Agency.

As soon as Professor Brown prepares his flood simulation, he wants to send it to Mr. Smith. He starts his simulation application and migrates its user interface (UI) to the PC of Mr. Smith. Mr. Smith is now able to follow the explanations of Professor Brown, because he sees all the changes the professor makes to the simulation in real time. In that moment Mr. Thompson calls to inform Mr. Smith that the traffic information is now available. Since he has to go to an important meeting, Mr. Thompson migrates not only the UI but his whole application to the PC of Mr. Smith. There the UI of the flood simulation and the traffic application merge together.

The EOC managing director now needs to closely examine the results he got from both experts. He asks Professor Brown to give him the control over the flood simulation and shows him eventual routes for the planned response activities. After the planning phase is over, Mr. Smith migrates each simulation back to the device of the respective expert.

3 Requirements

The motivating example just described incorporates the challenges mentioned in the introduction - the need for a flexible migratory service, synchronization, and fast response times. The example also helps to filter out a set of obligatory requirements for a migratory disaster management tool, some of which can be seen as general ones for migratory services.

1. There should exist the possibility to choose which parts of an application are to be migrated - if just several components are relevant to the current situation, only these should be selected and migrated. The migration of whole applications will be referred to as *total migration*, and the migration of application components as *partial migration*.

2. When more than one application is migrated to the same target device, the geographical data of these applications should be merged.

3. If an application is migrated to another device, it should be possible to keep the source and the migrated applications synchronized. In this way all changes made to the migrated application parts would be also reflected on the source device.

4. For the synchronized mode there should be a clear control policy, so that the users of both devices are aware of who is in control of the application or application component.

5. When migrating, the user should have the possibility to decide whether to quit the source application or leave it running.

6. Last, but not least, short response times for the migration service are vital for a seamless and smooth migration that does not disturb the ongoing discussions.

The fulfillment of these general requirements provides migratory tools with the needed functionality to support disaster managers and their co-workers.

4 Agile User Interfaces

On the basis of the requirements elicited from the motivating example in Section 2, we developed a migratory disaster management tool. The current section describes the agile user interfaces of the tool - its client server architecture as well as the internal architecture of the client itself.

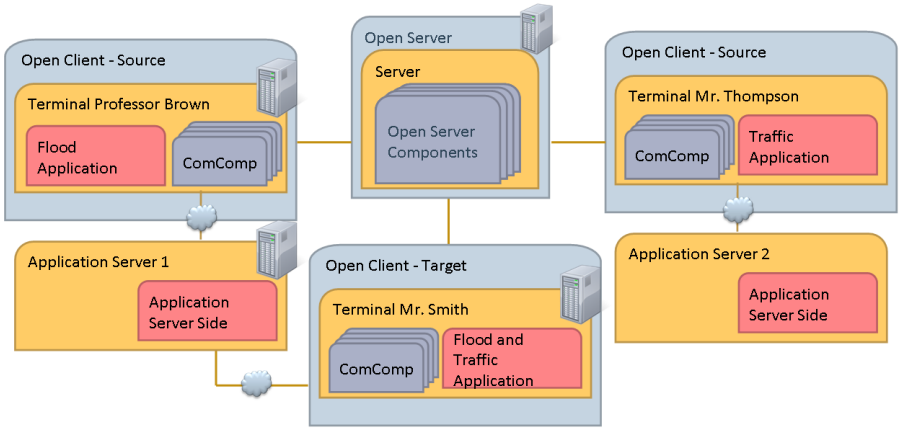


Figure 1: Client server architecture

The proposed client server architecture can be seen in detail in Figure 1. On the left of the figure the client that serves as migration source is depicted. The client is a terminal (e.g. the PC of Professor Brown) which hosts a simulation application, in this case the flood simulation, which is requested from an application server (e.g. Application Server 1). All migration requests are forwarded from the PC of Professor Brown (the source) to the target client (the PC of Mr. Smith) through the Open Server using the communication component. As soon as Mr. Smith has accepted the request, the current state of the selected flood application components or of the whole application running on the source client is sent to the target client.

When Mr. Thompson is ready with his traffic application, requested from Application Server 2, he also sends it to the PC of Mr. Smith. There the flood application is already running, so the migrated traffic components are integrated into the flood application.

Figure 2 depicts in more detail the internal architecture of the disaster management client. It is composed of application components and migration control; the application components are completely independent from the migration control, thus, any application that provides the necessary interfaces can be migrated.

For the domain of disaster management, the first application component, is a map that contains geographical imagery (streets, topographic, satellite imagery). It can receive incoming content and commands from the flood and traffic control components through the component communication bus. Due to its ability to display different layers, the content received from each control is represented by a separate layer. The map itself is stateless and does not have a dependancy relationship with other components.

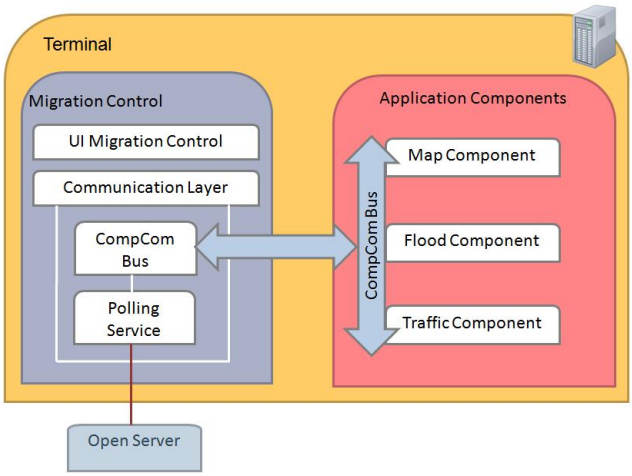


Figure 2: Client side architecture

The flood and traffic components control the flood, respectively, the traffic visualization. Both of them are dependent on the map component that displays their results visually when a content or a control command has been forwarded to it. For example, the flood simulation of Professor Brown is shown on the map and its visualization updates as the professor changes the map scale. The state management is done by the components themselves. Thus, the current state at migration time is transferred together with the corresponding component.

The migration control consists of a user interface and a communication layer. The UI enables the user to start a migration. The communication layer has a component communication bus that is responsible for the internal communication between the client application and the migration control, and a polling service that frequently and periodically polls information from the Open server. In such a way the polling service is also the communicator between server and clients that enables the synchronization of application components. All incoming and outgoing connections regarding migration are routed via the communication layer. The migration control is stateless and does not depend on any other components. It is the first component instantiated during application start and is responsible for instantiating all necessary additional components during runtime.

5 Agile UIs Implemented

As a proof of concept, we implemented the proposed architecture for a migratory disaster management tool together with a map application as a rich internet application (RIA). As a RIA, the application delivers desktop-like behaviour and interactivity. The implementation utilizes Microsoft Silverlight, because it offers a wide range of animation and multimedia capabilities as well as lean connectivity to the backend. The communication layer uses XML remote procedure calls for the internal bus as well as for the communication with the Open server. The migration tool we implemented facilitates total and partial migration as defined in Section 2. Each component can be selected for migration and the component that currently is in focus on the map is accordingly highlighted. In the migration settings one can choose to delete the to-be-migrated component from the source device after migration or to keep both devices, source and target, in synchronization. In synchronization mode there is a clear control policy. There is one synchronization master. At the beginning this is the source device, but it can be easily changed on request. The synchronization mode can be deactivated at any point of time, which is communicated to each participating side.

Figure 3 shows the application with the flood simulation activated. After migration of the flood and traffic component to one target device, both maps merge into one, inside which a smaller map in the right corner depicts an overview, as shown in Figure 4.

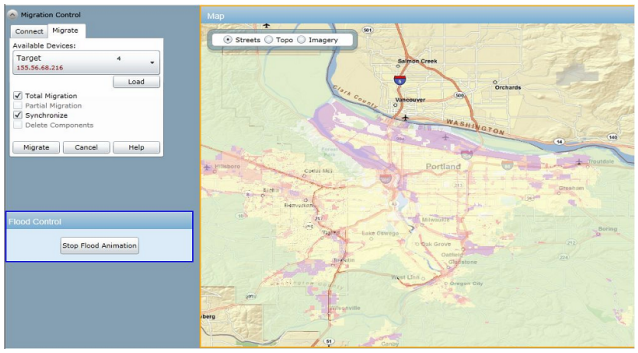


Figure 3: Application with activated flood component

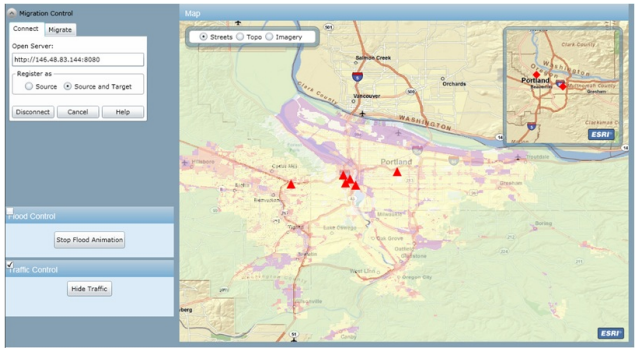


Figure 4: Merged traffic and flood components after migration

6 Agile UIs Evaluated

To determine the degree of maturity of our migration tool, we performed usability tests with 16 participants, all of whom were SAP Research employees. The tests were completed in pairs, so that a real emergency situation could be simulated, in which two participants, in the roles of a flood and a traffic expert, collaborated with each other. We elaborated a task list for the tests and used the System Usability Scale (SUS) questionnaire to determine the overall usability of the tool [5]. The SUS questionnaire was applied because it is an effective, easy to use and reliable low-cost usability

tool.

During the test sessions the screens of both participants were recorded and they were asked to think aloud and comment on their activities. Furthermore, we asked all participants to describe their understanding of total and partial migration in the given context and encouraged them to give us feedback.

The average result of the SUS questionnaire adds up to 58.44 out of 100. According to [6] most products score around 70, and a result of 58 is within the boundaries of "good" in the proposed adjective rating scale. The result makes it clear that users can work well with our tool, but there are some usability issues that still have to be repaired - e.g. a user suggested that "using the map it would be of assistance to have a zoom in / zoom out scale", a map annotation function was also proposed. One of the main problems of the tool reported in the comments was that users did not receive any feedback whether the migration was successfully executed or not. Half of all participants had a clear idea of how total and partial migration should work. All these issues give useful insights about potential future improvements to the migratory tool. Ont the positive side, 14 out of 16 participants said they would like to use the tool frequently.

7 Related Work

Research on migratory interfaces has been done in different projects. ICrafter is a service framework for a class of ubiquitous computing environments, also called interactive workspaces [3]. It aimed at allowing the users to flexibly interact with these workspaces. It generated UIs for services obtained by dynamic composition of elementary ones. ICrafter does not provide any support for migration or task continuity across different devices.

[7] presents an instant messaging infrastructure that allows developers to easily add application functionality that spans multiple personal devices. However, this approach differs from ours, since we take into consideration a wider set of devices, not limited to the user's personal ones.

Sousa and Garlan from the Carnegie Mellon University proposed an architectural framework for mobile users called Aura. It automatically reconfigures the computing infrastructure so that users can continue working on their tasks while moving to different platforms [1]. A drawback of that approach is the necessary switch of the current application to a different one that is supported by the new platform, but potentially does not provide all features of the source application.

The programming model presented in [2] supports developers in building migratory applications and has been integrated into the programming environment VisualObliq. The transmission of applications is implemented with the help of agents hopping from host to host. However, using VisualObliq, it is not possible to create migratory multi-user applications and ensure synchronization between target and source devices.

[8] describes solutions for total and partial migration, where Web applications are migrated using the transformation-based environment TERESA. The Web applications have to be developed using a model-based approach, otherwise the application cannot adapt (but can still migrate) to the new platform. However, this solution differs from our approach since it offers no merge functionality that is necessary for map applications in the disaster domain.

In general, migratory services address a complex set of issues [2]. Work done in the areas of process migration [9], virtual machine monitors [10], mobile agents [4], active networks [11], and dynamic reconfiguration in distributed systems [12] can be leveraged for migratory services. The migration idea also starts to raise industrial exploitation interest [13], even if the developed product is still in a preliminary stage in regard to migration.

8 Conclusions and Future Work

Migratory applications can be very useful in the domain of disaster management. They enable decision makers and experts to easily gain an overview of the disaster situation and share simulation results or forecasts. In such a manner, a lot of time, which is precious in disaster situations, can be saved. At the same time mistakes due to oversight can be more easily avoided.

In this paper we presented a general purpose migra-

tion tool that supports decision makers and experts, as it enables migration from one to another device, where whole applications as well as application components can be migrated. Migrated components merge with each other and can be kept synchronized with the source device.

Our evaluation results showed that the necessary time for executing a migration operation is quite short, and the response times of the migration tool itself lie under 3 seconds. The migration control tool has an average to high degree of maturity and future work will include the remediation of the usability issues that have arisen from the performed tests. Future work would also be dedicated to expanding the migration functionality to devices other than PCs, e.g. mobile or vocal devices.

The here tool proposed here provides disaster management experts with more flexibility in respect to analyzing and overlaying simulation data. In such a way, they can make better and faster decisions for possible response activities.

9 Acknowledgements

The disaster response tool presented here was designed and implemented within the VII Framework EU STREP project, Open Pervasive Environments for migratory iNteractive services (OPEN). It uses as a backend the OPEN platform [14], which is described in detail in [15].

References

[1] Joo Pedro Sousa and David Garlan. Aura: An architectural framework for user mobility in ubiquitous computing environments. *The Working IEEE/IFIP Conference on Software Architecture*, 2002.

[2] Krishna Bharat and Luca Cardelli. Migratory applications. *Proceedings of User Inteface Software and Technology (UIST 95)*, pages 133–142, 1995.

[3] Shankar R. Ponnekanti, Brian Lee, Armando Fox, Pat Hanrahan, and Terry Winograd. Icrafter : A service framework for ubiquitous computing environments. *Proceedings of UBICOMP 2001*, pages 56–75, 2001.

[4] James E. White. Mobile agents. In *Software Agents*, pages 437–472. MIT Press, Cambridge, MA, USA, 1997.

[5] John Brooke. Sus - a quick and dirty usability scale, 1986.

[6] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4:114–123, 2009.

[7] Jeffrey S. Pierce and Jeffrey Nichols. An infrastructure for extending applications user experiences across multiple personal devices. *UIST*, pages 101–110, 2008.

[8] Renata Bandelloni and Fabio Paternò. Flexible interface migration. In *Proceedings ACM IUI*, pages 148–157. ACM Press, 2004.

[9] Dejan S. Milojicic, Fred Douglis, Yves Paindaveine, Richard Wheeler, and Songnian Zhou. Process migration. *ACM Computing Surveys*, 32:241–299, 2000.

[10] Constantine P. Sapuntzakis, Ramesh Chandra, Ben Pfaff, Jim Chow, Monica S. Lam, and Mendel Rosenblum. Optimizing the migration of virtual computers. *SIGOPS*, 36:377–390, 2002.

[11] David Wetherall. Active network vision reality: Lessons from a capsule-based system. *SOSP*, 34:64–79, 1999.

[12] J. Kramer and J. Magee. Dynamic configuration for distributed systems. *IEEE Transactions on Software Engineering*, 11:424–436, 1985.

[13] Pluto, http://plutohome.com/index.php.

[14] Open website, http://giove.isti.cnr.it:88/.

[15] Open project deliverable 1.2: Initial open service platform architectural framework. http://giove.isti.cnr.it:88/documents/deliverables/d1-2-final.pdf, 2008.

Analysis of the UCD process of a web-based system

Carmelo Ardito, Paolo Buono, Maria F. Costabile,
Rosa Lanzilotti, Antonio Piccinno, Adalberto L. Simeone
Dipartimento di Informatica, Università di Bari
Via Orabona, 4, 70125 Bari, Italy
{ardito, buono, costabile, lanzilotti, piccinno, simeone}@di.uniba.it

***Abstract*— This paper analyzes how User-Centred Design (UCD) has been carried out in the creation of a web-based system, whose aim is monitoring air quality for sustainable industrial development. This distributed multimedia system has been commissioned by the Puglia region and it is used primarily by industries and regional government experts. Several lessons are learned from this analysis and hints about the effective application of UCD and the fruitful involvement of users for creating usable systems are derived.**

Keywords: System usability, User-centred design, Participatory design, User involvement, Prototyping, Software life-cycle.

I. INTRODUCTION

Usability is widely recognized by Human-Computer Interaction (HCI) experts, as well as Software Engineering (SE) experts, as a factor that affects not only the design of user interfaces but the software system development as a whole (e.g., see [10][16][18]). Users feel frustrated when they have to interact with poorly designed systems that are difficult to learn and complex to operate on, and end up by either misusing these systems or even abandoning them. This outcome is costly for the organization adopting the system and harmful to the reputation of the company that produced it [17]. This failure is the main reason why HCI research has been mostly concentrated on developing techniques and defining methodologies which can ensure that the final system satisfies users' needs and expectations.

Since the late 1970s, quality factors have been introduced in the software engineering field as part of software quality frameworks, developed to provide a measurement approach for software engineering [6], [19], [20]. Interestingly, already by that time one of these factors was usability, defined as the "effort required to learn, operate, prepare input, and interpret output of a program" [19], [20]. The fact that usability was already included as a quality factor did not really mean that software engineers paid much attention on it. Traditionally, they are trained to judge their work by criteria that may have little to do with the needs and constraints of the end users. Despite the fact that the importance of system usability and of the overall user experience has been heavily emphasized in the last couple of decades, software developers are still more sensitive to other factors, such as program efficiency and flexibility, which are certainly important engineering goals, but have little or nothing to do with creating a computer system that is usable and supportive of its users.

The main approach to creating more usable systems is user-centred design (UCD), which has been proposed as a key factor for the development of successful systems since the '80s [23]. In UCD, final users are involved from the very beginning of the planning stage, and identifying user requirements becomes a crucial phase. Many papers in the literature discuss the benefits of the user-centred approach, mainly related to the completeness of the system functionality, repair effort saving, as well as user satisfaction [1][5]. Involving users from the early stages permits the system core to be based on what is really needed. Poor or inadequate requirement specifications can later cause interaction difficulties, including lack of functionalities and usability problems. Even if late evaluations are useful to assess the usability of final systems, it is unrealistic to expect these results to bring about a complete redesign, since it would be too expensive at that time.

Many researchers acknowledge the above advantages of UCD, yet we have observed that many IT companies do not adopt this approach in their current practice of software development. There is also evidence that some techniques prescribed by UCD are not properly applied, and this prevents from achieving the expected system usability and user satisfaction results. This paper analyzes an on field experience in designing and developing a web-based system, which aims at monitoring air quality for sustainable development of industries in the territory. The design team claimed to follow a user-centred approach. Our analysis reveals that several mistakes were performed and provides useful hints about the effective application of UCD in current practice and the proper involvement of users for designing usable systems.

The organization of the paper is the following. Section II briefly reports about UCD. Section III presents the CET web-based system, whose design and development experience is described in Section IV. The lesson learned from the analysis of this experience is discussed in Section V. Finally, Section VI concludes the paper by summarizing the hints that can be derived about the effective application of UCD for creating successful systems.

II. USER-CENTRED DESIGN

Usability is acknowledged today as a fundamental factor affecting the quality of interactive systems, and designing for usability is considered critical for the success or failure of a product on the market [5], [14]. HCI issues, that affect the usability of interactive systems, are relevant within all the activities of the software lifecycle. It is not simply a matter of

adding one more specific activity to those already existing in the lifecycle, but it is the whole lifecycle that must be revised. Several models of the software lifecycle have been proposed in the software engineering community, with the aim of controlling the lifecycle and thereby producing high quality software products reliably, predictably and efficiently. On the other hand, the HCI community mostly agrees on the model of user-centred design, which refers to a design process in which end users have an influence on the design of software products.

UCD has been formalized in the ISO-standard 13407 Human-centered design processes for interactive systems [15]. At the beginning of the process, the system stakeholders are brought together to discuss and agree about how usability can contribute to the system goals and to ensure that all factors related to the use of the system are identified before the design work starts. Moreover, the methods to be used during the system development must be established. Then, the system context of use, i.e., technical, physical, and social or organizational conditions that may affect its quality, must be identified. Capturing contextual information is important to help specify user requirements, as well as providing a sound basis for later evaluation activities. After this, it is recommended to build mock-ups and prototypes of the system to support this iterative design lifecycle. Such mock-ups can be produced quickly and easily in the early stages of the system development cycle, to be followed later on by more sophisticated prototypes. This is a very important activity within the system development lifecycle; it can confirm how far the user and organizational objectives have been met, as well as provide further information for redefining the design. If the specified requirements are not met, the process is iterated and goes once again through a revision of the specifications and a proposal of a new prototype. The cycle is repeated until the particular usability objectives are met.

Various studies have demonstrated that designing for usability is economically viable since it increases sales [24], user productivity [25], while decreasing training costs and needs for user support [26]. However, none of these works seem to have succeeded in convincing companies and practitioners, since the current situation is that most IT companies still do not adopt UCD in their software development processes. The case reported in this paper refers to the developments of a system, for which it has been claimed to adopt user-centred and participatory approaches, but many mistakes have been performed, not only due to the incorrect application of some UCD techniques, but also because proper user involvement in the design process is intrinsically critical, as it will be discussed in this paper.

III. THE CET SYSTEM

The air quality is a pressing problem all around the world. Ever since the industrial era, human beings have been heavily polluting the environment and now we are aware that air quality must be carefully monitored. Emission inventories are the tools that allow governments to control the presence of emissions. Building effective emission inventories is still a challenging task but they are the basis for developing models that allow governments to monitor air quality and make good decisions about sustainable development.

The government of the Puglia region, in Southern Italy, commissioned to a consortium of research centers and local companies a system to monitor air quality for sustainable industrial development in the region; it provides information about the quantity and quality of emissions (pollutants) by processing data directly inserted in the system by employees of the industries in Puglia. The system focuses on industrial emissions in the atmosphere because they are the most relevant factor for air pollution; it actually represents an electronic cadastre of industrial emissions, thus it is called CET, being the acronym of "Catasto Emissioni Territoriali" (Territorial Emissions Cadastre). CET also supports regional government experts analysing the distribution of industrial plants and their emissions in the atmosphere.

Being a Web-based application, the system has classic client-server architecture with three tiers. An underlying database contains all data related to the monitored industries, their organizational structure, yearly production activity, plants emissions, etc. In accordance with the Italian and regional regulations, industries themselves must provide numerous complex data. To simplify users' input, data are inserted in different steps through several consecutive forms, made available by the system on the Web. CET offers the possibility of automatically compiling a series of detailed reports about the stored company data, their production processes, etc., that are very much needed by the regional government.

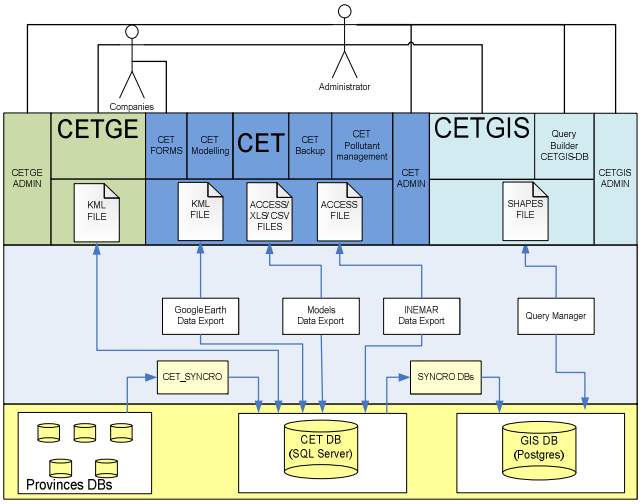


Figure 1 CET Architecture

CET includes two supporting systems: CETGIS and CETGE. CETGIS is a GIS-based system that handles spatial user queries, such as visualizing the distribution of a pollutant by town or by province; it supports arbitrarily complex views composed of various layers. Users are also offered the possibility to customize the maps of interest. CETGE has been developed to improve the system's ability to visualize the results of more complex queries on Google Earth [13]. It integrates the maps provided by Google Earth with the 3D models of elements of industries, such as chimneys, plants, etc. The CET architecture is shown in Figure 1. At the top of Figure 1 there are two main users: the employee in the company and the system administrator.

More users are involved in the system, which are not shown in this architecture. In Figure 2, some users of the CET distributed multimedia system and their main tasks are shown. As we said, CET is an electronic cadastre of industrial emissions, in which data are inserted directly by the industry employees who access the web application through a browser. They provide information about their company structure, pollution measurements, and produce their own reports. CET also supports environmental experts (chemists and regional officers) in their decision-making processes to monitor air quality; they are provided with several mechanisms to perform queries, to filter data and to visualize results. The CET server is located at the regional government where the system administrator, besides supervising the whole system, controls user registrations and manages all the industry data.

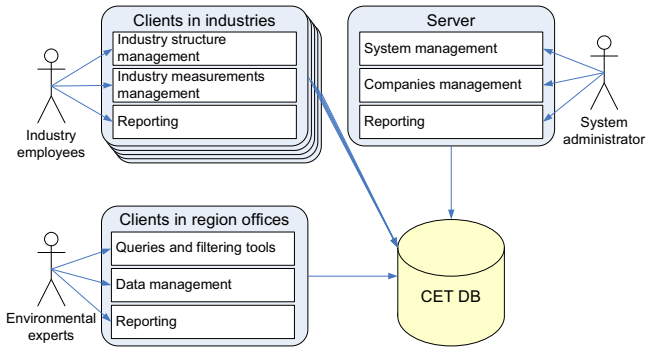


Figure 2 Users of the CET system and the main activities they perform.

In the next section, the description of the field experience in designing and developing this system is reported, focusing on the development of the specific CET component used by industry employees to input and update the required data.

IV. UCD IN CET

Data for the requirement analysis were collected from a survey of the state of the art, taking into account legacy systems and Italian laws on environment and pollution, as well as other systems having the same goals (e.g. [4], [11]). According to the user-centred approach, some contextual analysis involving domain experts and real end users was performed. Specifically, people working in industries and in the regional organization for environmental protection were visited and interviewed. It is now fully acknowledged that experts of the system domain, including end users of the system, must be involved in the design, since expertise is fundamental for the success of the final product. Thus, chemists and environmental experts were involved in a participatory design team from the very beginning [28]. Within this team, these experts explained in detail how the emissions inventories are usually built by considering aggregated data that come from multiple sources. The team analyzed advantages and limitations of other systems in order to create a very innovative product.

As we said above, a component of the system is devoted to companies that, by accessing it via Web, have to provide various data about their production processes and the type and quantity of emissions produced. Once these data have been inserted, they can create various reports for their own use as well as for fulfilling requests by the regional government. In order to involve end users as much as possible, industries in the Puglia territory were contacted and asked to provide information at different stages of the process to build the emission inventory of the Puglia region. The companies had different options for testing successive prototypes and providing useful information and feedback:

a) by filling in paper forms, which represented an initial paper prototype for the system to be designed;

b) by filling in electronic forms under the guidance of technicians from the regional office, which represented a successive running prototype of the system;

c) by filling in electronic forms that would be available on the Web, with no guidance.

These three different options aimed at involving end users right from the beginning of the development by having them work on a paper prototype first (filling in the paper forms as in option a), then on a running prototype (option b) and on a more refined, possibly final, system (option c). Seventeen companies chose option a, twelve chose option b, sixty chose option c. Based on indications provided by the domain experts, including a set of guidelines defined by the Chemistry Department of the University of Bari, which was part of the consortium developing the system, paper prototypes of the forms which would later be available on the Web were first created. The paper prototype was considered very important by the designers: it would allow users to see if the overall organization and layout were consistent with the users' mental model of organizing pollution emission data, and if all the required data, including captions and input formats, were easily understood by users.

Once the design team had agreed on a paper prototype, consisting of a set of paper forms, this was sent to the seventeen industries who chose option a, asking them to fill in the forms. Since what the designers mostly needed was feedback by end users, a Note field was inserted at the end of each form asking users to insert any comment about the form, e.g. possible problems in understanding the required data, data redundancy, missing information, etc. Moreover, users were given a telephone number to call, if they needed clarification or any help. Experts from the design team answered users' calls and, besides helping them, they took notes of all their requests, comments, complaints, etc.

Figure 3 shows the sixth one of the set of paper forms that companies employees had to fill in. This form, titled "Scheda 6 – Impianti termici", requires information about heating generators. A form must be filled in for each heating generator. In the example, the user manually inserted data about generator N. 1 and did not provide any comments in the Note field at the bottom. The other forms required several other data about the company, e.g. registry data, plant data, pollution emissions, etc.

The forms returned by the seventeen companies were carefully analyzed by the design team, which reported that special care was devoted to the comments sent by the users and to the data collected by the experts about the users' calls. On the basis of this analysis, the forms were carefully revised. An HCI expert in the design team was keen to test the revised

forms with end users, but the other team members strongly refused this proposal: they said that enough feedback had already come from the users, and the domain experts (chemists and environmental experts) were all agreed on the revised forms, which were consistent with their guidelines and mental model of how companies should operate. Moreover, time was running short and the system needed to be implemented. This is the usual concern of the software companies, which see user testing and other usability evaluation techniques only as a waste of resources. Thus, a high fidelity prototype of the system was produced, implemented in Java; it was an almost final software application since the design team was pretty sure it was the proper version complying with all defined requirements. This running prototype would be tested with the twelve companies who had chosen option b, i.e. they had agreed to fill in the electronic version of the forms under the guidance of experts of the Puglia region. A calendar of on-site visits was scheduled so that direct observation of users inserting data could be performed. During on-site visits, computer scientists and chemists briefly trained the company employees in using the prototype, then users were required to input the data.

SCHEDA 6 - IMPIANTI TERMICI
(compilare una scheda per ogni impianto termico)

N° impianto termico	1		
Denominazione impianto	GENERATORE DI VAPORE		
Descrizione del tipo di macchina termica	BUZZI		
Tipo (caldaia, forno, altro)	CALDAIA		
Potenza termica[15]	4,8		MW
Attività SNAP[11]	030103		
Codice ATECO [12] Indicare Classe e Categoria	15.11.1		
Codice NOSE-P[13]			
Modalità di svolgimento (barrare la casella corrispondente)	Continua X		Discontinua
Energia termica prodotta	500		GJ/anno
			kWh/anno
Energia elettrica prodotta	0		GJ/anno
	0		kWh/anno
Ore/giorno, giorni/anno, ore/anno	24	120	2880
Combustibile utilizzato	COKE DI PETROLIO TRATTATO		
Quantità di combustibile (consumi annui) (Migliaia m³ per i gas; tonnellate/anno per gli altri combustibili)	6630		Unità di misura (GJ, t, 1000m³, tep) t
% Zolfo[16]	5,57		%
% Ceneri [16]	0,25		%
Ore/giorno, giorni/anno, ore/anno	24	120	2880
Note	//		

Figure 3 An example of a paper form requesting data for a heating generator.

The data entry phase led to unexpected results:

- The prototype required data to be typed into specific labelled fields of the visualized forms. The labels chosen by the chemists in the participatory design team turned out to be too complicated for end users, even after they had been revised to comply with the user comments about the paper prototype. The chemists were often engaged in long discussions with the company representatives to explain the meaning of the various terms.

- The "idealized" company organization that the chemists envisioned during the design team discussions did not exist in reality. For example, the prototype considered only one emission point for each production plant, whereas in many plants there were actually multiple emission points. This caused a lot of problems to users who entered the data.

- Several data about industry production and consumption were not available individually but only aggregated per year; again, there was a difference between the data the prototype expected and the data available in the real world.

Because of these problems, company employees found it very difficult to fill in the required forms. It was clear that, in order to create a successful system that can satisfy end users' needs and expectations, the system would have to be re-designed, no matter how many new resources this might require. The design team went back to work on a complete system re-design, that took into account all problems raised during the visits to the twelve companies, when end users used the electronic prototype in their work places, as part of their real activities. Consistently with what it is reported in [30], this situation enabled end users to provide the most meaningful feedback about all their needs, their expectations, their problems with the system interaction.

Once the new version was ready, it was put online and all the one hundred industries initially contacted were asked to use the system and to input their data. This new version has now been used by all companies and no requests for help have been received. Only one complaint was received, due to a software bug that was easily fixed in one day.

V. LESSON LEARNED

As described in the previous section, the CET system was created by taking into account methodologies that consider usability as a fundamental quality. However, several mistakes were made. The participatory design approach stresses user involvement, requiring end users to participate themselves in the design team and not only domain experts and/or representative of end users. The end users of CET are both company employees and air quality monitoring employees working at the regional government. Such end users were only partially involved. In fact, some of them were interviewed at the very beginning of the analysis phase; other filled the paper forms sent to them and possibly provided comments in the Note field (see Figure 3), but were not observed nor interviewed by HCI experts about their testing experience. Another mistake was that the domain experts in the team claimed to know everything about the functionalities to be implemented, the data to be entered in the system, the way they should be entered, and so on. As it has been described, the test

with end users, who worked with the running prototype, demonstrated a big failure of the design. As it often happens to university professors, the domain experts of the Chemistry Department of the University of Bari had in mind an ideal company organization that did not exist in reality. The many difficulties end users encountered in entering the complex data in the system convinced the design team to re-design the system, as the only way to build a successful system, adapted to people who have to use it rather than requiring people to adapt to the system, as is usually the case with the software systems available on the market.

It is certainly true that end users are generally not capable of explaining their needs or envisioning new systems. They possess tacit knowledge and use it to carry out their activities, but they are unable to externalize it. Since part of this knowledge must be embedded in any system that aims at supporting users' activities, it is mandatory to acquire it in one way or another. On-site visits to end users work settings and interviews are important techniques, but, unfortunately, several IT companies underestimate their importance and consider discussing with end users a waste of time. In other cases, like the one of the CET system described here, the design team did some interviews at the beginning, but refused to involve again end users before creating a more complete prototype, claiming that representative of end users, i.e. chemistry and environment experts, knew everything. This experience provides further evidence that end users and other domain experts possess different types of knowledge, all important for designing successful systems.

Another important finding is not dependent by inappropriate application of UCD techniques, but is similar to what it is reported in [30]: end users provide the most valuable feedback about their possible problems once they get to work with the new system in real settings. End users gave main indications when observed interacting in their workplace with a running prototype pretty close to the final version. Only when a new system impacts their daily practices, end users are able to truly evaluate it and raise significant issues about its usability. This does not mean that involving users in early phases of the design process is of no value, because there is no doubt that they provide useful feedback. For example, the paper prototype used by the seventeen industries that were first contacted was certainly useful for the first re-design of the system. However, the comments received were a lot less numerous and meaningful than those collected during the tests with the end users working with the running prototype in their work settings. This is true not only for the comments indicated in the Note fields, but also for those discussed during phone calls with people who required assistance with filling in the forms. Moreover, various experiences reported in literature ([22], [29], [30]) as well as those of our own research group (e.g., [2], [3]) prove the value of involving users in evaluating mock-ups and early prototypes. The fact that users are very engaged when they put their hands on running software and perform real activities suggests that we have to re-think the different stages of system development and, as clearly said in [30], legitimize post-implementation activities that, instead of being considered a sign of system failure, can help to manage end users' expectations. After the analysis of the CET experience, we were able to convince the CET developers that, even if the CET system was not yet completed in all its components, it would be useful to make some modules available on the Web so that companies could use them to enter their data.

It is worth noting that some requests came from an expert of the regional government who needed to create some specific reports that were not envisioned during the design process. This is not surprising since it is well known that, during use of a system, people's initial requirements are likely to evolve. Already in [21], Nielsen affirmed that "using the system changes the users, and as they change they will use the system in new ways". The new uses of the system cause the working environment and organization to evolve and force the designers to adapt the system in order to meet these new requirements. The new interactive system should be developed to manage this coevolution of users and systems [7], [8]. In other words, system development should not end with its first release; it should be experimented by its end-users and further evolved on the basis of their feedback and/or possible requests for new software artefacts and functionalities.

Some new design methodologies which consider software design as an evolutionary and never-ending process are already being proposed [9], [12]. In other words, the system is in a "perpetual beta" version [27]. A key feature of the methodology described in [7] is that end users work with the successive system versions (starting from basic prototypes) in real settings to accomplish real activities, and the system is evolved on the basis of the results of this work. Thus, end users are not required to envision system use, they actually experience what the final product will be like and how it will impact on their work practice, and this can provide very valuable feedback for system evolution.

VI. CONCLUSIONS

Since the '80s, User-Centred Design (UCD) has been proposed as the basic approach for creating successful interactive systems, capable of satisfying the needs and expectations of continuously increasing user populations. Unfortunately, still today there are several examples of systems that fail to achieve even basic and not very ambitious usability goals, especially because many mistakes are made in involving users in the design and development of software, which is the basic principle of UCD.

This paper aims at providing insights on how UCD can be applied in a more effective way. Such insights are derived from experiences of our research group that analysed how design and development of interactive systems are carried out in various IT companies in our country, and in particular from the one described in this paper, which refers to the CET system, commissioned by the Puglia region government with the aim of monitoring the air quality. As discussed in this paper, some mistakes were made during system design and development. The lesson learned provides hints for a more effective application of UCD and the fruitful involvement of users so as to create successful systems. These can be summarized as follows:

a) To make UCD effective, contextual analysis must be performed that includes interviews not only with domain

experts but also with end users, and on-site visits to observe end users' practices, in order to correctly capture users' requirements and users' tacit knowledge.

b) End users, namely those people that will actually use the system, must be involved in the software development cycle. Domain experts, who are often considered users' representatives, are in most cases unaware of the working practices that need to be captured and embedded in the system, to make it truly effective.

The above two remarks are not certainly new in the UCD methodology, but they are worth mentioning since, unfortunately, UCD is still misunderstood by IT companies, at least in Italy. A final, less obvious, remark is:

c) End users can raise significant issues about system usability only when they get down to using the system, or even a running prototype, in their real activity settings.

Overall, the experience with users involvement in UCD we have reported is very much in line with what it is discussed in [30]: in order to design successful systems, it is fundamental to engage end users at the right times, when they can really provide useful contribution. The CET team failed to involve the end users, and this badly affected the developed system. The failure was evident only when end users put their hands on the system in their working place, because they could then provide the right indications about what was working well and what was not. This does not mean that users should not also be involved in early phases of the design process, but it is an indication that the whole life cycle of software systems must be revised, viewing software design and development as an evolutionary process [9], [12], [27]. This because user feedback is very valuable when people work with the system in their working place; this feedback helps evolving system in order to manage end users' expectations.

ACKNOWLEDGMENT

Partial support for this research was provided by EU through grant L4ALL and through funds provided to the Consortium which developed the CET system.

REFERENCES

[1] C. Abras, D. Maloney-Krichmar, J. Preece, User-Centered Design, in W.S. Bainbridge, (Ed.), Berkshire Encyclopedia of Human-Computer Interaction, Great Barrington, MA: Berkshire Publishing Group, 2004.

[2] C. Ardito, P. Buono, M. F. Costabile, The Challenge of Visualizing Patient Histories on a Mobile Device, in: M.F. Costabile M., F. Paterno' (Eds.) Human-Computer Interaction - INTERACT 2005, vol. LNCS 3585, Berlin: Springer, Germany, 2005, pp. 942-945.

[3] C. Ardito, P. Buono, M.F. Costabile, R. Lanzilotti, A. Piccinno, Enabling interactive exploration of cultural heritage: an experience of designing systems for mobile devices. Knowledge Technology and Policy, 22(1), 2009, pp. 79-86. doi:10.1007/s12130-009-9079-7.

[4] ARPA Regione Lombardia, Progetto INEMAR, http://www.ambiente .regione.lombardia.it/inemar, accessed on June 2009.

[5] R.G. Bias, D.J. Mayhew, Cost-Justifying Usability, Second Edition. Morgan Kaufmann, San Francisco, CA, 2005.

[6] B. Boehm, Characteristics of Software Quality, North Holland Publishing Co., New York, 1978.

[7] J.M. Carroll, M.B. Rosson, Deliberated evolution: stalking the view matcher in design space, Human-Computer Interaction, 6 (3), 1991, pp. 281-318.

[8] M.F. Costabile, D. Fogli, A. Marcante and A. Piccinno, Supporting Interaction and Co-evolution of Users and Systems, in: Proc. of International Conference on Advanced Visual Interface, ACM Press, Venice, May 23-26, 2006, pp. 143-150.

[9] M.F. Costabile, D. Fogli, P. Mussio and A. Piccinno, Visual Interactive Systems for End-User Development: a Model-based Design Methodology, IEEE Transactions on System Man and Cybernetics Part A-Systems and Humans, 37(6), 2007, pp. 1029-1046.

[10] A. Dix, J. Finlay, G. Abowd, R. Beale, Human-Computer Interaction, Third ed. Pearson-Prentice Hall, Essex, UK, 2003.

[11] European Pollutant Emission Register, EPER Database, http://eper.ec.europa.eu/eper/, accessed on June 2009.

[12] G. Fischer, Seeding Evolutionary Growth and Reseeding: Constructing, Capturing and Evolving Knowledge in Domain-Oriented Design Environments, Automated Software Engineering, 5(4), pp. 447-464, 1998.

[13] Google Earth, http://earth.google.it/, accessed on June 2009.

[14] K. Hornbaek, J. Stage Editors. Special Issue on The interplay between usability evaluation and user interaction design. Int. J. of Human-Computer Interaction, 21 (2), Nov 2006, pp. 117-123.

[15] ISO (International Organization for Standardization), ISO 13407: Human-Centered Design Process for Interactive Systems, 1999.

[16] N. Juristo, A. Moreno, M.I. Sanchez-Segura, Guidelines for Eliciting Usability Functionalities, IEEE Transactions on Software Engineering, 33 (11), Sept. 2007, pp. 744-758.

[17] M. Maguire, Methods to support human-centred design, Intl. J. Human-Computer Studies, 55,(4), 2001, pp. 587-634.

[18] D.J. Mayhew, Usability engineering lifecycle, Morgan Kaufmann San Francisco, CA, 1999.

[19] J.A. McCall, P. Richards, J. Walters, Factors in Software Quality, three volumes, NTIS AD-A049-014, 015, 055, 1977.

[20] J.A. McCall, Quality factors, in: Encyclopaedia of Software Engineering, J.J. Marciniak (Ed.), John Wiley & Sons, New York, 1994, pp. 958-969.

[21] J. Nielsen, Usability engineering. Academic Press, Cambridge, USA, 1993.

[22] E. Mumford, Redesigning Human Systems. Hershey, PA: Information Science Publishing, 2003.

[23] D.A. Norman, S.W. Draper, (Eds.), User-Centered System Design: New Perspectives on Human-Computer Interaction, Lawrence Earlbaum Associates, Hillsdale, NJ., 1986.

[24] C.M. Karat, A comparison of user interface evaluation methods, in: J. Nielsen, R. L. Mack, (Eds.), Usability Inspection methods, John Wiley & Sons, New York, NY, 1994, pp. 203-233.

[25] C.M. Karat, Cost-justifying usability engineering in the software life cycle, in: Handbook of human-computer interaction, Elsevier Science Inc., Second Edition, 1997, pp. 767-778.

[26] S. Reed, Who defines usability? You do!. PC Computing, Dec. 1992, pp. 220-232.

[27] G. Stevens, V. Pipek, V. Wulf, Appropriation Infrastructure: Supporting the Design of Usages, in: V. Pipek, M. B. Rosson, B. de Ruyter, V. Wulf, (Eds.), End-User Development, vol. LNCS 5435, Springer, Germany, Berlin, 2009, pp. 50-69.

[28] D. Schuler, A. Namioka, Participatory Design: Principles and Practices. L. Erlbaum Associates Inc. Hillsdale, NJ, USA, 1993.

[29] E.L. Wagner and S. Newell, Exploring the importance of participation in the post-implementation period of an Enterprise System Project: A neglected area, Journal of the Association of Information Systems, 8(10), 2007, pp. 508-524.

[30] J.L.Wagner, G. Piccoli, Moving beyond user participation to achieve successful IS design, Communications of the ACM, vol. 50, no. 12, 2007, pp. 51-55.

Proceedings
International Workshop on
Distant Education Technology
(DET 2010)

Co-Editors

Rita Francese, *University of Salerno, Italy*
Mahbub Syed, *Minnesota State University, USA*
Kazuo Yana, *Hosei University, Japan*

Enforcing Team Cooperation: an example of Computer Supported Collaborative Learning in Software Engineering

Mauro Coccoli[1], Paolo Maresca[2], and Lidia Stanganelli[1]
[1]Department of Computer, Communication and Systems Science, University of Genoa
Viale Causa, 13 - 16145 Genova, Italy
{mauro.coccoli, lidia.stanganelli}@unige.it
[2]Department of Computer and Systems Engineering, University of Naples, Federico II
paolo.maresca@unina.it

Abstract - *In this paper the OTRE experience is presented, as a case study in a larger project named Enforcing Team Cooperation (ETC). Cooperation is one of the keywords in education and in computer assisted instruction, and thus e-learning platforms provide users with specific tools, enabling them to collaborate and/or to cooperate to reach a common objective. Collaboration is considered as a teaching strategy but, in software engineering classes, it has to be a learning outcome too, since students must acquire ability in team working. A specific working environment is needed and the authors propose an experimental set-up based on the Eclipse and Jazz technologies.*

Keywords- *Team Cooperation, E-learning, Knowledge Management, Software Engineering, Eclipse-Jazz*

Introduction

This paper deals with the problem of teaching software engineering in higher education. The idea of team cooperation is not part of the syllabus of software courses, thus developing team cooperation skills in students at University level is strongly needed. To this aim, a specific project (OTRE) has been launched within the Italian branch of the Eclipse Community, with the objective of enforcing and enlarging cooperation activities among the large number of students involved in software engineering courses at the Universities of Italy. OTRE[1] is the acronym for *"On The Road to Eclipse-IT 2010"*. The Eclipse-IT 2010[2] is the fifth workshop of the Eclipse Italian community, to be held in Savona (Italy), on the forthcoming Sept. 30TH and Oct. 1ST, 2010. The workshop will be the place for the scientific and academic community to meet the participants, to share their research and experience with the Eclipse developers, and to meet the public institutions and industries that use, or are interested in adopting, the Eclipse framework. The main topic of the OTRE project is the creation of an effective Computer Supported Collaborative System (CSCL), to be used for higher education on team cooperation, in software engineering classes for the analysis, design, and development of software programs along their lifecycle. The idea arises from the observation that, in many cases, software engineering suffers from the lack of a really full integration among different development teams. The proposed solution is teaching to work in team, together with languages and tools for programming computers. Based on the novel programming paradigms and tools specifically created for supporting team-work, a specific didactic strategy based on social constructionism and collaborativism is needed. Moreover, a suited software platform that enable effective working is also necessary, especially in order to coordinate cooperation in developing for people that usually work in remote sites, even in different countries, and belonging to different Universities or Research Institutes. Many systems already exist for the coordination of work, but recent Integrated Development Environments (IDEs) and the last research trends are overcoming the idea lying behind them; more importance is given to the remote developing and cooperation as the way to cut time and money. Such a kind of environments need to have something like an "orchestra Director" over the process of developing. The orchestra Director is the one who knows when and how, each instrument has to be better used, when it has to spend its contribution, and how to leverage the quality of all execution. A realization of such an environment exists and it is named Jazz[3]. The

[1] An *otre*, in Italy, is a leather bottle, more precisely a water-bag, in the past used during travels; from an *otre* one can drink and in the *otre* one can protect something precious like the water. In this respect, the authors like to think that one of the aims of the *OTRE* Project is to protect knowledge and experience and give access to others, sharing the content.

[2] See http://2010.Eclipse-IT.org

[3] See http://jazz.net

Jazz system is a complex development platform, released by the brand rational of IBM, that is usually used in developing by co-operating from the IBM researchers themselves, in a large world-wide distributed organization. Up to now, such a kind of complex platform has been considered "professional" and it has never been used before in a University educational activity. For the present OTRE experience, the Eclipse IDE, together with the Jazz environment, is the working platform that has been chosen.
The reminder of the paper is the following: first the OTRE project is presented, then CSCL is considered as a general purpose learning strategy and applied to the case of software engineering. Next, the Eclipse framework as a learning environment is presented while giving some more details about the OTRE project. Conclusions will follow.

The OTRE Project

The OTRE Project can be briefly described as an experiment of realization of a collaborative project centered around the Eclipse community and tools; at the end, the results will be collected, published and shared, in a workshop in which the teams that will have collaborated remotely, finally will have the opportunity of meeting themselves as well as comparing their work and results with other teams involved in the same project. Something like a peer-to-peer conference in the conference, that wants to be much more then a student contest. Each team of students will have to complete a group of tasks whose outcome will be the realization of an Eclipse plug-in that will have to be integrated with the Eclipse plug-ins developed by the other teams. The OTRE project is part of the program *"Enforcing Team Cooperation using rational software tools into software engineering"*, an innovative academic project sponsored by the Italian branch of IBM, *IBM-Italia*. At the actual stage, a restricted number of Universities in Italy have been involved in the pilot project and the participants are the following: Naples, Federico II Naples; Milan, Bicocca; Alma Mater, Bologna; Bergamo; Genoa and his Savona; Bari and his Taranto. The students involved, and signed-up within the platform, are 263 and they all will work in their initial first semester. Each University has gone forming teams of students from different courses such as: software engineering courses from the Universities of Naples Federico II, Bologna and Milan Bicocca; web design and development courses from the Universities of Genoa and Bari; advanced programming course from the University of Bergamo. Heterogeneous and distributed teams will be composed, with students from different cities and with one teacher tutor for each. In addiction, for each University, a Champion Student (usually a computer engineering or computer science, PhD student) will support the corresponding teacher. Moreover, the Champion Student is the responsible of the local group. One only computer engineering PhD student (administrator) has the technical direction for the overall ETC platform, both software and hardware. The topics addressed in the project will be directed to a re-engineering of the actual Eclipse Italian Community web site. New services will be designed and developed, with the objective of making the site itself the perfect place to repeat the experimentation after this first group of pioneers students and teachers. Future classes will participate, adding further functionality to the system, integrated with the ones already available. After the first running phase, the OTRE will represent a repeatable model so that new experiences can be done without the start-up period and with quick set-up time and a smooth learning curve. According to the specifications given by the teachers, the newly realized Eclipse-IT portal will have to integrate audio and video contents, paying attention to accessibility and making multi-format information available to the visitors. In this present edition of the OTRE project, the activity will be centered around a set of modular components that will be designed and realized to manage events. Possible services will be: a live-conference system (live streaming of the talks given at the conference); news management; live-blogging; interactive boards supporting chat sessions and live discussions; collaborative systems such as wiki for the documentation management (user, developer, and reference manual, documentation, etc.) as well for the organization of the conference agenda and the topics in a bar-camp style; discussion boards for exchanging opinions; sharing systems; publication tools addressed to different targets such as scientific community, developer community, generic and specialist press; user data and profiles management; mash-up tools for the data fusion and aggregation; a ”chairman” system to manage submissions and review of the papers for a conference; a repository and indexing system for the archiving of the work done and keeping track of interactions; content management; advanced search capabilities; semantic systems with specific handlers for software libraries and the Eclipse plug-ins in particular.

Computer Supported Collaborative Learning

The adoption of a collaborative learning strategy can be useful in many situations, and it can be realized with or without the use of technology. In such a student-centered model, learning is achieved through the interactions that happen among peers, while cooperating to cope with a given problem. It appears obvious that this approach may have weaknesses and strengths [1], mainly connected to the responsibility that, compared to a traditional model, is shifted from the lecturers to the students themselves, which play a more active role in their own learning process. Summarizing, CSCL is a powerful didactic strategy for both the learners and the teachers, when communicating and sharing are considered added value and where the social aspects go beyond the learning activity itself. By means of the Information and Communication Technology (ICT), communication, collaboration, and thus the learning experience, can be greatly enhanced so to achieve better results in education.

CSCL in Software Engineering

Traditional CSCL is based on the use of standard collaboration tools integrated in Learning Management Systems, (e.g., synchronous and/or asynchronous

communication systems, discussion boards as well as forum, blog, wiki, and social networking, resource sharing systems such as teleconference tools and shared whiteboards). A slightly different approach is needed when students are supposed to work in a collaborative environment where the collaboration is part of the task to be achieved, not just a part of the learning process. That is the case of software engineering in which code development is a collaborative process and teams of developers have to work together to design, solve problems, and produce code as well. Team cooperation is at the basis of the development of any software product [2]. Moreover, recent trends are driving software engineering towards the use of non-traditional development methodology (e.g., agile programming, eXtreme programming) [3] where computer-based collaborative tools are needed to support coordination of work as well as communication, both structured or unstructured, and synchronous or asynchronous. In complex projects, as well as in simpler ones, different persons have to interact in a group (e.g. analysis, design, development, graphics, documentation, user interface, accessibility, interoperability, APIs) and the outcomes of independent groups have to be integrated so that they can work together in just one product. This can be achieved by using the unified software development process [4] and the Unified Modeling Language (UML) [5] so that different groups, at different levels of detail, can use a common language to give each other instructions as well as to exchange information. At engineering schools, in the field of the education of future software engineers, collaboration has to be considered in terms of the construction of collaborative parts in a collaborative environment that can realize the spiral model used for software development [6]. Software engineering and the use of the CASE (Computer Aided Software Engineering) model, make available specific tools (e.g., common versioning systems, source control, bug tracking systems) and specific development environments offering advanced collaboration functionalities. The process is meant to be the collaborative construction of one "big thing" for which objects are developed in a collaborative way; moreover, the developed objects them-selves have to be collaborative.

To this aim, the Eclipse framework offers to software developers and designers a set of tools for empowering collaboration and enforcing team cooperation.

Communication in Collaboration

A good communication is the key-point in collaboration and in cooperation. For software engineers the development platform is expected to be the enabler for a good communication to take place at different levels: developers working on the same application; practitioners across the life cycle; business analysts; architects; developers; testers; project managers.

Also, communication takes place amongst many different subjects: content (technology, business domain); project (timelines, tasks, resources); process (sequences, dependencies); organizations (responsibilities).

Collaboration in a workgroup is made possible by means of technology enabling the people in that workgroup to a seamless work, as if they were co-located, in order to: improve communication; create a common context; share information; notify events.

Eclipse as a Learning Environment

In this Section the Eclipse IDE is presented and it is pointed out how it can be considered as a learning framework. Eclipse is an open universal platform for tool integration. It is an open and extensible Integrated Development Environment, and, in the same time, an open source community. The Eclipse community creates royalty-free technology as a universal platform for tools integration. The aim of Eclipse developers, and hence of the Eclipse-based tools, is giving developers the freedom of choice in a multi-language, multi-platform, multi-vendor environment supported by multiple vendors. The operating system platforms that Eclipse has been targeted to, include MS-Windows, Linux, Solaris, HP-UX, AIX, and Mac OS. In addition, Eclipse provides a unique environment for members of the academic community to build new tools for teaching, research, and further growth of the Eclipse community [7]. Collaboration is an integral part of software development, occurring through tools inside and outside the IDE. The Jazz project, which seeks to integrate collaborative capabilities into the Eclipse IDE, enables small teams of software developers to work together more productively. A very clear image of the Jazz philosophy is that *"Jazz is based on the metaphor of an* open office *approach to application development. [...] Within the office, each developer will have his or her own workstation, while elsewhere in the office will be space for team meetings, shared whiteboards, schedule information, etc. A key aspect of this setting is team awareness. Even while focusing on their personal work, all team members have a sense of what is going on elsewhere in the room, who is talking to whom, what others are working on, etc. Communication among team members is also facilitated in this environment, whether in the form of a question shouted out to the team in general, or in calling a colleague over to consult on an issue at a particular workstation"* [8].

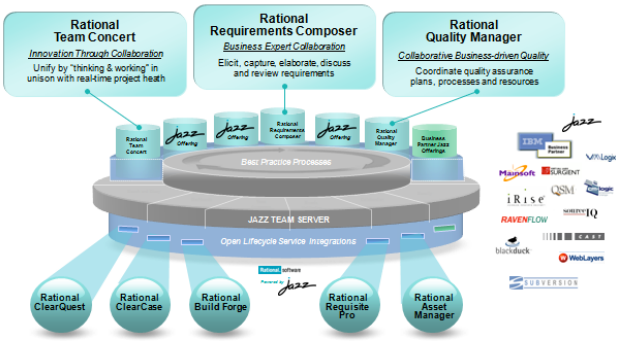


Figure 1. The Jazz architecture

Team cooperation

Team cooperation is enabled by Jazz platform [9] as shown in Figure 1. The Jazz platform is an Eclipse application composed by other three tools: the *Rational Team Concert* (RTC), the *Rational Quality Manager* (RQM), and the *Rational Requirement Composer* (RRC). The three tools assist different teams to co-operate in

developing software specification respecting quality constraints and realizing a hybrid between an architecturally layered (three tiers) and a service oriented architecture, based on the paradigm of development cooperation projects, thus it has both the means for conception of projects for its implementation and verification of quality. The architecture of RTC is innovative because it combines both the process of developing the tools and the process needed to implement them. Above all have built-in a number of process models of development, between these agile ones. It is believed that the latter may be the subject of the trial given the limited time for completion of tasks (9 weeks). In particular, we plan to take as SCRUM agile development process.

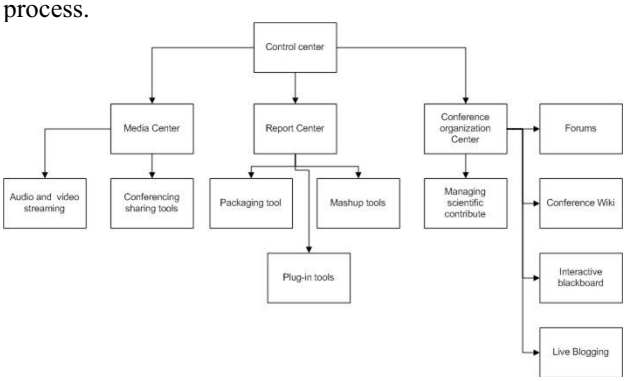


Figure. 2 The OTRE Architecture

OTRE Software Architecture

A specific task in the project, involves the University of Naples Federico II and the University of Genoa (and his Savona), with the aim of creating web-services in order to manage the Eclipse 2010 event. The system handling the event will be composed of the following abstract functions: control center with coordination role; media center that aims to manage all media documents (voice, video, streams, etc) including conferencing tools (live presentation, Q&A, etc); report center able to extract documents and whatever will be useful as documentation of event; conference organization center aims to make available such a useful tools from organization point of view. For the sake of brevity the blocks shown in Figure 2 will be no further detailed; just some additional information about some strategic components will be given. Live-conferencing tools for streaming audio and streaming video of the talk of the conference is a good choice for a practice community that plans to grow; these information will be accessible from both the companies and the students. Connected to these tools there are other cooperative tools such as live blogging, blackboard and wiki, finalized to support live-narration during the conference and post-conference discussion. It is worthwhile to outline that wiki tools could be useful also for collaborative construction of topics for the first call for papers and agenda of the day after, by the program Committee. For not-attending people, sharing tools, browser for remote viewing of presentations by the various channels through video streaming are available for use. Also it is planned the inclusion of tools to package and sharing web technical articles and essays in order to extract data in order to publish them onto journal or magazine. A more specific resource is represented by a mash-up activity that aims to manage conference data (both video and audio) in order to extract and abstract behavior of virtual communities and groups during both the development of the project and the community activities (conferences, sharing knowledge, etc.). Classical conference managing tools include the ones that manage the papers presented as the model of the *easychair* conference system[4]. Finally, tools for archiving and cataloging the plug-ins developed by students and professionals submitting to the community in the annual workshop will be available. These will own capabilities of semantic content search and classification with plug-in for document management and software libraries management. The project will involve groups of students with different roles predefined and geographically distributed throughout Italy between universities that are part of the experimentation.

Conclusions

In this paper, the outline of a running project has been presented. Results from the activity of the students (and involved teachers) will be soon available and will be able to give a measure of the effectiveness of the chosen approach. Besides, a more general framework for the development of team-working skills is presented, based on the Eclipse and Jazz framework and tools. Such a test-bed can be considered as a learning tool itself and it can be re-used in many different applications, even different from the software engineering we have been referring to along the whole paper. Especially in Public Institutions and large Organizations, cooperating to reach a common objective can be a critical point to achieve results and meet the milestones. The availability of a framework enabling users to cooperate through advanced software systems can greatly empower such scenarios and interesting perspectives can be considered in many other fields.

Bibliography

[1] S. Wiliams and T.S. Roberts, "Computer Supported Collaborative Learning: Strengths and Weaknesses," *Proc. Int. on Computers in Education*, pp. 328-331, Dec. 2002.

[2] G. Booch and A. Brown, "Collaborative Development Environments," *Advances in Computers,* vol. 59, Academic Press, Aug. 2003.

[3] C. Ghezzi, M. Jazayeri, and D. Mandrioli, *Fundamentals of Sotware Engineering.* Englewood Cliffs, NJ: Prentice Hall, 2002.

[4] I. Jacobson, G. Booch, and J. Rumbaugh, "The Unified Process," *IEEE Software*, vol. 16, no. 3, pp. 96-102, May/June 1999.

[5] OMG - Object Management Group, "The Unified Modeling Language," http://www.uml.org, 2010.

[6] B. Boehm, "A Spiral Model of Software Development and Enhancement," *Software Engineering Notes*, vol. 11, no. 14, pp. 14-24, ACM, New York, 1986.

[7] IBM, "Eclipse," http://www.research.ibm.com/eclipse. 2010.

[8] L. Cheng, S. Hupfer, S. Ross, and J. Patterson, "Jazzing up Eclise with Collaborative tools," *Proc. of the 2003 OOPSLA Workshop on Eclipse Technology Exchange*, pp. 45-49, Anaheim, CA, 2003.

[9] R. Frost, "Jazz and the Eclipse way of Collaboration," *IEEE Software*, vol. 24, no. 6, pp. 114-117, Nov./Dec. 2007.

[4] See http://easychair.org

The assessment of knowledge requirements in a network of goal oriented organization

G. Franco
IBM Italia
Segrate (MI), Italy
giacomo_franco@it.ibm.com

P. Maresca
DIS Università di Napoli Federico II
Napoli, Italy
paolo.maresca@unina.it

G.Nota
DMI Università di Salerno
Fisciano (AV), Italy
nota@unisa.it

Abstract—**Many organizations that conceive their work in terms of goal to pursue wish to determine the knowledge requirement in order to assess resources, steps and necessary efforts to reach the goal. It is indeed a complex process that can receive support from other organizations available to share their know how with other organizations participating to a network of cooperating organizations. In this work we propose a distributed framework oriented to the knowledge management within a network of organizations that pursue similar or shared goals. Starting from the goal identification, the assessment of knowledge requirements to acquire is obtained using the Goal/Question/Metrics approach. The organization that wants to acquire useful knowledge toward the achievement of shared goals, evaluates the necessary effort and provides to other organizations, in explicit form, its experience for what concerns the aspects knowledge as objects and knowledge as processes.**
Two case studies show possible applications of the proposed framework.

Keywords- knowledge engineering; knowledge management; knowledge sharing; cooperative organizations; knowledge sharing assessment; Goal/Question/Metrics (GQM);E-learning.

I. INTRODUCTION

The determination of knowledge requirement to achieve goals is a perennial challenge in the activities of organizations today as it impacts on the acquisition strategy of new resources, on kind and level of competencies their members must have to face activities and on which extent of collaboration and cooperation they and/or their members must establish with the outside world to fill needed competencies.

The determination is even more difficult in many cases, for example when: 1) actors performing activities driven by objectives are part of a set of cooperating organizations, 2) a single organization is unable to entirely meet this demand of knowledge and it needs to define how much of that knowledge is available in "neighboring" organizations available for cooperation, 3) a single network of organizations that needs to use the knowledge in neighboring organization networks to achieve shared goals. Daily working by objective can be seen as a practice of modern organizations and there is a lack of models that help to determine knowledge requirements to achieve the goals. The impact of the knowledge requirement determination problem is higher in a knowledge network (KN) that comprises a network of organizations working by objectives. The KN comprises individuals and organizations willing to cooperate with each other; this type of models should provide a structure of KN available to organizations, it should facilitate the dissemination of requirements related to objectives, research and quantification of that knowledge in inter/intra-organizational networks to feed the measurement process.

The paper offers a systemic approach that integrates various levels to determine goals and Knowledge Management (KM) for the assessment of knowledge-related requirements: individual level, intra-organization level and inter-organization level. As a matter of fact, to identify and assess the need of knowledge useful to reach goals means:

a) To identify a priori schema of goals and sub-goals together with related activities and knowledge necessary to reach the goals.

b) To assess efforts for knowledge acquisition.

c) To be aware that these goals may change over time as a result of strategic changes.

d) To provide ways for the assessment of high order knowledge-related requirements starting from low order assessment data.

It might also happen that new targets emerge in the course of activities (new goals linked to previous one, e.g. evolving requirements in software development projects) eventually augmenting or replacing those previously established and that this scenario requires change management capabilities. The impact of dynamism of objectives can be captured only by a vision and a systems approach to their definition and measurement. To support the achievement of these goals a

systems approach is at least useful, if not essential, in the design and structuring KN too.

Through studies of recent decades on both the real concept of the knowledge and on its management and distribution inside organizations, a window has been opened on their important role as a corporate asset. Groupware and management information systems [7], [8] have made concrete theories and processes of KM through the use of technology, contributing to interoperability of organizations that work together by exchanging knowledge. However they have not sufficiently addressed the emerging nature of knowledge [16]. Indeed, the concept of knowledge involves implicitly emerging properties and systemic characteristics that cannot be assessed or collected in advance, since knowledge, unlike data or information, emerges by human interpretation and by the complex interactions between human beings [18].

Although the accidental nature of knowledge is increasingly seen as an issue in the management of organizational knowledge, many studies indicate that the emerging issues and self-organization of knowledge are a vital source of creative ability and flexibility of the strategic organizations [21], [20], [3], [18].

The assessment of knowledge requirements can be approached considering two important points of view: "K = O" where the dominant, typical of technological contexts, is "knowledge as object", and the "K = P", which establishes the dominance of a "knowledge as process", typical of sociology and other human sciences (which also contain the visions "knowledge as interpretation" and "knowledge as relationship"). These two points of view have to converge in a single model which can assess both aspects "objective" and "subjective" associated with the knowledge. A desired effect of this model is the chance to design, predict, measure and control the spread of new knowledge and the effects of change that it induces in an organization.

Moreover, using the metaphor of complex systems, the network of organizations may be perceived as a meta-virtual organization (corresponding to the knowledge network), with new properties emerging from interactions among the organizations participating in the network. Some of the emergent properties are useful to knowledge selection and dissemination to achieve common goals in the network of organizations or individuals [22], [23], [24], [14]:

- increased total knowledge of the network,
- growth of competencies of individuals who populate organizations,
- reduced time and work necessary to raise the level of skills in the distributed network.

In this paper we address the problem of how to assess the knowledge requirement in a network of goal oriented organization. To this end, an approach based on the paradigms K=O e K=P is used to identify the Objects of Knowledge and the Processes of Knowledge that must be measured.

We first discuss the elements of a Distributed Knowledge Management Framework (DKMF) for a systemic approach to KM able to represent common goals and knowledge. Then, we will describe how the well known Goal Question Metrics (GQM) approach can be integrated in the proposed DKMF in order to assess Objects of Knowledge and Processes of Knowledge. The integration of DKFM with GQM provides an increased awareness level on the knowledge requirements necessary to reach common goals in a network of goal oriented organizations.

In the following it will be discussed:

1) How the framework DKMF allows to take into account the emergence of the knowledge within networks of goal oriented organizations.

2) Which is its role in sharing and dissemination of knowledge.

3) The elements of the model Goal / Question / Metrics (GQM) [27] as a guideline to implement a process of knowledge assessment.

4) How are integrated the framework and GQM into a single systemic vision.

The validation of the framework integrated starting from DKMF and GQM has been pursued studying its application first to the evaluation of an e-learning platform with major emphasis on the K=O perspective and then to the process devoted to the design and construction of plants for alternative energy.

II. SHARING KNOWLEDGE IN A DISTRIBUTED NETWORK OF GOAL ORIENTED ORGANIZATION

The research results obtained in the last years in the science of networks [2] and in the field of Knowledge Management [24], encourage further efforts to cope with problems and perspectives not yet addressed by the literature in the context of distributed networks of goal oriented organizations.

We believe that the above mentioned works are reference points to start research on how KM can be better understood in networks of cooperating organizations. In particular, the *top-level conceptual framework* shown in [24] points out the foundation of a KM discipline identifying its four pillars: *leadership*, *organization*, *technology*, and *learning* that assure the practicality of the KM life cycles usually expressed as activities of generation, codification, use, and transfer.
In our previous works we explore some variant of the work due to Stankosky [24]. In [34] a *Distributed Knowledge Management Framework* (DKMF) that uses the top-level KM framework as a building block in a distributed network of cooperating organizations is proposed. In particular, aspects that characterize the dimension of "social influence" are expressed and characterized in terms of identity, negotiation and trust. The common identity for a network of cooperating organizations is evaluated as one of the enabling factors

toward knowledge exchange and sharing. It become stable by means of a negotiation activity pursued at several levels: individual, organization, virtual meta-organization (referred as Virtual Community or VC in the following) and creates the humus for trust and cooperation between the organizations participating to the VC.

Absorptive capacity has been defined as "the firm internal ability to recognize, assimilate and utilize new information and knowledge" [6]. The concept of *absorptive capacity of a DKMF* is discussed in [35] where it is shown how the exploitation of absorptive capacity and adaptive behavior in a scale free network can enhance significantly the capability of a social network to acquire new knowledge. Other contributions in the direction of social network support in knowledge sharing can be found in [28], [29].

The DKMF shown in "Fig. 1", is an augmented version of those presented in [34] as it considers the management of shared goals. It is built starting from three fundamental components:

1) Local Knowledge Managers (LKM): They are the holder of local knowledge (i.e. an organization) that are available to share their knowledge in a specific application domain with other community participants in order to reach common goals. A LKM is organized according to the top-level conceptual framework for KM.

2) Virtual Community Supervisor (VCS): It is the entity of fundamental importance during the start up phase in a KN; apart from the usual functions assigned to LKMs, it can: design a new network structure, assume the leadership for the governance of a VC, construct the identity of a KN stating the rules that allow the participation, collect from LKMs the goals to be discussed as possible candidate for goal sharing.

3) Virtual Knowledge Repository (VR): Is the virtual place that stores the shared knowledge; the VC participants usually own "chunks of knowledge" tacit or explicit and are encouraged to produce/consume knowledge from the VR which is perceived as a centralized knowledge base although it is really distributed in many virtual or physical places.

The VR is the structure that allows to store as well as to communicate the knowledge software. It represents a particular instance of organized knowledge in a given application domain, usually stored in many places, and is fed by human being that participate to the Virtual Community.

If we consider the scenario of goal oriented cooperating organizations, then remarkable *chunks of goal oriented knowledge* to be shared are:

a) The structure of goals and subgoals that the cooperating organizations try to reach.

b) The knowledge to acquire in order to reach the goals (gap of knowledge).

c) The assessment of acquisition efforts.

d) The common strategies.

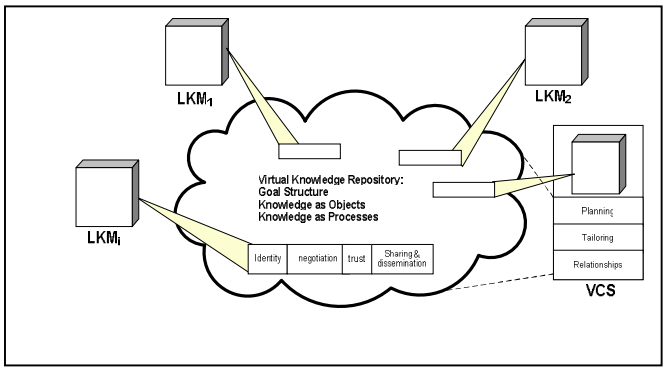


Figure 1. The Distributed Knowledge Management Framework

In the following sections we discuss how the integration of a DKMF together with a methodology for the measurement of entities such as work products, activities and processes can support the assessment of knowledge acquisition efforts in a network of goal oriented cooperating organizations. First, we outline the fundamental steps necessary to build a DKMF in a given application domain as an instance of the general framework, then we will turn to the knowledge assessment problem.

The practical use of DKMF in the process of construction of a knowledge network requires the activation of four distinct phases:

4) Planning: the four KM pillars (leadership, organization, technology, e learning) are defined at the metalevel of virtual community. This includes the social and communication structures, the identification of protocols for the knowledge transfer.

5) Tailoring: the application context as well as the common goals is chosen. The models for the knowledge representation are identified and the first artifact are produced and shared.

6) Relationships: the identities of VC participants are first acquired; then, trust relationships and negotiations modalities are established. The rules that allow LKM to consume/produce knowledge from/to the VR are also specified.

7) Use: a LKM share its knowledge with other VC participants and receives available knowledge from the VR according to the rules and to the transmission modalities (e.g. on demand, planned transfer, asynchronous communication, announcement, notification, etc.).

III. GOALS ASSESSMENT IN A KNOWLEDGE ORIENTED PARADIGM

Real domains present a big challenge to quantification of transferred knowledge due to their peculiarities. However, in order to gather data on transferred knowledge, it is possible to use best practices normally used to evaluate quality of internet services. In fact, the process of knowledge transfer to increase the qualitative level of network of organizations could be compared to a service, each organization inside the VC can supply or use it trough the DKMF.

Quality evaluation of this service is tightly related to an expensive process of harvesting many data resulting from assessments made to satisfy target objectives. There are not general norms to standardize this process, but Goal/Question/Metrics (GQM) [26], [27], [31], [32] is a good guideline to implement an assessment process using a systematic and top-down approach.

GQM provides guidelines to develop a methodology to define and to build an "ad hoc" measurement program. This measurement program will complete the system that satisfies the specific organizational needs. GQM helps to define: 1) goals for which we wish to determine the achievement degree; 2) measurement process detailed through questions supporting metrics, surveying modalities and scheduling of surveys and operation.

The entire GQM process is shown in "Fig. 2", it is a top down approach that starts from the goals to the metrics used to evaluate the degree of goals achievement. It is independent from the application dominion and it allows you to get very specific results for the context under test. GQM is based on the following questions:

- What do we want to measure? (Goals)
- Which factors do they engrave? (Questions)
- Which data have I to pick up and with which metrics have it appraise them? (Metrics)
- What are the points of view through which can I look my measure activity? (Quality Focus)

Therefore, we must define the goals, get some questions or hypothesis from each one, quantify the goal by collecting data or simply answering questions and finally with the collected data validate our hypothesis. It is important considering that the questions must cover all the goals. Sometimes, in order to answer questions, we must get measures (the Metrics). These metrics, when compared with threshold values (if available), gives you a measure of goals satisfaction. In [10] the concept of cluster is introduced. A cluster is a macro goals assemble, a set of goals about the same argument. For instance, in the cluster Communication Tools all the goals about communication tools (synchronous and asynchronous) are joined. A full GQM plan example has been given in [9], [11], [12].

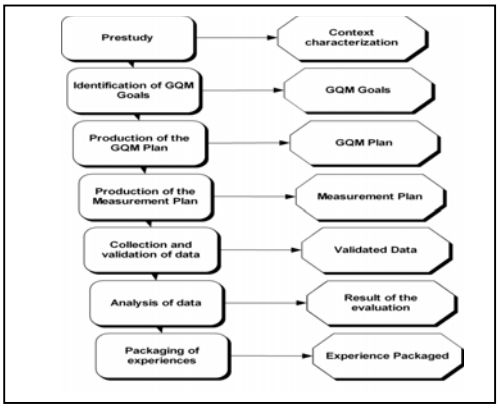


Figure 2. The GQM Process

In the process of construction of a DKMN that supports the achievement of common goals within a network of organizations, the tailoring phase discussed in the previous section is further specialized in order to handle goal management and the assessment of knowledge needs. These capabilities arise integrating the GQM approach within a DKMF as follow:

1) tailoring the DKMF to the problem (include the first three DKMF phases: Planning. Tailoring, Relationships)

2) goals identification, goals structure, goal sharing (include the first three GQM steps: Prestudy goal identification, GQM plan production)

3) Evaluation Effort necessary to overcome the knowledge gap (include step 4 of GQM: production of the measurement plan)

4) Performance Monitoring in order to reach the goals (include step 4 of DKMF (Use) and the last three steps of GQM: data collection, data analysis and packaging of experiences).

IV. CASE STUDY 1: E-LEARNING PLATFORM EVALUATION

The real domain of E-learning platform evaluation concerns several organizations (e.g. universities) that share their knowledge (e.g. technical and user experiences) on e-learning platforms and each of them try to find the best-fit solution for its needs. This case shows how much the integration of DKMF and GQM covers the approaches of KM represented by the opposite paradigms "K=O" and "K=P". In the e-learning platform evaluation problem the "knowledge as object" can be considered the platform itself and is the primary focus of this case study; whereas the "knowledge as process" can be referred to: the process we use to produces the formative act, the process we use to evaluate the contents, and the process we use to check the achievement of educational goals. Note that the GQM itself employed for the assessment of e-learning platforms can be considered as a process that can be measured. The evaluation of e-learning platforms requires an assessment of the evaluation scenario. The scenario can be specified clarifying some key points normally related to the competencies and the knowledge of evaluators as well as to the context of educational acts. For example, technical and organizational impacts of implementing a software package in a specific organization (or network of organizations) is a critical factor that might undermine the successful introduction of e-learning platforms into learning organizations. Keeping in mind the scenario it is possible to identify several indicators to measure and to quantify characteristics of e-learning platforms to evaluate globally how much it fits the kind of formative process we need.

Table 1 exemplifies how the activities 2 and 3 for the integration of GQM within a DKMF can be pursued. For each goal, its description, the K=O characteristic, and who inserted it are identified and published into the Virtual Knowledge Repository of a DKMK instance.

Then, the "evaluation effort to overcome the knowledge gap", in this case study the comparative evaluation of e-learning platforms, is assessed and published in the Repository.

The state of the GQM process aiming at the evaluation of a given characteristics is also shown (approved, proposed, pending, etc). The columns "Measure details" and "Results" have pointers to documents that describe, respectively: a) the set of measures that are connected to the questions that, in their turn are relied to goals; b) the results of their evaluation (see table 2 for the results about the comparative evaluation of this case study).

TABLE I. GOAL STRUCTURE AND EFFORT ASSESSMENT FOR THE OF E-LEARNING PLATFORM EVALUATION.

Goal	Description	K=O	Inserted by	K=P Learning and/or Design Effort	Process state	Measure details	Used by	Results
G	SCORM compliance							table 2
G_1	General SCORM compliance	SCORM standard	$Actor_1$	3 month/men prestudy 3 month/men identification 7 g/u GQM plan 7 g/u measurement plan 1 g/u collection data 3 g/u analysys data 1 h /u packaging	Approved			
G_{11}	Benchmark compliance		$Actor_2$			mset 1,	$Actor_6$ $Actor_7$	
G_{12}	compliance details		$Actor_2$		Proposal	mset 2	$Actor_3$ $Actor_4$	
G_2	SCORM packages managing	SCORM workload definition		2 month/men prestudy 1 month/men identification 4 g/u GQM plan 9 g/u measurement plan 2 g/u collection data 5 g/u analysys data 2 h /u packaging				
G_{21}	SCORM packages managing analysis		$Actor_4$			mset 3		
G_{31}	SCORM creation & searching utilities analysis		$Actor_5$			mset 4		
…								

It is worthwhile to observe that the "initial" information is stored in the Virtual Knowledge Repository by the organization that assumes the role of leader for the application domain. However, when other organizations wish to contribute, they are enabled to insert K=O and K=P in the repository repeating the steps for goal identification, structuring and sharing above mentioned. After the knowledge insertion in the Repository an organizations relies its identity to the chunk of knowledge inserted (column: Inserted by) other organizations can declare the use of that knowledge (column: Used by) so that trust can be placed on the published contents.

The results have been structured according to the cluster concept, useful to "slice" the platform and focusing on a specific part of it; this choice allows the comparison between the analyzed platforms considering high level characteristics.

Table 2 shows the scores got by the platform in each cluster of goals including the SCORM cluster shown above. All the data are normalized to 10.

Experiences are encouraging so far, mainly thanks to in-the-field data gathering campaigns, regardless the intrinsic complexity of the target problem. The actual GQM plan for an e-learning platform includes so far 48 goals, 79 quality focuses, 495 questions and 91 metrics. This experience can be shared in a KN but also enriched from the others who are interested in enlarge the GQM evaluation and don't want reinvent the wheel by using as is the GQM process only adding the Quality Focus they are interested in reducing the effort and the cost of the process itself.

TABLE II. GOAL SCORES GOT BY THE PLATFORMS MOODLE, DOKEOS AND A-TUTOR.

	Moodle	*Dokeos*	*A-Tutor*
Installation and Standards Compliance	9.12	8.25	7.75
SCORM Compliance	4.07	9.17	7.98
General Characteristics	8.5	7.25	6.375
Content Management	7.11	7.2	7.33
Courses and Users Management	6.33	6.14	5.83
Communication Tools	7.17	8.88	7.1
Learning Tools	8.33	4.66	4.83
Co-working tools	8.87	5.1	0
Tracking e Reports	8.02	8.5	4.5
Help and Documentation	9.5	8	5.5
ISO 9216	9.18	8.17	8.17

GQM process can be conducted using iterative–enhancement with respect to the whole process or the single phase. As an example, during the definition phase for a metrics we can refine the relative question using the packaging experiences that the Knowledge Network can offer.

V. CASE STUDY 2: REDUCING DEPENDENCY FROM FOSSIL FUEL

Many are the motivations behind the exchange, dissemination and sharing of knowledge. The scientific progress, the cooperation between public agencies involved in the management of a common process, the temporary alliance made of a set of enterprises that participate to a public competition for the realization of a project are just examples of such motivations.

In this section we discuss a possible scenario in which some universities cooperate in order to exchange knowledge concerning the ideation, the design and realization of systems whose purpose is the reduction of dependency from fossile fuel.

We resume below a case study arising from the incentive that are stated by Italian laws in the field of energy savings. The focus of this case study is on K=P.

The law "DM 19 Feb. 2007" states the criteria and modalities to incentivize the production of electrical energy from photovoltaic systems can be interpreted as a K=O that is available in the Virtual Repository. On the other side, the K=P can be either prescribed because already appearing in, laws, rules, ecc., or part of the "process experience" that a VC participant is available to share with other participants. Examples of K=P are:

a) "access procedure to the incentive schema for energy saving" (art.5 del d.lgs 19 feb. 2007);

b) BPMN Models for the management of: proposal, decision processes, public competitions, project monitoring, ecc.

c) Table 3 shows the output of activities 2 and 3 applied during the integration of a GQM within a DKMF customized to the case study.

TABLE III. GOAL STRUCTURE AND EFFORT ASSESSMENT FOR THE REDUCTION OF ENERGY NEEDS.

Goal	Description	K=O	Inserted by	Learning and/or Design Effort	K=P (Process Models)	Process duration	Measure details	Inserted by
G	Reduction of Energy needs							
G_1	Photovoltaic	D.lgs 19/2/ 2007 …	$Actor_1$	3 d. legal aspects 7 d. building integration: architect 15 d. system design: electrical engineer				
G_{11}	Realization of 1MW plant	design document		1 months	Proposal	2 days: clerk	$mset_1$	$Actor_4$
					Approval	1/2 day: director	$mset_2$	$Actor_2$
					Competition Management	3 months: legal office	$mset_3$	$Actor_5$
					Project Realization	8 months: competition winner	$mset_4$	$Actor_6$
…								
G_2	Microgeneration	…	$Actor_2$					
G_{21}								
…								
G_3	Green building	…	$Actor_3$					
…								

The general goal is divided in other subgoals (Photovoltaic, Microgeneration, ecc.) further divided in subgoals in their turn. The advantage of the tailoring phase of a DKMF is twofold. On the one hand, it circumscribes the context of activities such as creation, exchange and knowledge sharing; on the other hand, the models for the knowledge representation and the relied artifact can be shared also in the preliminary ideation activity. After the social network is built, all LKMs, including the supervisor, can take advantage of experiences done by other LKMs. For example, design and realization experiences can be reused easing the reaching of common goals and the assessment concerning the aspects K=O and K=P provide useful references about the effort necessary to acquire knowledge or the realize systems.

The model below shows how a BMPN process for the procurement of goods and services can be used to assess a process managed by a Public Agency.

For example, considering the goal of "acquisition of a new photovoltaic system", the "call for tender" process owner could decide to measure the contribution of the activity "tender evaluation" to the whole achievement process.

According to the GQM method, the following are examples of meaningful question concerning the K=P paradigm with respect to goals oriented to the reduction of energy needs:

G1. What is the medium process duration for the realization of a photovoltaic system of 1MW?
G2. What is the time contribution, in percentage terms, of the design activity with respect to the whole acquisition process?
G3. What is the contribution, in terms of workload, of each participant to the acquisition process?

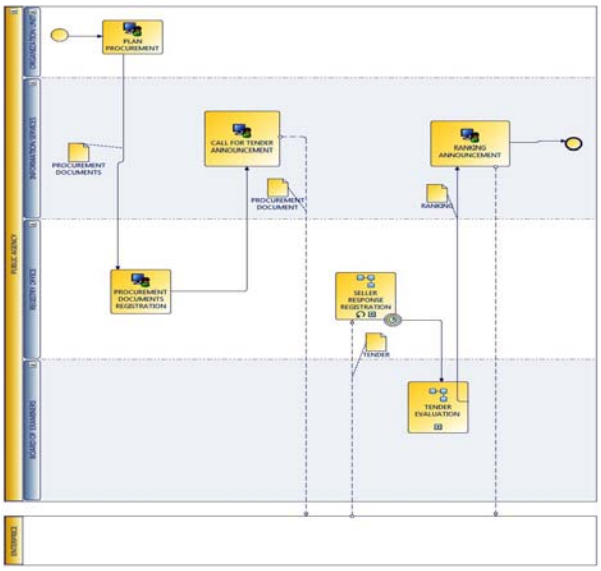


Figure 3. BPMN model for the "call for tender" process

The following measure can be adopted to find the contribution of an activity to the execution of its enclosing process. The measure could be, for example, be used to answer the question G2.

$$activityTimeContribution(activity, process) = \frac{duration(activity)}{duration(process)} * 100$$

VI. CONCLUSIONS

The motivation that drive modern organization to share their knowledge are manifold. On the one hand, public research agency have the mission to share, transfer, and disseminate knowledge at the large extent possible in order to pursue social growth and welfare. On the other hand, enterprises are sometime interested to share knowledge either when a joint project is designed and realized by an alliance of enterprises or when open knowledge is perceived as a value that can lead to new opportunities.
Since the distributed knowledge management can increases the value of knowledge asset of organizations, we believe that the demand for guidelines aiming at the assessment of quality and quantity of knowledge within a network of organizations will grow in the future.
The work presented here combines the DKMF with the GQM model in order to provide an extended model and guidelines useful to assess knowledge concerning the aspects K=O and K=P in a distributed knowledge network. Although the experience done so far is limited to universities contexts, the results are encouraging. Besides the two case studies described in the paper, a third experience has just begun by the authors

of this paper and deals with the development of cooperation in an open environment where students belonging to different universities have to cooperate and increase their knowledge on common issues concerning software engineering projects. A pilot project (enhancing team cooperation using rational software tools in software engineering academic projects - ETC), sponsored by IBM, is in progress. Students' competences have been collected in input by means of questionnaires and check lists and the dynamics of cooperation between students are observed and traced. At the end of the ETC project we expect that data to be collected in output will provide interesting prospects for the development of this work.

Several roads will be explored in future research. One is the refinement of the integrated model DKMF+GQM and the development of tools that support data collection, metric definition and evaluation. This is a complex task because a plan to measure a K=O and K=P objects requires to consider qualitative as well as quantitative aspects and many data to collect even for small networks. The other is the understanding of the role that plays identity, negotiation and trust as facilitators of learning processes that receive the support of an ICT system that implement the model described in the paper.

[1] V.R. Basili "Software modelling and measurement: The Goal/Question/Metric paradigm," Technical Report CS-TR-2956, Departement of Computer Science of Maryland, 2002.

[2] A.-L. Barabasi, "Linked. The new Science of Networks," Perseus Publishing, 2002

[3] C. U. Ciborra, "Teams, Markets and Systems: Business Innovation and Information Technology." Cambridge: Cambridge University Press. 0-521-57465-X, 1993

[4] C. U. Ciborra, Notes on Improvisation and Time in Organizations. Accounting, Management and Information Technologies, vol. 9, pp. 77-94, 1999.

[5] F. A. Calabrese, "The Early Pathways: Theory to Practice – a Continuum- in Creating the Discipline of Knowledge Management," Michel Stankowski ed. Elsevier, 2005.

[6] M. Cohen and D.A. Levinthal, "Absorptive capacity: a new perspective on learning and innovation - Tchnology, Organizations, and innovation," Administrative Science Quarterly, March,1990.

[7] T. H. Davenport, S. L. Jarvenpaa, and M. C. Beers, "Improving Knowledge Work Processes." Sloan Management Review, vol. Summer, pp. 53-65, 1996.

[8] T. H. Davenport and L. Prusak, "Information Ecology: Mastering the Information and Knowledge Environment," Oxford University Press, New York, 1997.

[9] B. Fadini, P. Maresca, P. Prinetto, C. Sanghez and G. Santiano, "A GQM Based E-Learning Platform Evaluation," Conference: Proceedings of the Conference: Distributed Multimedia Systems (DMS2005), invited session on E-Learning, 77-82. Banff , Canada, 2005.

[10] B. Fadini, P. Maresca and S. Russo, "Un approccio alla valutazione di piattaforme per l'e-learning basate su standard di qualità ISO," Proceedings of conference Didamatica 2002: E-Learning metodi, strumenti ed esperienze a confronto 107-132, Napoli, Liguori Press, 2002.

[11] B. Fadini, P. Maresca, P. Prinetto, C. Sanghez and G. Santiano, "GQM for e-learning platform," CINI Internal Report, 2005a.

[12] B. Fadini, P. Maresca, P. Prinetto, C. Sanghez and G. Santiano, "GQM Plan for E-Learning Platform," 2005b.

[13] A. Fuggetta, L. Lavazza, S. Morasca, S. Cinti, G. Oldano and E. Orazi, [Applying GQM in an industrial software factory. ACM Transactions on Software Engineering and Methodology, 1998.

[14] B. Holldobler and Wilson E.O "The Ants Belknap," Press of Harvard Univ.Press, Cambridge, Mass., 1990.

[15] P. Maresca, G. Santiano, B. Fadini and P. Prinetto, "Validation Criteria for a GQM Plan applied to e-learning platform evaluation," Journal of e-learning and knowledge society, 3 (2) , 413-419, 2006.

[16] I. Nonaka and T. Nishiguchi, ed. "Knowledge Emergence: Social, Technical, and Evolutionary Dimensions of Knowledge Creation," Oxford University Press, Oxford, 2001.

[17] T.B. Riley, "Knowledge management. An Evolving Discipline," A paper for National Research Council Canada, Riley Information Services Inc, Ottawa, 2000.

[18] R. D. Stacey, "The Emergence of Knowledge in Organizations," Emergence, vol. 2, no. 4, pp. 23-39, 2000.

[19] L. A. Suchman, "Plans and Situated Actions: The Problem of Human-Machine Communication," University Press, Cambridge, 1987.

[20] G.. von Krogh and J. Roos, "Organizational Epistemology," St. Martin's Press. 0-312-12498-8, New York , 1995.

[21] K. E. Weick, "Organization Redesign as Improvisation," Organizational Change and Redesign, ed. G. P. Huber and W. H. Glick, Oxford University Press, New York, 1993.

[22] L. J. Priestley, "Knowledge Transfer within Interorganizational Networks," Encyclopedia of Communities of Practice in Information and knowledge Management, Idea Group Reference, 2006.

[23] J. Roberts, "From Know How to Show How?," Technology Analysis and Strategic Management, 12 (4), 429-443, 2000.

[24] M. A. Stankosky, "Creating the Discipline of Knowledge Management," Elsevier, 2005.

[25] T. Winograd, F. Flores, "Understanding Computers and Cognition," Addison-Wesley, 1986.

[26] V.R. Basili, "Software modelling and measurement: The Goal/Question/Metric paradigm," Technical Report CS-TR-2956, Departement of Computer Science of Maryland, 1992.

[27] V. R. Basili, "Models and Metrics for Software Management and Engineering," IEEE Computer Society Press, IEEE Catalog No. EHO-167-7, p 343, 1980.

[28] S.J.H. Yang, I.Y.L. Chen, J. Huang and T. Fan, "Social Network Supports in Knowledge Sharing," Proceedings of the Second International Conference on Innovative Computing, Informatio and Control, IEEE Computer Society, Washington, DC, USA, 2007.

[29] B. Yao and S. McEvily, "Absorptive capacity and social network: internal ability and external opportunity for product innovation," Management of Innovation and Technology, vol 2, 2000.

[30] M. Yasuhiro and V.R. Basili, "Using the GQM Paradigm to Investigate Influential Factors for Software Process Improvement," The Journal of Systems and Software, 36 (1), 17-32, 1997.

[31] V.R. Basili, "Quantitative Evaluation of Software Methodology," invited paper, National Communications Forum '86 Chicago, IL, 1986.

[32] V.R. Basili, "Applying the Goal/Question/Metric Paradigm in the Experience Factory," 10th Annual CSR (Centre for Software Reliability) Workshop, Application of Software Metrics and Quality Assurance in Industry, Amsterdam, Holand, 1993.

[33] J. Y. Mashiko and V. Basili, "Using the GQM Paradigm to Investigate Influential Factors for Software Process Improvement," Journal of Systems and Software, vol. 36(1): 17-32, January 1997.

[34] G.Franco, P.Maresca and G.Nota, "Negotiation and Trust in a Distributed Knowledge Management Network," Distributed E-Learning Technologies, Boston 2008.

[35] G.Franco, P.Maresca and G.Nota, "Un approccio quantitativo alla condivisione della conoscenza in una rete di organizzazioni cooperanti," MTISD 2008 – Methods, Models and Information Technologies for Decision Support Systems, Università del Salento, Lecce, 18-20, September 2008.

A Model for Teaching Distributed Computing in a Distance-based Educational Environment

Petra le Roux[1], Alta van der Merwe[12], Aurona Gerber[12]
[1]School of Computing, University of South Africa, Pretoria
[2]Meraka Institute, CSIR, Pretoria
lrouxp@unisa.ac.za, alta@meraka.org.za, agerber@csir.co.za

Abstract - Due to the prolific growth in connectivity, the development and implementation of *distributed systems* receives a lot of attention. Several technologies and languages exist for the development and implementation of such distributed systems; however, teaching students in these new technologies remains a challenge. Even though several models for teaching computer programming and teaching programming in a *distance-based educational environment (DEE)* exist, limited literature is available on models for teaching distributed computing in a DEE. Here our research we examine how distributed computing should be taught in a DEE in order to ensure effective and quality learning for students, specifically by investigating both the specific characteristics of distributed systems technologies and the models used for teaching programming in DEE. The required effectiveness and quality should be comparable to those for students exposed to laboratories, as commonly found in residential universities. This led to the identification of the factors that contribute to the success of teaching distributed computing and determine how these factors can be integrated into a proposed distributed systems distance-based teaching model we call the Independent Distributed Learning Model (IDLM).

I. INTRODUCTION

The design and implementation of software is a difficult and expensive process [1]. This process is also complex, even in a homogeneous environment, that is, an environment consisting of a single, stable platform, using a single operating system and a single programming language, usually from a single vendor. The complexity increases immensely if software is developed in a heterogeneous environment in which the vendors of the hardware and system software may differ. Such an environment adds a number of additional complexities to new software development [2, 3, 1, 4]. The new software must run on a *network* of hosts, and these *hosts* may run different *operating systems*. Furthermore, components of the new software will probably have to be integrated into *legacy* systems. Adding to the complexity is the probable use of different *programming languages* for different components of the new software system. All these complexities generally cause an increase in the cost of developing and maintaining software.

To address some of these complexities, [1] proposes that new software needs to be designed and developed as a set of integrating components that can *communicate* across the boundaries of a network, different operating systems and different programming languages. Furthermore, legacy code has to be *wrapped* so as to resemble *components* to ensure that they can be integrated into new systems. These components also have to be *transferable to new environments* to ensure that they can be integrated into various applications.

Distributed computing focuses on the hardware, software and middleware that allows for a collection of autonomous hosts connected through a computer network to coordinate their activities in such a way that users perceive the system as a single, integrated computing facility. Distributed programming and distributed computing frameworks assist in and simplify the design and development of systems that communicate across the boundaries of a network, running on different operating systems and which were written in different programming languages. These frameworks (commonly referred to as *middleware)* offer flexibility and new ways of integrating existing and new technology, as well as new ways of facilitating communication between systems.

Distributed computing frameworks have raised expectations that these highly functional systems can solve the majority of computing problems. Experience, however, has shown that it is challenging to build such distributed applications [5]. Apart from the difficulties and challenges involved, distributed computing is increasingly being used as a basis for the World Wide Web and distributed network-related software developments.

Furthermore, the use of distributed computing frameworks is complex because the use of distributed computing components (plug and play) presents various challenges. Besides the embedded complexities of this subject, the teaching thereof poses various problems to both student and teacher. In a residential institution, students can be exposed to the various elements of a distributed computing environment under laboratory conditions. This might simplify some of the complexity associated with the learning of distributed computing. In a DEE, however, a different approach is needed to ensure that students receive the same quality of teaching and are able to experience the same degree of learning as under laboratory conditions. Teaching distributed computing without laboratory sessions is more challenging and might fail completely if a set of clear guidelines is not available to direct the teaching and learning process.

We are concerned with the question of how distributed computing should be taught in a DEE to ensure effective

learning for students. The required effectiveness should be comparable to those for students exposed to laboratory conditions such as commonly found at residential universities. Therefore, this paper focuses on determining the factors that contribute to the success of teaching distributed computing and subsequently how these factors can be integrated into a proposed model for the teaching and learning of distributed computing in a DEE. We call this model the Independent Distributed Learning Model (IDLM). Section II provides the background by motivating reasons for teaching an undergraduate course of this nature, as well as discussing existing systems and why there is a need for the proposed IDLM. Section III addresses the approach and method followed to define the IDLM. Sections IV and V discuss the architecture, components, functionality and characteristics of the IDLM. Section VI discusses a case study to validate the proposed Independent Distributed Learning Model. A discussion of experiences gained followed in section VII and a conclusion in section VIII.

II. TEACHING DISTRIBUTED COMPUTING

The rationale for teaching distributed computing as a university elective course is highlighted in the final report of the Computing Curricula 2005 project[1] [6] and are twofold: changes in the computing field and the need for advanced programming courses.

Technical changes in computer science are both evolutionary (exponential increase in available computing power) and revolutionary (the rapid growth of networking after the appearance of the World Wide Web). This rapid evolution of the computer science discipline has had a profound effect on computer science education, affecting both content and pedagogy. Computing education is also affected by changes in the *cultural* and *sociological* context in which it occurs. The following changes, for example, have all had an influence on the nature of the educational process: changes in pedagogy enabled by new technologies, the dramatic growth of computing throughout the world, the growing economic influence of computing technology, greater acceptance of computer science as an academic discipline and broadening of the discipline.

Advanced courses in undergraduate studies serve three purposes: to expose students to advanced material beyond the core, to demonstrate applications of fundamental concepts presented in the core courses, and to provide students with a depth of knowledge in at least one subarea of computer science. One of the advanced programming courses included in the CC2005 is distributed computing.

To determine the components of a model that would contribute to the success of teaching and learning distributed computing in a DEE, we explored the state of affairs when teaching and learning *programming languages*, *programming in DEE* and *distributed computing*.

[1] CC2005 is a joint undertaking of the Computer Society of the Institute for Electrical and Electronic Engineers (IEEE-CS) and the Association for Computing Machinery (ACM) to develop curricular guidelines for undergraduate programmes in computing.

A. Teaching and Learning of programming languages

The available models for the teaching and learning of programming languages are mainly aimed at contact or laboratory sessions, as is the practice in residential universities [7,8]. Typical problems and solutions are identified. All research efforts and development of models which aid in the teaching and learning of programming, especially at introductory level, have one purpose: to ease the learning of the programming language by eliminating all complexities that do not directly contribute to the achievement of the learning objectives [6,7]. For this we developed a generic model consisting of two components: a teaching module and a student module (see Figure 1). The *teaching module* consists of an *expertise module*, which presents the domain knowledge that the teacher intends to be mastered by the student. The text that represents the command of the computer programming languages to be taught is maintained in the *tutoring text*, whereas the *semantic rules* contain the structure of commands to be taught. The s*tudent module* comprises a *GUI*, the *tutoring text*, and the *student profile.*

B. Teaching and Learning programming in Distance-based Educational Environments

Systems used in the teaching of programming in a DEE are classified as intelligent tutoring systems (ITS) [9,10]. An ITS for effective teaching and learning in a DEE is built on the client-server model. In its basic form, the client side consists of a *teaching agent* or *agents* and *student agents*. The server provides the content and infrastructure needed to present a course. A *tutoring agent* and the *Internet* are added to the generic model. The tutoring agent represents the knowledge to be taught; it encompasses modelling of the knowledge, and management and coordination of the learning activities. The Internet serves as a vehicle for communication.

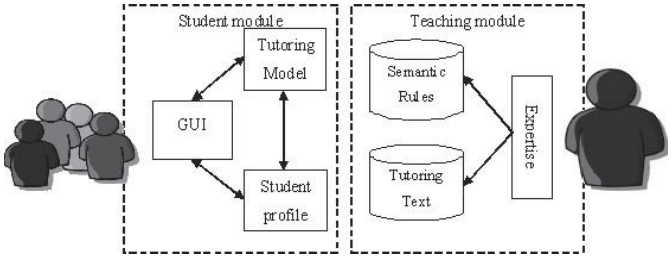


Figure 1. Generic Model for Teaching Programming Languages

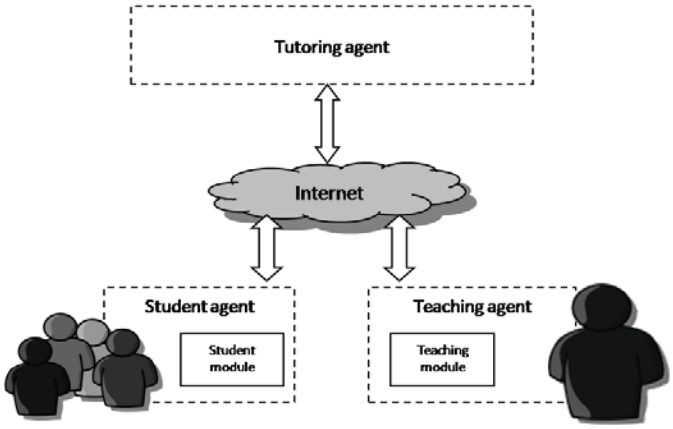


Figure 2. Generic Model for Teaching in a DEE

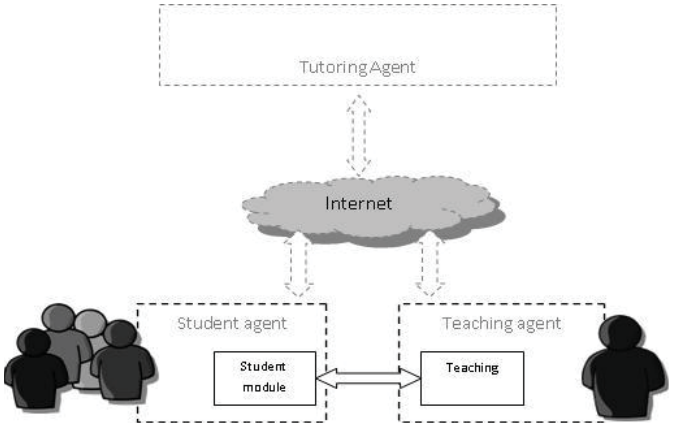


Figure 3. Generic Model for Teaching Distributed Computing

The student and teacher communicate with the tutoring agent through the Internet, or WWW. Thus, the teaching module is contained in a teaching agent, whereas the student module is contained in a student agent. The student agent retrieves the tutoring dialog that a student wants to learn through the Internet. The teaching agent communicates with the tutoring agent to maintain the knowledge to be taught. Figure 2 depicts the updated generic model.

C. Teaching and Learning Distributed Computing

Since distributed algorithms are difficult to grasp and also to implement and debug, models that aid in the demonstration of complex relationships and dynamic processes are identified that have the potential to support teachers and learners [11]. Also, when the Internet is used, use of XML, Java applets, etc. help to ease the complexities, since explicit installation is not needed on the student side; only a Java-enabled browser (available on most computers) is needed. The model consists of two components, a *teaching module* and a *student module*, but the content is presented in a distinctive way. This model is depicted in Figure 3.

D. A Consolidated Approach

When the teaching and learning of distributed systems in a DEE takes place, a number of additional requirements become apparent: (i) a need for a system that is not dependable on a laboratory environment, (ii) a system that addresses the asynchronous and geographically dispersed nature of DEE, and (iii) a system that addresses the economic realities of students. Synchronous communication, broadband and commercially available software may be out of reach for an average student.

The focus is therefore on a model that has the characteristics for effective and quality teaching and learning to take place as well as addressing the above-mentioned challenges and economic realities.

III. METHOD

A three phase approach was used for identification of the proposed distributed computing model. Phase 1 focused on the identification of the elements of distributed computing in order to compile a body of knowledge needed to be present in a model for effective teaching and learning of distributed computing to take place. Phase 2 included an investigation of available models for the teaching of computer programming, programming in a DEE and distributed computing in order to identify success factors for teaching distributed computing in a DEE. Lastly, Phase 3 was used to establish and verify a model for the effective teaching and learning of distributed computing in a DEE.

The research approach used was a *qualitative research* method, which was developed in the social sciences to enable researchers to study social and cultural phenomena; situations in which people and different processes are involved. The research approach employed was *design research,* where the purpose was the creation of an artefact.

The artefact developed was in the form of the independent distributed learning model (IDLM). For data collection and verification of the IDLM, a *case study* approach was followed. The selected case-study environment was a single-case design. A *survey* was also used to gather perceptions on the approach suggested and followed by the researchers.

IV. INDEPENDENT DISTRIBUTED LEARNING MODEL (IDLM)

In order for effective teaching and learning of distributed computing to take place in a DEE, learners need a way to create their own space, where distributed computing concepts can be simulated and studied asynchronously as economic realities do not always allow for synchronous communications. Learners must also be able to work at their own pace and time, with the ability to take full advantage of the synchronous functionalities available in a DEE. These requirements are depicted in Figure 4.

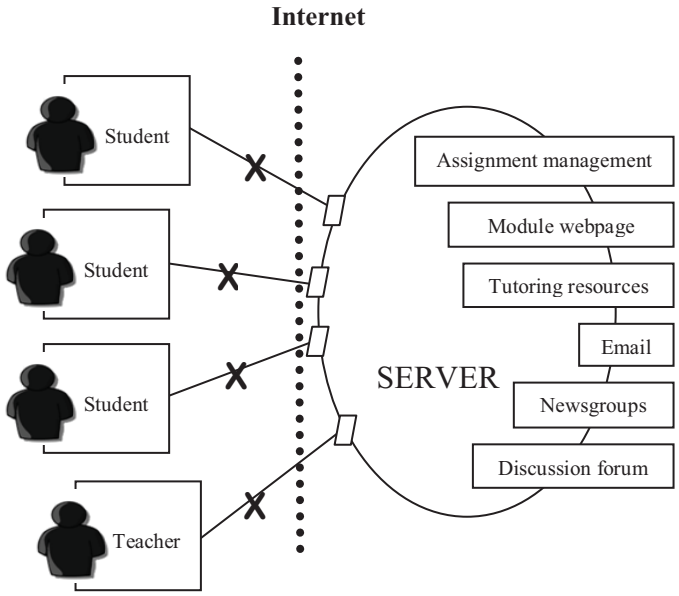


Figure 4. Schematic representation how teaching and learning have to take place in a DEE

Thus, the suggested IDLM consists of different spaces, the resource space, broker space and the learning space (Figure 5). Firstly, the **resource space (RS)** acts as a server and facilitates an environment in which all the resources and functionality needed to accommodate the learning experience reside. These resources are available to the learning space, in which the student and the teacher reside, through the broker space. The **broker space (BS)** acts as a *middleman,* which pairs requests from the learning space with the resource space. The learning space makes its functionality and needs known to the broker space. The main responsibility of the broker space is to identify and match these requests to the resource space. Thus, the broker space is responsible for communication between the resource space and the learning space. The **learning space (LS)** facilitates an environment in which the teacher and student can execute their respective tasks.

Any learning environment requires support infrastructure; so much more an e-learning environment. The infrastructure needed to support e-learning includes inter alia, remote servers, databases and software systems to create the learning space in which teaching and learning can take place. This environment might be synchronous, necessitating continuous remote resource support, whereby the learner and teacher interact in the same time frame with their respective learning or teaching environments and with each other. The environment might also be an asynchronous learning environment in which remote resources are responsible for initial set-up of the learning environment without continuous monitoring and support of this environment. In an asynchronous environment, students work offline for most of their learning session, but might at any time choose to reconnect to the resource space to interact with the resources for tasks such as queries, assignment submission, etc.

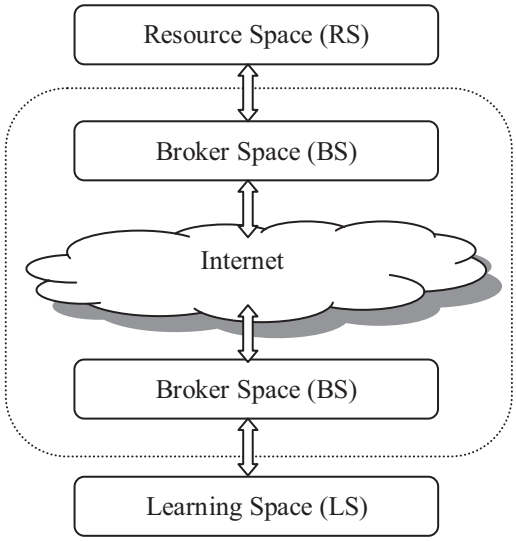


Figure 5. The Independent Distributed Model

A. The Resource Space

The significance of the resource space is that it combines the different resources into an area in which technical support is maintained. Although the infrastructure and the systems within this space might be heterogeneous and distributed as regards specific location, this space consists of the different resource components. Preservation and, consequently, maintenance of a well-defined and specialised space is less complex than when these resources are conceptually scattered throughout every space. The components of the resource space are tutoring resources, assignment management and communication, which includes the course webpage, newsgroups, email and a discussion forum as depicted in Figure 6.

B. The Broker Space

The broker space acts as a *middleman,* which pairs requests from the learning space with the resource space. The learning space makes its functionality and needs known to the broker space. The main responsibility of the broker space is to identify and match these requests to the resource space. Thus, the broker space is responsible for communication between the resource space and the learning space.

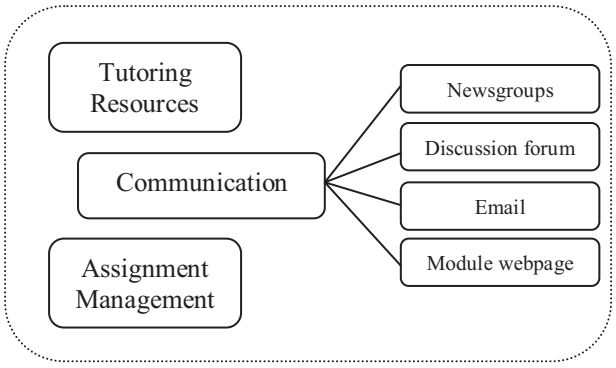


Figure 6. Resource Space

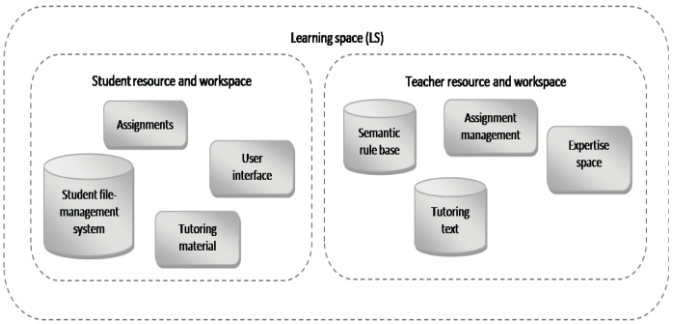


Figure 7. Learning Space

C. Learning Space

The learning space consists of both the s*tudent resource and workspace* and the *teacher resource and workspace* and can be viewed as a facility that allows the student and teacher to fulfil their tasks. The significance of the learning space is that it localises the workspace and resources of both the teacher and the student. The student resource and workspace facilitates an environment in which the student can communicate with the resource space, via a user interface, through the Internet. The components of the student resource and workspace are a user interface, tutoring material and a student file-management system. The teacher resource and workspace facilitates an environment in which the teacher can communicate with the resource space through the Internet. The components of the teacher resource and workspace include an expertise space, semantic rule base, tutoring text and assignment management. See Figure 7.

V. IDLM DETAIL LEVEL

For each of the spaces on the IDLM, more detail is available in [12]. In the specification of the IDLM, UML diagrams were used, more specifically *use-case diagrams,* which model the users and their interactions with the system at a very high level of abstraction, and *activity diagrams,* which are used to describe the workflow behaviour of a system. One example of a use-case diagram of the resource space is depicted in Figure 8.

VI. CASE STUDY: UNIVERSITY OF SOUTH AFRICA

The IDLM was used in a distance educational environment and as guideline in presenting the module to third-year students at the University of South Africa (Unisa). The module is offered by Unisa as part of the undergraduate studies towards a Bachelor in Computer Science or Information Systems. It is an advanced computer science elective module called *Advanced Programming.*

Essentially the software used in the course consisted of a C++ compiler and a CORBA environment. There were a number of permutations of compilers, CORBA ORBs and operating systems which could be used to achieve the objectives of the practical part of the course. However, a balance was struck between allowing students to work in their preferred environment, providing a set of freely distributable tools and providing detailed guidance and support for students and their environments. The prescribed tools were the *Mico ORB*, the *Dev C++ editor*, the *minGW compiler* and the *make facility*. Communication with students took place in the form of *tutorial letters*, the *module webpage*, the d*iscussion forum* and the *module email*.

The student numbers varied from 125 to 200[2], and two lecturers were responsible for the module. The data were collected via questionnaires and were delivered to the students via email, a facility offered by the University's Administrative Department. The number of students registered for the module in question was 247. Seventy-eight (78) students cancelled their studies during the year, leaving a total of 169 students. The number of students who received the questionnaire via email was 151. Thirty (30) of the emails were returned as undelivered mail, possibly because students did not update their personal information via the existing administration channels. Therefore the sample size that received the questionnaire was 121. The number returned was 33, resulting in a response rate of 27%.

After the 1 year period, a survey was conducted on the experiences of the implementation. Questions covered the effectiveness of the teaching methods used, experiences on the usefulness of the tools employed, the effectiveness of the means of communication, suggestions for enhancements and to what extend distributed computing was applicable in the working environment of the student.

Feedback from the students indicated that the methods described in the IDML were found to be helpful by 61%, opposed to 18% of students who found the methods not to be helpful. The number of students who did not have strong feelings about whether the methods contributed to successful learning was 21%.

The data gave strong evidence that the majority of the students made use of the software provided, but found the editor to be least helpful and made use of their own editor. After an initial installation process, which some students found to be cumbersome and difficult, the majority again found the software easy to use.

The results indicated that the discussion forum was found to be most helpful by 91% of students, whereas the module email was found to be least helpful and was used by 24% of students. Tutorial letters, as opposed to the module website, remained the preferred means of communication.

The survey further indicated that the practical component of the module needed more attention, both in respect of tutoring text (additional exercises, more detailed discussions, etc.) and software tools and most of the students did not work in an environment in which the development of distributed systems was applicable.

[2] Enrolment figures from 1997 to 2005.

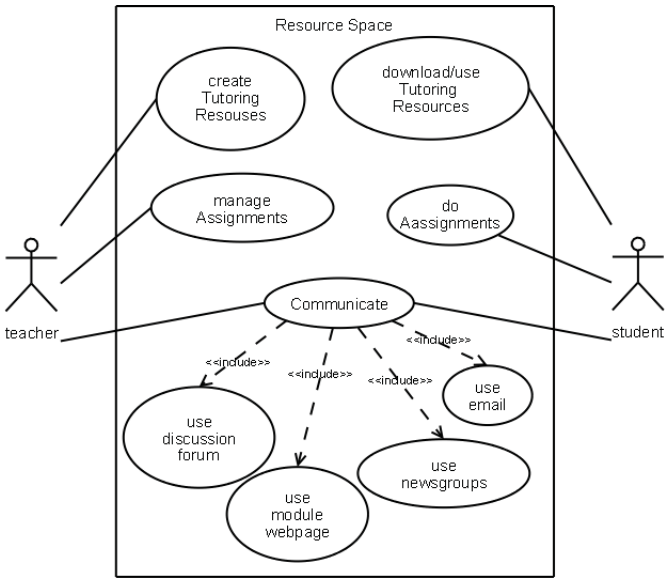


Figure 8. Use-case Diagram – Resource Space

VII. DISCUSSION

This research contributed to the field of teaching a complex subject, namely, distributed computing in a DEE by determining the factors that contribute to the success of teaching distributed computing and subsequently integrating these factors into a model. The Independent Distributed Learning Model (IDLM) is used for the teaching and learning of distributed computing in a DEE. The complexity lies in the fact that presenting this subject matter requires a *specialised developing environment* to develop *specialised softwar*e considering certain *economic realities.*

The IDLM was found to be valuable because it addressed all the challenges experienced. During the case-study in a one-year experiment, the most pertinent challenges included the creation of a specialised development environment. This environment, which was given to the students upon registration, had to be affordable, understandable and easy to use.

The second challenge was to provide supportive material on a regular basis. These supportive materials were presented in various formats, such as tutorial letters, module webpage announcements and forum discussions. The supportive material contributed to the continuity of the module.

The third challenge was to create the student's initial environment and to get the student to write his/her first application. This challenge was due to the various platforms that exist. Since the forum proved to be the single most useful communication medium between students and teacher and students and students, it was used to effectively and efficiently address this challenge. It had to be monitored on a regular basis, especially when assignments or the project was due, which placed a strain on the available resources. However, without the discussion forum, the students felt that the successful completion of distributed computing in a DEE would not have been possible.

VIII. CONCLUSION

The IDLM presented enables teachers within the field of distributed computing and specifically in distanced-based education to present complex subject matter in such a way that students can work *asynchronously* in their own *space*, at their *own time* and *pace*. Therefore, use of IDLM in presenting courses of this nature in this way enabled students to master the complexity.

During use of the suggested model in a one-year experiment, the challenges were the creation of a specialised developing environment, provision of supportive material on a regular basis, initial creation of the student's working environment and the significant value of the forum. All of the above were addressed by the IDLM. Therefore, the use of the proposed model in presenting distributed computing in a distance-based educational environment contributed to the mastering of a advanced and complex course in distributed computing.

References

[1] Condi, S. "Distributed Computing, tomorrow's panacea - An introduction to current technology" *BT Technol Journal* , pp. 13 – 23, April 1999.

[2] Baker, S. *CORBA Distributed Objects Using Orbix.* Addison Wesley Longman, Inc. 1997.

[3] Balen, H. *Distributed Object Architectures with CORBA.* Cambridge University Press. 2001.

[4] Rock-Evans, R. "Component Architectures." *Enterprise Middleware*, 7 – 20, 1999.

[5] Benns, S. B. "What's in the middle?" *BT Technol Journal , 17* (2), pp. 32 – 52, 1999.

[6] IEEE-CS, A. A. *Computing Curricula 2005The Overview Report.* www.acm.org/education/curric_vols/CC**2005**-March06Final.pdf 2005.

[7] Xinoglos, S. "An Integrated Programming Environment for Teaching the Object-Oriented Programming Paradigm." *EurAsia-ICT, LNCS 2510 EurAsia-ICT, LNCS 2510*, pp. 544- 551, 2002.

[8] Kelleher, C. (2005, 37(2)). Lowering the Barriers to Programming: A Taxonomy of Programming Environments and Languages for Novice Programmers. *ACM Computing Surveys* , pp. 83-137.

[9] El-Khouly, M. F. Expert tutoring system for teaching computer programming languages. *Expert Systems with Applications* (18), pp. 27-32. 2002

[10] Hartley, J. S. "Towards More Intelligent Teaching Systems." *International Journal of Man-Machine Studies , 5* (2), pp. 215- 236, 1973.

[11] Schreiner, W. A. " Java Toolkit for teaching Distributed Algorithms. *ItiCSE'02,* 2002.

[12] Le Roux, P. *Towards a model for teaching distributed computing in a distance-based educational environment.* MSc UNISA, 2009.

How do archaeologists work? A multimedia digital object to learn their investigation process

C. Ardito, R. Lanzilotti
Dipartimento di Informatica, Università di Bari, Italy
{ardito, lanzilotti}@di.uniba.it

Abstract - **The aim of the CHAT ("Cultural Heritage fruition & e-learning applications of new Advanced (multimodal) Technologies") research project was to provide on smartphones context-sensitive multimedia services accessible through several interaction modalities. The multimedia content of such services was implemented as Digital Objects managed by the Fedora web service. In this paper, we present a multimedia digital object that, through images, videos, 3D reconstructions, sounds, informs people about the scientific value of the archaeological investigation and about the complexity of the overall archaeological research process. Archaeologists follow a rigorous method, called stratigraphic excavation, to reconstruct the past civilizations lifestyles. People tend to think that the archaeologist work consists only in the excavation and research of valuable elements, while it has its scientific rigor. By exploiting the CHAT infrastructure, the multimedia digital object about the archaeologist work can be accessed in different context on several devices.**

Keywords — **digital objects, cultural heritage**

I. INTRODUCTION

"Paul is a child 9 years old. This year, his school curriculum included the study of prehistory. He learned from the teacher Francesca that the discovery of Otzi, the man who came from the ice, is the work of archaeologists. Paul, intrigued, asked his dad if nearby there is a museum or an archaeological park like what he saw on the history book. Thus, last Sunday, he visited with his dad the Archaeological Park of Egnathia. Ilaria, the park guide, during the path through the ancient Roman city showed him the archaeological finds: the Furnace, the Foro Boario, the Civil Basilica and the Episcopal church, and so on. But Paul is still curious: how did they understand that the Civil Basilica was not the Episcopal church? It seems almost the same! And how can they decide where to dig? Where to look? In the museum there are many finds but how did they understand their origin and use?"

The above scenario shows that archaeologists should let people know their scientific work. This need came out during the work carried out for the CHAT ("Cultural Heritage fruition & e-learning applications of new Advanced (multimodal) Technologies") research project. We developed a software infrastructure for providing services supporting and enhancing visits to cultural sites, such as archaeological parks. We adopted a participatory design approach involving experts in different domains: computer science, history, archeology, cultural heritage, etc. In particular, archeologists highlighted that people tend to think that their work consists in the excavation and in the research of valuable elements, without any scientific rigor. The excavation is only one of the phases of a complex scientific process, called archaeological investigation, which starts with the analysis of how archaeological remains were formed and preserved, and concludes with a historical interpretation of what has been retrieved.

In this paper, we present a multimedia digital object that, through images, videos, 3D reconstructions, sounds, informs people about the scientific value of the archaeological investigation and the complexity of the overall archaeological research process. This work is based on the CHAT software infrastructure, described in the paper presented at the DET 2009 workshop [3].

The paper has the following organization: Section II summarizes the CHAT software infrastructure. Section III illustrates the scientific work of the archaeologists during the excavation process of a site. Section IV presents the digital object describing the work of the archaeologists. Section V closes the paper.

II. THE CHAT INFRASTRUCTURE

The software infrastructure developed for the CHAT research project aims at providing, through a client/server architecture, context-sensitive and personalized services accessible through thin clients

such as cellular phones or PDAs [3]. The dialogue with the user is adapted according to several factors: the service provided, the context in which the user is acting, the task being executed, the user and her/his device. All these factors are managed by a specific software component that constitutes the Adaptive Dialogue Manager (see Fig. 1), which receives input events from different channels (tap or sketch on the screen, voice, gesture, RFID or visual tag scan, etc.) through the MultiModal User Interface (MMUI) Manager of a client. The server components communicate with mobile thin clients through HTTP requests over wireless networks [1]. The Adaptive Dialogue Manager composes the output to be delivered to the user device by retrieving or generating suitable representation of the (possibly multimedia) content and takes care of its delivery and synchronization aspects. In particular, the Adaptive Dialogue Manager builds an XML-like file, which describes the overall interface to be displayed on the user device.

The client application displays the multimedia contents sent by the server and captures user interaction and contextual data through the mobile device carried by the user and/or received from other sensors in the surrounding environment.

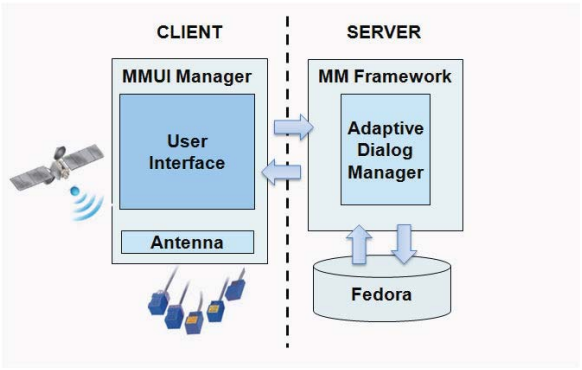


Fig. 1. A simplified representation of the CHAT software infrastructure, depicting the components of interest in this paper.

The multimedia contents are stored in Digital Objects (DOs) managed by an open-source content management software called Fedora (Flexible Extensible Digital Object Repository Architecture) that runs as a web service within an Apache Tomcat web server. Fedora provides tools and interfaces for the creation, "ingest" (a Fedora specific term to indicate the operation of inserting the DO and its own components in the repository), management, and dissemination of the content [6]. Application of different stylesheets to the data and metadata of a Fedora object allows multiple views of the object's content and metadata. Because of this inherent strength and flexibility, it is simple to add new views and data transformations over time as the implementer's and user's requirements change. An additional system component, the Authoring Tool, has been developed to allow end users lacking computer programming expertise to develop multimedia contents for the different services provided.

An example of service provided by the CHAT system is a game-based m-learning service, which supports the visit of middle school students to historical sites, like an archaeological park [3]. The overall information about this service is stored in Fedora through five digital objects (DOs). The first three DOs, namely "SITE", "PLACE" and "CONTEXT", contain information describing the overall site, its places of interest and contextual information (e.g. contextual sounds reproducing real life activities at that site in ancient times). The other two DOs, "SCHEMA" and "INTERFACE", are used to define the game storyboard and some user interface elements.

In the rest of the paper, we present a multimedia DO that has been recently developed and added as a new service in Fedora. It is implemented as a unique DO available in the CHAT repository managed by Fedora. This DO is called "Archaeologists' Work" since it illustrates the work of archaeologists during the archaeological investigation process. While the CHAT infrastructure considered smartphones as presentation devices, people can now access the content of the multimedia DO on different devices: personal computer or large displays installed in the archaeological park; on a mobile device during the visit; at home through the park official web site.

III. THE ARCHAEOLOGISTS' WORK

The idea to create the Archaeologists' Work DO came when we used the m-learning service described in [3] to support the visit of several classes of the middle school "Michelangelo" in Bari, Italy, to the archaeological park of Egnathia, located about 60 Km South of Bari. Students enjoyed a lot playing with the system, which was able to stimulate very much their curiosity about the park and its places of interest. They wanted to know more about how such places were identified, since it is very difficult to figure out their original aspect from the ruins in the park. Unfortunately, archaeologists were not presented during these visits, so that it was not possible to answer to students' questions. As a consequence, we thought it

would be useful to illustrate the archaeologists' work.

The Archaeology studies past human cultures and civilizations and their relationship with the environment, through the collection, documentation and analysis of the traces left during the time. The goal of the archaeologists work is to reconstruct the past civilizations lifestyles, through a rigorous method, called stratigraphic excavation. The archaeologist analyses what is found and determines the relationships among the findings.

In designing the services proposed by the CHAT infrastructure, we collaborated with archaeologists of the Department of Ancient History of the University of Bari. They told us that usually people see the archaeologists as the legendary Indiana Jones, but it is not as an archaeologist works! Thus, we decided to develop a multimedia digital object aiming at learning people (e.g. visitors of archaeological parks) how places and objects they look during their visit have been discovered.

In the early phase of the design, interviews and focus groups involving the archaeologists and experts of the archaeological park of Egnathia were performed to capture information about the overall archaeological investigation process. Three main phases were identified: 1) archaeological site identification, 2) excavation, 3) historical interpretation (see Fig. 2).

The archaeological site identification phase aims at locating the areas where past human civilizations left their tracks. Based on ancient documents and preliminary evidence it is possible to define the areas to look for tracks. The identification of these areas requires the adoption of appropriate techniques, through which the archaeologists are able to read the tracks preserved in various ways in the current landscape and to formulate hypotheses about the nature and appearance of the site is going to dig.

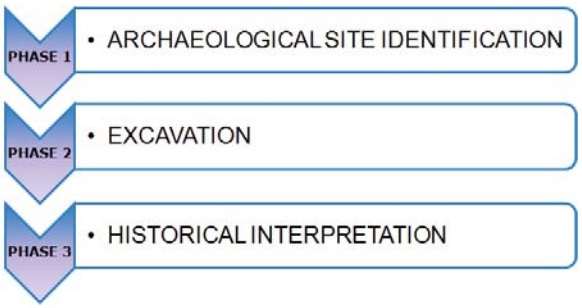


Fig. 2. The three main phases of the archaeological investigation process.

The excavation phase represents the central phase, during which archaeologists collect and organize the unearthed tracks. In full compliance with the rules of the stratigraphic excavation, this task requires the documentation of findings and the production of a stratigraphic chart, which graphically shows the development of an archeological stratification. It highlights each stratigraphic unit found in the area and the relationship among different units.

Finally, during the historical interpretation phase, archaeologists reconstruct the "life" of the site. By adopting specific dating methods, archaeologists can understand the history of the findings by referring them to the culture and habits of the ancient civilizations that have populated such site.

The first prototype of the multimedia digital object was designed as a set of three distinct video documentaries, each one describing a phase of the archaeological investigation process (Fig. 3). A narrator explains how an archaeologist works by commenting the images shown in the video. The key concepts are emphasized by text: this combination of media helps users to focus on the information and to recognize various objects when they see it in the real environment.

The prototype evaluation highlighted some disadvantages. First, the narrative form of documentary makes the product unable to capture and maintain the users' attention high, conversely they are distracted. Another problem emerged from the evaluation was that using three videos, the phases of archaeological investigation process appear disjointed.

In order to make the product more interesting and attractive, a new version of the multimedia digital object, described in Section IV, was implemented.

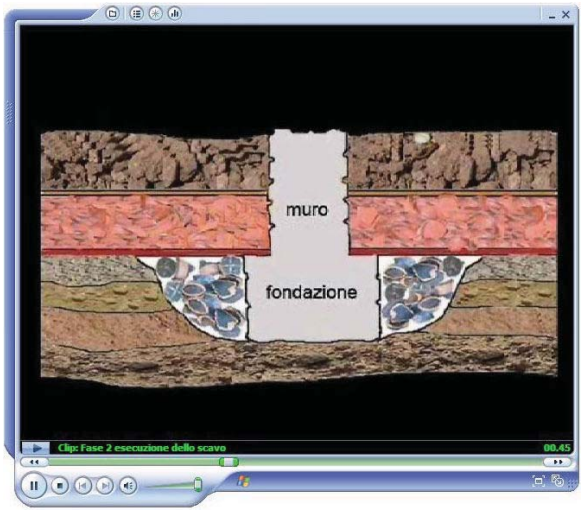


Fig. 3. A video illustrating an activity performed during the excavation phase.

IV. THE "ARCHAEOLOGISTS' WORK" DIGITAL OBJECT

It is well acknowledged the advantage of the animated characters in the learning process 0, especially if a team of characters are involved [4]; thus a dialogue between two avatars has been implemented in order to describe the archaeologists' modus operandi.

The multimedia digital object proposes the dialogue between Gaius and Miranda. Gaius embodies a fictional character of the excursion-game Explore!, through which students visit the archaeological park of Egnathia [2], [5]. Gaius is a citizen of the Roman Empire, who lived in Egnathia in the second century AD, who in 2010 comes back to his city and meets Miranda, a young archaeologist involved in the Egnathia excavations. Miranda drives Gaius through the ruins of the ancient Roman city showing what archaeologists have unearthed and illustrating how the process has occurred.

The purpose of Miranda is to describe the difficulties of the work of archaeologists asking Gaius confirmation of conjectures they have done and helped them to rebuild the city Roman life. Gaius is interested in understanding how archaeologists work and confirms or declines interpretations and reconstructions of the archaeological findings illustrated by Miranda.

The narrative dialogue aims at emphasizing the concepts of the domain from two different viewpoints: that one of Miranda, who is the expert in the archaeological field, and that one of Gaius, who personifies the unaware visitor of their modus operandi. In this perspective, Miranda acts as a teacher, a guide to learning, emulating the role of Virgil in The Divine Comedy.



Fig. 4. Miranda leads Gaius through Egnathia. Their current position is shown on the map of the ancient city.

Miranda leads Gaius in the city through various places. The characters move on the virtual map of Egnathia shown behind them. In this way, the user focuses on the places and remembers their names and physical locations. In Fig. 4, the characters are located near the Furnace, whose plan is highlighted in yellow.

The dialogue between Gaius and Miranda is implemented as an animation sequence divided into ten scenes in the archaeological park of Egnathia. Each scene represents a place: the arrival of Gaius to Egnathia and the meeting with Miranda, the Furnace, the Trajan Way, the Civil Basilica, the Temple, the Foro Boario, the shop, the residential area, the Episcopal Basilica and the ending.



Fig. 5. Miranda and Gaius discuss about the functionalities of the places retrieved in Egnathia.

During each scene, the characters provide historical notions by illustrating the characteristics of the places and their intended purposes (Fig. 5). Miranda reports the discoveries the archaeologists have done and Gaius tells the stories that join him to that places, thus validating or adjusting Miranda's conjectures.

Then, behind the characters a 3D reconstruction of the visited place is shown (Fig. 6) and contextual sounds reproducing the noises of the ancient daily activities are played e.g., carts running on the paved Trajan Way, lawsuit announcements in the Civil Basilica, cows lowing in the Foro Boario, etc.

When the dialogue between Miranda and Gaius about a specific place is over, they reach the next one. Again their movement is shown on the map.

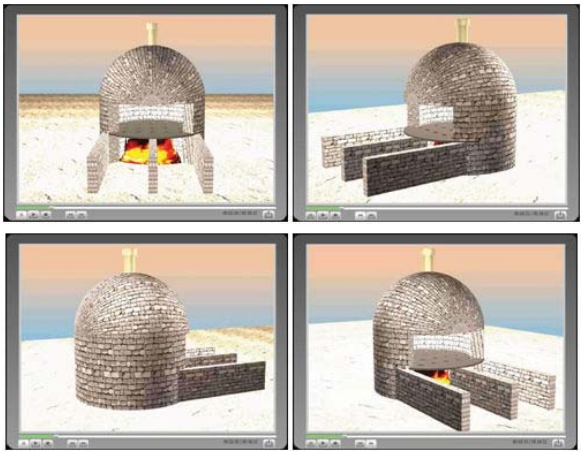


Fig. 6. During the dialog between Miranda and Gaius, a 3D model of the visited place is displayed behind them.

During the path between some of the places, Miranda explains to Gaius the archaeologists research method: when she needs to illustrate more complex concepts she exploits snippets of the video documentaries realized for the first prototype (Fig. 7).



Fig. 7. Short documentaries help Miranda in explaining some concepts related to her work.

V. Conclusions

The work presented in this paper continued the work carried out during the CHAT research project. Together with archaeologists, we created a new service to illustrate the scientific value of the archaeologists' work. This service is implemented as a digital object, called "Archaeologists' Work"; it exploits several types of multimedia contents (i.e., images, videos, 3D reconstructions, sounds).

This multimedia digital object was firstly designed as a set of videos in which a narrator explains how an archaeologist works. A first formative evaluation highlighted that it was almost boring and not able to capture the user attention; furthermore, the phases of archaeological investigation process appear disjointed. To make the product more interesting and attractive, a the new version of the digital object implemented a dialogue between two avatars has been developed. This approach emphasized the presented concepts from two different viewpoints: that one of Miranda, who is the expert in the archaeological field, and that one of Gaius, who personifies the unaware visitor.

In the CHAT infrastructure the content of the various services is organized as DOs. This allows us to reuse the multimedia resources created for other CHAT services within the cultural heritage domain, by combining them with new content in order to create new DOs. In particular, in the Archaeologists' Work DO, we have reused the map of the archaeological park of Egnathia, the 3D reconstructions of important places, the contextual sounds, which are the more expensive resources to be created.

Acknowledgements

Partial support for this research was provided by Italian MIUR (grants CHAT and L4ALL). We are indebted with all people that were involved in the participatory design of the developed cultural heritage applications and, in particular, to archaeologists Raffaella Cassano and Silvio Fioriello, and the director of the Egnathia Archeological Park. We thank Professor Maria Francesca Costabile for her valuable guidance in our research.

References

[1] C. Ardito, P. Buono, M.F. Costabile, R. Lanzilotti, T. Pederson. Re-experiencing History in Archaeological Parks by Playing a Mobile Augmented Reality Game. *In LNCS 4805*, 2007, pp. 357-366. Berlin, Springer.

[2] C. Ardito, P. Buono, M.F. Costabile, R. Lanzilotti, T. Pederson, A. Piccinno. Experiencing the past through the senses: an m-learning game at archaeological parks. IEEE MultiMedia, 2008. pp. 76-81.

[3] C. Ardito, R. Lanzilotti. Organizing the multimedia content of an m-learning service through Fedora digital objects. *In Proc. of DET 2009*, pp. 228 – 233.

[4] G.H. Bower, et al. Hierarchical retrieval schemes in recall of categorized word lists. Journal of Verbal Learning and Verbal Behavior 8, 1969, pp. 323–343.

[5] M.F. Costabile, A. De Angeli, R. Lanzilotti, C. Ardito, P. Buono, Pederson T. Explore! Possibilities and challenges of mobile learning. *In Proc. of CHI 2008*, pp. 145-154.

[6] Fedora web-site. http://www.fedora-commons.org/

[7] R. Moreno, R. Mayer, J. Lester. The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedagogical Agents? Cognition and Instruction, 19(2), 2001, pp. 177-213.

An Adaptive Test for Learning Objects: item calibration

P. Di Bitonto, M. Laterza, T. Roselli, V. Rossano
Department of Computer Science
University of Bari "Aldo Moro"
Via Orabona, 4 – 70126 Bari, Italy
{dibitonto, marialaterza, roselli, rossano}@di.uniba.it

Abstract— **Learning Objects differ from the usual digital resources because they include both learning contents and assessment.**

In this paper we propose a methodology that can make an adaptive assessment of the knowledge/ability level achieved by the learner, and is thus less boring for primary school pupils than most of those produced up to now. The methodology, based on the ideas of Adaptive Systems, Computer Adaptive Test and, Item Response Theory allow tests to be created that can adapt the difficulty level of the questions to the individual learner.

The preliminary phase in building a test using the Item Response Theory is the calibration of the items in the bank item. This consists of assigning a level of difficulty to each item to be proposed to the student. Different psychometric models are used to scale the items. Usually, the IRT is employed because it uses the same metric to measure both the item difficulty and the student's level of ability, so the two measures are comparable.

Usually, the main problem in building a quiz game is the choice of what question to pose to the student. The CAT theory allows automatic selection of the question to ask, according to both its difficulty and the individual student's abilities.

Keywords-component: Learning Objects; Game Based Learning; Computer Adaptive Test; Item Response Theory.

I. INTRODUCTION

Information and Communication Technology (ICT) has led to a revolution in learning and teaching methods. Thanks to the use of the Internet, nowadays it is possible to integrate traditional educational processes with the newest technological innovations. In this scenario Learning Objects (LOs) have played an important role in the definition of digital resources. Since 2000, e-learning research efforts have been focused on the concept of Learning Objects. In depth discussion about them has led to several definitions [2, 12]. The results of this process have been identification of the main LO characteristics (reusability, autoconsistency, interoperability, accessibility) [4, 9, 10] and the main LO components (learning content and assessment). In the e-learning literature a lot of research has been focused on the effectiveness of the learning content, and some of the studies offer guidelines to build LOs. But there are few works about the effectiveness of comprehension tests, that are a key LO component. An interesting theory in the field of testing is the Computerized Adaptive Testing theory (CAT), that increases the statistical accuracy of test scoring by selecting the best item (question) for the examinee's ability level [13].

Our research aims at building an adaptive test addressed to pupils of primary schools. It is based on Game Based Learning and the Item Response Theory. Both methodologies allow to comprehension tests to be built that are less boring for the learner thanks to selecting the question difficulty level according to the learner's abilities. The first phase of the IRT process is the calibration of the bank item, that is the definition of the difficulty level for each question. This paper presents the calibration process of the bank items of the GreEnergy quiz.

The paper is structured as follows: section II illustrates the background theory of the research work; section III describes the Item Response Theory and the calibration process that is used to build the adaptive test presented in the paper; section IV describes the quiz and the construction of bank items. Finally, some conclusions and future research directions are proposed.

II. THEORETICAL BACKGROUND

The core issue of our research is to define a methodology to build adaptive tests that can be included in LOs. Building an effective test is a complicated process. Usually comprehension tests are very boring for learners. The Game Based Learning approach is based on the idea of using games in learning processes. A quiz game could be a very interesting way to propose an assessment test.

Moreover, the main difficulty for instructors is to define the right level of questions, that allow the learner's learning gain to be accurately measured. Often, in fact, tests (or part of them) can be very difficult for learners with low abilities or very easy for learners with high abilities. A good compromise is to build an adaptive test, using the CAT theory, that is able to choose the right question for the individual learner's ability, and thus the right score for it.

A. Game based learning

Game Based Learning (GBL) aims at introducing the benefits of playing games, especially video games, in educational processes. The general idea of GBL software applications, also named "Serious Games", is to exploit the

game properties to improve students' learning outcomes in educational environments. Nowadays, one of the most commonly used approaches to build GBL applications is fully immersive environments, or augmented reality (see [1] just to make an example). These environments allow students to act in a virtual world like they would in a real one, thanks to the 3D graphics capabilities, thus enriching their learning experience. Moreover, the use of GBL makes the learning process less formal and more pleasant [7, 8] At the same time, simpler quiz games are also widely used to support learning in higher and further education contexts [11].

Therefore, the starting point of our work is to exploit the potential of playing games to introduce a more engaging evaluation in Los, that can increase the motivation of learners, offer another way to interact, and make learning fun.

B. Computerized Adaptive Testing

Computerized Adaptive Tests (CATs) [3, 13] are tests that are able to adapt the difficulty level of the questions to the student's ability level. In particular, a CAT selects the question to be posed from a set of possible questions, according to the answers provided by the student during the evaluation process.

Standard traditional fixed tests pose the same number of questions to all users, but not all of these are necessary to assess their learning; the questions may be too easy or, on the contrary, too hard and so the test result may provide little information about the student's real learning gain.

Instead, a CAT selects only those questions that are necessary to assess the student correctly, because it is able to choose the level of the question according to the student's level of ability. In other words, a CAT poses easier questions to low ability learners and harder questions to high ability learners. The scoring is calculated on the basis of the learner's ability and the item difficulty. For example, if two students answer the same number of questions, the one who answered the hardest questions will have the highest score. Thanks to the use of CATs it is possible both to reduce the time required to answer the test, by dynamically selecting the question to be posed, and to maximize the precision of the learner's evaluation. For these reasons a CAT could be less boring for students, because high ability learners do not have to answer useless questions (too easy for her/his ability) and, on the contrary, low ability learners do not have to try to answer too difficult ones.

To build a CAT requires a prior classification of the items and the learner abilities. The psychometric literature proposes different approaches, the most widely used of which is the Item Response Theory (IRT).

III. Item Response Theory

Item Response Theory (IRT) [6] is a paradigm adopted to build a CAT, that measures the student's ability level, the difficulty level of items, the test score, etc. It overcomes the Classical Test Theory (CTT) limits. In a CTT test there is a set number of questions and the final score is computed on the basis of the correct answers given. Instead, in the IRT test the number and difficulty of the questions is evaluated in run-time and the test score is calculated on the basis of these parameters.

A. Main IRT models

The IRT methodology uses a statistical model to measure the learner's ability level and some item characteristics. In particular, it computes the test score on the basis of latent traits (denoted by **θ**) that represent the learner's ability, and some item parameters such as difficulty (denoted by **b**), discrimination (denoted by **a**), and guessing (denoted by **c**). There are several IRT models, that differ in terms of the type of mathematical function and the number of parameters considered: 1PL models that employ only one parameter (difficulty), 2PL models that employ two parameters (difficulty and discrimination), and 3PL models that use three (difficulty, discrimination and guessing). During the design phase of GreEnergy we used this last one because it is more complete.

B. 3-parameter logistic model

In the 3-parameter model the relation between the user's ability and the likelihood that s/he will give the right answer is expressed by means of a function, defined for each item rather than for the whole test. In other words, the 3PL model measures, for each item, the probability that a user will give the correct answer. The probability varies according to the user's ability θ as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1,2,\ldots k$$

where parameters a_i, b_i and c_i are the discrimination, difficulty and level of guessing of item i, respectively. In order to build an Item bank from which the CAT can select the questions to be posed, it is necessary to evaluate the actual relation between the item difficulty and user's ability. For this reason a calibration phase is necessary to define the right b parameter for each item.

The values of the item parameters are calculated using statistical software that, according to a set of answers given by a sample of examinees, calculates the *a, b,* and *c* parameters of each single item. The $P_i(\theta)$ values computed for each item i, with changing θ, represent the Item Characteristic Curve (ICC) that allows the relation between the user's latent ability and the item difficulty level to be defined.

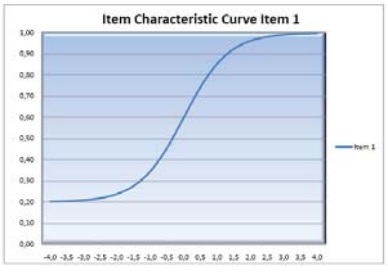


Figure 1. The Item Characteristic Curve

Usually, it is assumed that the user ability range is [-4, 4]. In figure 1 the ICC for the item shows that an examinee with a

very low ability (-4) has a 20% likelihood of giving the correct answer because of the guessing parameter, while for an examinee with a high level of ability (>2) the probability that s/he will give the right answer is about 90%. This means that for users with a low ability the item is too difficult, on the contrary for users with a high ability the item is too easy. The difficulty of item *i* is the point where $P_i(\theta)$ is equal to (1+c)/2. In this case b=0.0, that is the value of θ *where* $P_i(\theta)$= (1+0.20)/2=0.60.

After defining the item parameters, it is possible to define the Item Informative Functions (IIF). This expresses the precision according to which the item measures the learner's latent ability, in other words the IIF allows measurement of the item quantity of information for each user's ability. Formally, the function of IIF is:

$$I_i(\theta) = a^2 \left[\frac{Q_i(\theta)}{P_i(\theta)} \right] \left[\frac{P_i(\theta) - c^2}{(1 - c^2)} \right] \qquad i = 1,2, \ldots k$$

where $I_i(\theta)$ is the IIF value of item i, a_i is the discrimination value of item i and c is the guessing parameter, $Q_i(\theta)$ is the probability that a user with θ level of ability will give the wrong answer to item i.

Figure 2 shows the IIF graph for three different items. Item 1 is meaningful for a wide range of user abilities (from -2.5 to 2.5) and the maximum is reached at θ = 0.5. The IIF values are used to create the item bank. The IIF values that have too high maximum θ values (around +4) or too low (around -4) are discarded because these items are too difficult or too easy.

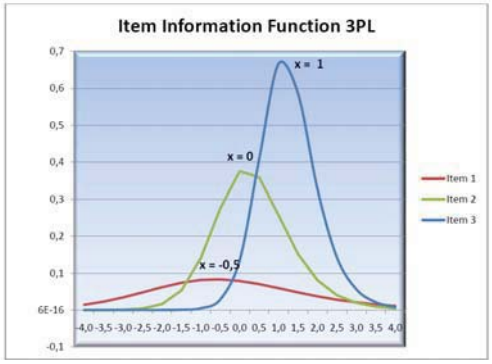


Figure 2. The Item Information Function in a 3 parameter model

The IIF values are used to compute the test score, that is very different from a traditional score. It will be a weighted score obtained using the IIF and the discrimination parameters. In each CAT the scoring is defined according to the specific goals.

IV. THE GREENERGY QUIZ

The GreEnergy Quiz is based on the idea of integrating Game Based Learning (GBL) and Computerized Adaptive Testing (CAT) theories. All primary school teachers know that quiz game activities are good tools for measuring students' learning gain. The interaction with a computer allows the student feel more relaxed and even more concentrated on the questions.

A. The Domain

The learning purpose of GreEnergy Quiz is to foster cultural renewal based on the principles of sustainability and to induce the changes in young people's behavior needed to face the challenges posed by the global era. Environmental education is one of the most recent areas of educational studies and practices. It has been introduced as a result of the increase in phenomena causing environmental disasters (the latest one was the Gulf of Mexico oil spill - 29th April 2010). For this reason, environmental education has been focused for many years on ways to defend the natural environment or endangered species. The introduction of this subject in primary schools aims at improving young people's sensitivity to the Earth problems and helping them to learn to live and act more responsibly than their forebears.

B. Users and Questions

The Greenergy Quiz is addressed to pupils attending third/fourth-grade at the primary school. In order to ensure that the questions, as well as the terminology, were suitable for the pupils' scholastic level, a primary school teacher and a domain expert helped us to define the quiz content. In accordance with their suggestions three tests were defined: how to save energy; photovoltaic, solar and thermal power; and wind power. Each test was composed of multiple choice items each offering three possible answers.

In CATs the number of questions is not usually pre-defined. It depends on the user's answers and interaction. Moreover, the adaptation algorithm allows the system to choose the question to pose according to the student's level of ability and the level of difficulty of the question. For this reason, CAT requires fewer items to assess a student than a traditional multiple choice test.

The GreEnergy quiz described in this paper is a preliminary, non-adaptive version realised in order to support the calibration process. For this reason the number (15) of items proposed in each test was pre-defined as well as the type (difficulty).



Figure 3. Screenshot of GreEnergy Quiz: a) the welcome/login screen, b) a test page

C. User Interaction

The interaction has been designed to be as simple as possible and, at the same time, attractive to young users. The

first interaction requires the user to insert Name and Surname, in order to keep track of the results (Figure 3 – a). Then the user can start the test. For each question the system presents three possible answers and the related buttons to submit the answer (Figure 3 - b), after which feedback is supplied. The feedback contains information about the good or bad answer and some clues to the question topic. After the test an overall feedback is presented to the user.

D. Technology issues

In order to build an attractive quiz for young users, Adobe Flash ® was used to build the interface. Moreover, in order to make the quiz scalable and adaptable to any domain the xml was used to memorise the item bank. The data organisation is depicted in figure 4. Each question is defined using the tag `<Item>` and is identified by an id attribute; the child `<Domanda>` contains the text of the question to be posed; the tags `<RispostaN>` contain the three answers, for each of them the attribute id defines whether the answer is right or wrong; the tag `<Sfondo>` defines the image to use as screen background. In this way, it is possible to adapt the screen layout according to the domain question; the tag `<Difficolta>` contains the difficulty of the question, in our case the range is 1 (easy) to 3 (difficult); the tag `<Approfondimento>` contains the feedback text to be presented after the user's answer.

```
<Items>
        <Item id="0">
                <Domanda>Quali sono le fonti di energia rinnovabili?</Domanda>
                <Risposta1 id="sbagliata">La luna</Risposta1>
                <Risposta2 id="corretta">Il sole, il vento e l'acqua</Risposta2>
                <Risposta3 id="sbagliata">Il petrolio</Risposta3>
                <Sfondo>imgRisp</Sfondo>
                <Difficolta>3</Difficolta>
                <Approfondimento>Testo Approfondimento</Approfondimento>
        </Item>

<Items>
```

Figure 4. XML structure for the questions (Item)

V. ITEM CALIBRATION

As stated above, one of the main steps in Item Response Theory is the scaling of items. The purpose of this process is to scale the difficulty of the items and the learner's ability on the same metric, using different IRT models,. In this way the difficulty of an item and the ability of a person can be meaningfully compared. In designing and implementing an adaptive test, therefore, it is important to measure the difficulty of the items, scaling them in order to choose the right question for the user's ability. This section describes the method used for item calibration of the GreEnergy quiz.

A. Procedure

The item calibration aims at measuring the difficulty of a set of items in order to build the bank item from which the adaptation algorithm will select the item to be proposed. The approach used is the IRT and, in particular, the 3-parameter logistic model is used. There were 45 items to be calibrated. In order to avoid boring users, we subdivided the items into three different tests of 15 questions each. Each test contains only questions related to a single topic, as defined previously (section III-B).

The three tests were submitted in three different sessions. In each session the pupils had to attend some lessons on a specific topic, and then they had to assess the knowledge acquired using Greenergy quiz. No time limit was defined.

The results obtained for each test were memorized in a database and used for the item calibration. First of all, the three parameters – Discrimination (a), Difficulty (b) and Guessing (c) – were calculated and then starting from these, the Item Information Function (IIF). Finally, the IIF was used to assign the difficulty level to each item. The freeware tool used for the item calibration was the PARAM-3PL (Calibration Software for the 3 Parameter Logistic IRT Model) [5].

B. Sample

The sample consisted of 75 pupils attending the third grade at the primary school "XXVIII Circolo Didattico G. Rodari – Japigia 2" in Bari. Although very young, all the pupils have a good familiarity with information technology, because they often use the computer labs and the Interactive Whiteboard.

C. Results

The data collected were 75 strings of 15 values each. For each pupil and for each test the information recorded in the db were strings such as: 111101100110100, where 0 means "wrong answer", and 1 the opposite. The strings were the input for the PARAM-3PL software. For each test, it calculates the three parameters. Table I reports the data calculated for the first test.

TABLE I. ITEM PARAMETERS ESTIMATED USING PARAM 3PL IN THE FIRST TEST

Item	Parameter a (Discrimination)	Parameter b (Difficulty)	Parameter c (Guessing)
1	2	0.054	0.7
2	0.361	-3.092	0.4
3	0.34	-2.975	0
4	2	1.076	0.8
5	1.028	-0.347	0.2
6	2	0.348	0.6
7	0.249	-3.897	0
8	2	-2.273	0
9	2	1.784	0.2
10	2	0.856	0.6
11	2	-0.893	0.3
12	0.698	-0.17	0.4
13	2	-0.476	0.1
14	2	1.277	0.3
15	0.356	-2.405	0

For each item, the three parameters calculated allow to calculate the IIF values (Table II).

The maximum value of the IIF is the basis on which to calculate the user's latent ability, that an item is able to measure. On the basis of the IIF, it is possible to assign the difficulty level to each item. In our item bank we chose three levels: easy, medium, and hard. The easy range was [-2.5; -0.5[; the medium range was [-0.5; + 0.5], and the hard range was]+0.5; +2.5].

TABLE II. IIF MAXIMUM VALUES AND RELATED ABILITY LEVEL IN WHICH THE IIF HAS THE HIGHEST QUANTITY OF INFORMATION (FIRST TEST)

Item	Max(IIF)	Ability level of Max (IIF)
1	0.19	-4
2	0.018	-2.5
3	0.22	-3.5
4	0.03	-2.5
5	0.95	-2.5
6	0.53	-0.5
7	0.81	-0.5
8	0.18	0
9	0.05	0.5
10	0.19	0.5
11	0.27	0.5
12	0.27	1
13	0.12	1.5
14	0.55	1.5
15	0.67	2

All items with an IIF value less than the minimum value (-2.5) are too easy and were discarded. Similarly, all items with an IIF greater than the maximum value (+2.5) were discarded because of being too hard.

On the basis of this classification the adaptation algorithm chooses the question to ask to the user, and the test score is calculated to assess the knowledge or skills the pupil has acquired.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

The introduction of the newest technologies in learning environments has led to the construction of more accurate and meaningful assessment systems. The CAT theory allows shorter, less boring, and more accurate assessment tests to be built. One of the mainly used psychometric approaches to build the item bank is the IRT, that requires a item calibration phase. The paper presents the calibration of a bank item for a quiz game createdf for use with pupils attending primary school. In this case, more than in others, it is important to build a test which does not cause either boredom, if for example the questions are too easy or the test is too long, or frustration if the questions are too difficult.

The future directions of our research will be to define the adaptation algorithm to select the question to be posed and then to test the quiz with larger samples.

ACKNOWLEDGMENT

Many people were involved in this work and we would like to briefly thank them: Valeria Rizzello (domain expert) and Mavi Ferramosca (teacher) for their valuable contribution to the field studies; teachers and pupils of the primary school "XXVIII Circolo Didattico G. Rodari – Japigia 2" in Bari, for participating in the study; and Patrizia Petrera who designed and implemented the GreEnergy quiz in her thesis.

REFERENCES

[1] C. Ardito, P. Buono, M. F. Costabile, R. Lanzilotti and T. Pederson, Re-experiencing History in Archaeological Parks by Playing a Mobile Augmented Reality Game, Lecture Notes in Computer Science, Springer Berlin / Heidelberg R. Meersman, Z. Tari, P. Herrero et al. (Eds.): OTM 2007 Ws, Part I, LNCS 4805, 2007, Heidelberg: Springer-Verlag Berlin, pp. 357–366.

[2] D. A. Wiley, The Learning Objects Literature. In J. M. Spector, M. D. Merrill, J. van Merrienboer, M. P. Driscoll (Eds.) Handbook of Research on Educational Communication and Technology, Routledge, 2007, pp. 345-353.

[3] H. Wainer, Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum; 1990.

[4] P. Polisani "Use and Abuse of Reusable Learning Objects". Journal of Digital Information, Vol. 3, Issue 4, 2003.

[5] Rudner, and M. Lawrence, PARAM-3PL Calibration Software for the 3 Parameter Logistic IRT Model (freeware), 2005. Available at: http://edres.org/irt/param

[6] R. K. Hambleton, and H. Swaminathan Item response theory: Principles and applications. Boston, Mass: Kluwer-Nijhoff; 1985.

[7] S. de Freitas, Learning in immersive worlds. A review of game-based learning, 2006. Available at: www.jisc.ac.uk/whatwedo/programmes/elearning_innovation/eli_outcomes

[8] S. de Freitas, C. Savill-Smith, and J. Attewell, Educational games and simulations: Case Studies from Adult Learning Practice. London: Learning and Skills Research Centre, 2006. Available at: www.lsneducation.org.uk/research/centres/RCFTechEnhanceLearn/computergames

[9] S. Downes, Resources for distance education worldwide. International Review of Research in Open and Distance Learning (July - 2001). http://www.irrodl.org/content/v2.1/downes.html

[10] T. Heins, and F. Himes, Creating learning objects with Macromedia Flash MX. Retrieved 1 February, 2003, from http://download.macromedia.com/pub/solutions/downloads/elearning/flash_mxlo.pdf

[11] T. H. Wang, Web-based quiz-game-like formative assessment: Development and evaluation, Computers and Education vol. 54, issue 4, May 2010, pp. 1157-1166.

[12] V. Rossano, M. Joy, T. Roselli, and E. Sutinen, A taxonomy for definitions and applications of LOs: a metaanalysis of ICALT papers. Journal of Educational Technology & Society, vol. 8(4), 2005, pp.148–160.

[13] W. J. van der Linden, C.A. W., Glas, Computerized adaptive testing: theory and practice, Kluwer Academic Publisher, Netherlands, 2000

Barriers to Distance Learning in Web Accessibility for Persons with Blindness and Visual Impairment

Kathleen A. Stirbens, Angela Guercio, Charles Haiber, Joseph Williams
Department of Computer Science
Kent State University at Stark
Kent, Ohio, USA
e-mail: kstirben@kent.edu, aguercio@kent.edu

Abstract *- Browsers and search engines are essential tools for retrieving information in distance learning but this technology is still a challenge for blind and visually impaired persons. Deciding to read or not read a page often requires the listening to an abundance of undesired information. While sighted people have the ability to filter information in a fast and non sequential way, blind persons rely on tools that process the information in a sequential way. Learning is slowed down by screen readers which do not interact well with web pages. In this paper we introduce WAVES, a tool for the fast retrieval of information in a web page for blind and visually impaired people. WAVES uses visual cues and query content to scan web pages and to produce an excerpt of information that helps the blind user to decide if a page is worth reading. Data from the webpage are presented to the reader in a concise and sorted-by-weight format.*

Keywords – ***Assistive technology for distance learning, blind, visually impaired, web accessibility.***

1. INTRODUCTION

Searching for relevant information on the web is an important aspect of distance learning. This activity is a challenge for visually impaired distance learners. The number of visually impaired distance learners is large and is going to increase. According to the Center for Disease Control, more than 3.4 million (3%) Americans 40 years and older are either blind or are visually impaired and millions more are at risk for developing vision impairment and blindness. By 2030, the number of blind and visually impaired people is predicted to double [16].

Many of these individuals are now or will become users of computers in one form or another. For these persons, the computer is a tool and a bane. For the blind and visually impaired, learning with computers can be frustrating and intimidating. While the sighted use mouse driven GUI (graphical user interfaces), visually impaired must rely on command line, keyboard shortcuts and screen readers (i.e. JAWS [1], Windows-Eyes [2], VoiceOver [3], IBM Easy Web Browsing[4]) for interface with their computers. Many screen readers claim to have ways to untangle web pages for ease of use, but even with the latest innovations, major problems with websites arise quickly and consequently learning is slowed down [19]. The problem is that most sites do best when they are mouse controlled. The 2 dimensional page layout and the use of frames are sometimes confusing to screen readers and it may be meaningless to a person who is blind from birth. Colors are used to show pertinent data. Popup windows redirect the page and cause confusion. Forms that are to be filled out can timeout or refresh themselves before the persons have time to fill them in. All these situations raise barriers that make it difficult for a blind user to decide if a page is worth reading.

In distance learning where much information is conveyed through web pages and/or applications that are shown via a web browser the speed and the accuracy of deciding if a page is worth reading constitutes an important step in the speed up of the learning process.

In this paper we introduce WAVES (Web Accessible for Visually-impaired Extraction System), a system designed for visually impaired and blind users that provides faster web navigation. WAVES is a tool for the fast retrieval of information in a web page which uses visual cues to filter web pages and produces an excerpt of information that helps the user to decide if a page is worth reading. Data from the webpage are presented to the reader in a concise and sorted-by-weight format which reduces the access time of the screen reader to specific information.

The paper is organized as follows. In section II we present a state of art of the existing tools and technologies for blind users. In section III we give the rational for our approach and in section IV we introduce the WAVES system while in section V we describe the methodology used for the data extraction. In section VI we discuss the experimental results obtained from the system. Finally, conclusion and future research topics are presented in section VII.

[1] http://www.freedomscientific.com/products/fs/jaws-product-page.asp
[2] http://www.gwmicro.com/
[3] http://www.apple.com/accessibility/voiceover/
[4] http://www-03.ibm.com/able/accessibility_services/EasyWebBrowsing-accessible.pdf

2. ACCESSIBILITY AND ASSISTIVE TECHNOLOGY

The problem of web accessibility is well known and many recommendations and much legislation [1] has been produced to support the needs of blind and visually impaired persons. The American Foundation for the Blind, the National Federation for the Blind, and the Web Access Initiative (WAI) [25] part of the World Wide Web Consortium, have recommendations that spell out what a blind user of a web site needs for a site to be perceivable, operable, understandable and robust but leave the actual strategy to accomplish this to the web developer.

In spite of the many existing recommendations, a jungle of difficult to navigate pages that raises barriers still exists [2, 5, 10, 13, 23]. In distance learning, assistive technology is necessary for blind and visually impaired people since it attempts to overcome the barriers and issues in the use of computer-based technology. This technology includes screen enlargers and magnifiers, screen readers, text scanners, speech synthesizers, Braille displays, Braille digital converter [9], and speech browsers. Other software solutions include Adaptive Browsing Shortcuts [17] which offer streamlined adaptive interaction with web sites. This approach, however, requires a full set of different keyboard shortcuts added to the ones already used by the screen readers and other existing web browsing software.

Research in voice-based web services accessible through web browser has produced software such as WebAnywhere [4] an online software for reading of web pages. This research does not add anything to the ability of the existing screen reader except their online availability. Also Content Management System has been used [22] to support the creation of fully accessible websites.

In spite of the many existing solutions the problem of accessing web services for blind and visually impaired users is still a reality. Blind users actually interact with websites in a completely different way than visual users [11, 14]. Creating the web page specifically for aural delivery [6] may eliminate many of the glitches that appear in and plague the more visual sites.

3. BROWSING AND SUMMARIZING

In this section we explain our rationale for designing and creating WAVES.

There is a difference between how sighted people read and how visually impaired people "read". We know that sighted people produce ad-hoc summarization of a page by scanning the page with the irregular motion of their eyes. When a sighted person opens a book, the process of browsing and scanning begins. Sighted persons do not look at each and every letter, word, picture, or paragraph. Often one look can push a sighted person immediately away from a site. Blind people, in contrast, because of Braille or audio files, must read every single word on any page before getting some feelings of what the page contains [20].

When the transition is made to the computer, the differences become even more drastic. Sighted people scan a web page looking for visual clues to determine the content at a glance. Synopsis, large print, placement on the page, and even color can be visual clues used to indicate that the content deserves to be read in more detail. For blind and visually impaired people, this aspect is lacking. No visual clues are readily available, and the screen reader that helps them to peruse the pages reads one word at a time. Web pages which are poorly developed can leave the visually impaired frustrated, discouraged, or just defeated in their searches. We believe that for visually impaired and blind users, additional software can be beneficial to simulate such a browsing process in order to speed up the decision point. Summaries designed with blind persons in mind can be very useful to reduce the amount of time required to read the data in the page. In general summaries have been considered good decision tools to determine whether to access a page or not [15].

Browsing: Browsing is an essential activity to speed up the reading process. Many researchers agree on the fact that browsing is fundamentally scanning, examining, and sampling and that this action is based on the experience of environmental perception and cognition [7]. How do blind users experience that perception and cognition to perform an effective browsing? How can a non-linear world be effectively communicated to a linear experienced user? While we understand that there is not a unique answer to such a question, we believe that small and simple summaries can be the best way to quickly decide whether something is relevant to the topic being researched.

Our goal is to provide a similar opportunity to our blind and visually impaired users.

Summarization: There are many ways to summarize, some more helpful than others. Boiling down the entire contents of a website into a small and coherent set of words is not an easy task. Many automatic techniques exist [3], and some work better than others. Some methods provide deep summarization by using techniques that incorporate knowledge-based methods with cues from natural language processing. These summarizations are more accurate, require heavy processing and return a larger amount of information to the user than shallow techniques [12, 15, 21]. Therefore they are not appropriate for a fast decision process.

Shallow techniques are faster to achieve and to select the content. They are typically based on the query content [18, 24, 26], or on links [8, 27]. Some shallow techniques are hyperlink based and use hyperlinks to extract the context from related pages rather than just examining the single page content. With this type of approach the quality of the summary depends on the amount of the referential links. While we think that adding more to the page can facilitate the decision process it also adds to the information overload and then slows the decision. In the attempt to keep the information

overload at a minimum we believe that the summary should be built without crawling, i.e. simply with the most essential data of the page.

In query based summaries query-biased sentence extraction methods are indentified [26], then a composition phase builds a summarization of the appropriate sentences. In [18], query-biased summaries that use machine learning techniques have been used for training and ranking data for sentence extraction. This approach has the disadvantage that it is often difficult or expensive to obtain training data. And again, if not appropriately skimmed, tends to produce heavier data than desired.

We believe that summarizations are a great support to overcome information overload. Whether sighted or visually impaired, there is a tendency to be overwhelmed by the sheer volume of data that is contained in a web page. Sighted people select relevant information while browsing, however none will deny that the large amount of superfluous information makes this process harder and longer even for a sighted person. For the visually impaired and blind people the problem becomes huge since the screen reading technology tends to defeat browsing or scanning the text and the user has to fight back the screen reader by manually discarding information via shortcuts.

We believe that automatic data extraction specifically designed with visually impaired and blind users in mind, can assist these users in deciding if the page is worth reading. As a consequence it is our desire to produce a minimal and effective set of information that could help the user in the initial approach to the navigation. Further perusal of the page can be obtained from the summary via the links.

4. THE "WAVES" SYSTEM

Offering to a blind or a visually impaired person an excerpt of a web page that is similar to what would be produced by the browsing and scanning activity of a sighted person is our goal. Our system, called WAVES, produces three different types of "summarizations" of the page according to the user's request: the first, called the static summary, uses only the intrinsic content of the page; the second, called the query-specific summary uses the query content together with the retrieved page content; the third, called the customized summary, utilizes a user's preferences to tailor the viewing experience of the page for the specific user.

The static summary is responsible for the drastic reduction in the information overload by eliminating any information that is not essential or useful for a blind user. For example, web sites that have pictures (poorly named or as links to elsewhere) are a problem for visually impaired users. One of the purposes of the summary is the elimination of such elements. If pictures are hyperlinked to some URL address, the hyperlink is retained while the actual image is discarded. The static summaries is responsible for identifying visual cues such as colors and font size and for retaining sentences that stand out in the page. In web pages colors and font size are not translated by the screen reader since they are meaningless for blind users however they offer valuable cues to predict the importance of the information on the website.

Links are considered elements of interest since they delve into deeper levels of information. Some links should be ignored in the summary, wherever possible. For example, navigational bar links are meant to move the reader out of the page rather than providing additional insight into the content. Such links are typically located at the very top or at the left hand side of the page however are very hard to identify without some heuristic and knowledge-based approach.

When all the elements of interest within the page are identified and ranked by level of importance, they are extracted and sorted for presentation. The good features about the static summary are they are fast to produce and if the page does not change the static summary does not change. This offers the opportunity to produce such summaries offline and to cache them for fast access.

The query-biased summary adds to the technique used to produce the static summary the use of a specific query as additional parameter for the summary generation. This summary is more refined with respect to the static one and must be dynamically generated and constructed at query time. The element of interests within the page are identified and ranked according to both the visual cues and the query content. Finally sorting-by-weight is used to create the final excerpt.

The customized summary is currently obtained by adding aural features to the excerpt. Specifically, elementary sounds selected by the user are added to specific elements, such as links, buttons, etc. for the screen reader to voice. Other customization will be considered in the future.

5. THE STRATEGY

In this section we describe the strategies for automatically identifying the elements, extracting these elements and constructing the final summary.

Identification of Elements: The process of summary generation starts with the preprocessing of the web page and the scanning of the file (.htm, .html). The WAVES system searches through the Cascading Style Sheets internally or externally attached to the page and it ranks each class, id or other meta-elements or tags. Specifically, fonts with their attributes of size, weight and color as well as headers are used to identify visual cues and titles within the page, respectively. Ranking is obtained by giving positive increasing values to larger fonts and negative decreasing values to smaller fonts in a range that goes from +5 to -5. Similarly font weight assigns negative value to lighter fonts and positive value to darker, bold fonts. Colors are ranked with respect to the distance from the background color and their frequency of use. If we consider the most frequently used color as the default color of the page, colors that differ are considered elements of attraction and therefore are scored higher. This scoring

technique is to identify and rank elements within all three types of summaries.

Extraction of Elements: When score values have been assigned to identified elements, the source is scanned for identification and the scoring of the page content is produced. In query-biased summary, additional weight is given to elements such as links, titles or sentences that contain the query.

During the extraction WAVES discards pictures or video, but retains their links where available. Hyperlinks are all retained since they are considered important elements of navigation however they are classified according to visual cues. In general hyperlinks, which are rendered with larger font size, darker font weight, or font color which differs from the background, are considered more important and, therefore, should be scored higher than other elements during the browsing action. Additional heuristics can be used to eliminate links that are not significant for the relevance judgment, but this has not been currently implemented in our system. Titles are retained while other textual information is retained only in query biased or customized summaries.

Optional sound is associated with specific website tags as an auditory cue for the blind user. An initial set up of the system that associates a sound to each web element, such as links, anchors, buttons, text, etc. is required for the production of the desired associated sound.

Finally the extracted elements are freed from possible duplications and presented to the user in a sorted-by-weight fashion that gives higher priority to the elements that scored high and lower priority to the elements that scored low. In the cases of identical values the order of the elements in the page prevails.

Figure 5.1 describes the architecture of the current prototype of the WAVES system. Currently the prototype has been implemented in C++ under Visual Studio 2008. The current software interacts with the JAWS screen reader.

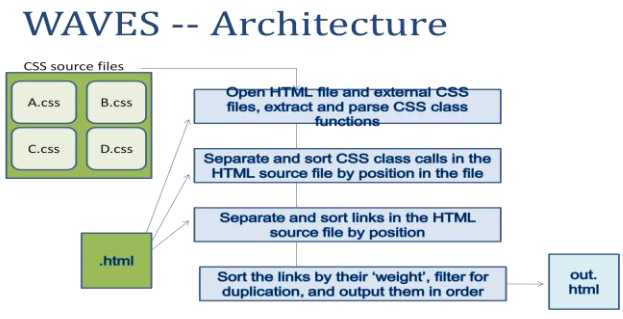


Figure 5.1- The architecture of WAVES prototype

6. EXPERIMENTAL RESULTS

In order to prove the ability and the quality of the products of our system an evaluation test has been performed. While informal, we tried to maintain all the features of a formal evaluation wherever possible. The aim of our testing was to establish if, given a list of retrieved web pages, our system would speed up the decision making process and still produce a correct result. In other words, if the user is making a decision faster while using our excerpts, is it an effective decision?

In [24], query-biased summaries are proved to be more accurate types of summaries than predefined static ones. Since we expect to obtain more effective decisions from them with respect to the simpler static summary, we have tested only the static summary. Static summaries are faster to produce and therefore a more attractive decision tool if we prove that the percentage of correctness is still very high. We did not test the customized summary since it is not our intention to prove that the addition of sounds or other features can make a significant difference in the outcome.

The Experiment: In the experiment a group of 14 subjects were used which included sighted, blind and visually impaired persons. The JAWS screen reader was used to read the page for all subjects. Sighted subjects were exposed to the screen reader before the test for familiarity. Environmental conditions, such as experimenter's behavior, noise background, lighting conditions, etc, have been maintained as similar as possible during the experiment.

A list of retrieved web pages typical of those encountered in a distance learning course was generated from a group of 5 randomly selected queries. The list was given to the subject together with only 1 randomly chosen query from the selected set of 5 queries. The subject received all the required instructions to perform the experiment from the test administrator. In no case were the subjects informed of the hypothesis we were trying to prove before starting the experiment. Two groups were formed. The first group was allowed to listen only to the full text web pages from the retrieved list while the second group was allowed to listen only to the static summary of the pages generated by our system from the retrieved list.

Both groups were asked to mark on an answer sheet if the page had to be rejected or accepted based on their assigned query. For blind and visually impaired people the answer sheet was filled out by the test administrator. A maximum of 10 minutes was given to each subject to identify as many pages as possible in the list that related to their query. If the subject completed the task before the end of the time limit, the time of completion was retained by the test administrator and written on the answer sheet. All the answer sheets were collected immediately after the time limit. After the sheets were collected, the test administrator shared the purpose of the experiment and encouraged the subjects to express their opinion about the experiment.

The Results: The accuracy of the decision was measured by counting the number of pages correctly identified by the subject with respect to the total number of relevant pages the subject was able to process from the retrieved list within the time limit. The speed was determined by the total number of

web pages examined by the subject in the given amount of time with respect to the entire list.

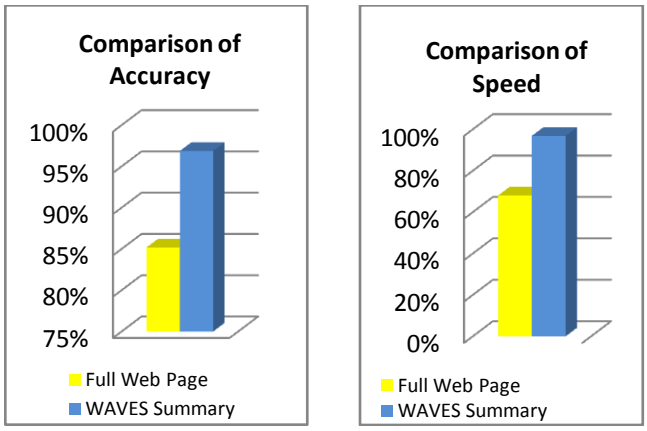


Figure 6.1- Comparison of accuracy and speed

The results of the test are shown in Figure 6.1. As expected the speed of using the static summary is higher than that of browsing with full web pages (68.06% vs. 96.8%). This confirms our hypothesis and was expected since the excerpts produce a reduced amount of data to be read compared to the full webpage. More interesting is the result of the accuracy which highlights that there is a dramatic difference between the two methods. The level of accuracy in the use of the summary was much higher (85.19% vs. 96.88%). We believe that this is due both to the quality of the summary, which filters and sorts significant information, as well as to the fact that in the case of full web pages, elements that indentify the page relevance are often embedded in volumes of unrelated information, such as advertisements, buttons, reading of frames and graphics which can discourage the user from processing the entire page. In fact, we observed that in the case of full text web page reading often the subject was frustrated by the length of the page, the amount of extraneous data, and the process in general. The ability of identifying and relocating relevant information towards the top of the page made the summary very attractive and it facilitated the identification process as the results show.

7. Conclusions and Future Research

Searching for relevant information on the web is an important aspect of distance learning. This activity is a challenge for visually impaired distance learners. Learning is slowed down in blind and visually impaired people by barriers raised by web accessibility. In this paper we have introduced a system for the creation of page excerpts that are effective for browsing of web pages for blind and visually impaired users. The system effectively identifies a minimal set of relevant information that speeds up the decision process if a page is worth reading. The prototype has been designed and tested and the results are very encouraging. It is our intention to create a plug-in for Mozilla Firefox that can be made accessible to blind and sighted as well. Even though we did not separately test blind persons from sighted persons the informal feedback that was recovered after the test, shows that this excerpt is also very beneficial for sighted people. In the next version of the system, more attention should be given to pop-ups, buttons and forms. We also intend to use the features of the Semantic Web for better solutions.

8. References

[1] *A Guide to Disability Rights Laws*" U.S. Dept. of Justice Civil Rights Division Disability Rights Section, (2005), http://www.ada.gov/cguide.htm

[2] Abou-Zahra, S., Brewer, J., Arch, A., "Towards Bridging the Accessibility Needs of People with Disabilities and the Ageing Community", *Proc.of W4A'08*, Beijing, (2008):83-86.

[3] "*Advances in Automatic Text Summarization"* Ed. by I. Mani and M. T. Maybury. Cambridge, MA: The MIT Press; (1999): 1-442.

[4] Bigham J.P., Prince C.M., Ladner R.E., "WebAnywhere: A Screen Reader On-the-Go", *Proc. of W4A'08 - Challenge*, Beijing, China, April 21-22 (2008): 132-133.

[5] Brajinik, G., Yesilada, Y., Harper, S., "Guideline Aggreagation: Web Accessibility Evaluation for Older User" *Proc. of W4A'09*, Madrid, Spain, Apr. 20-21 (2009): 127-135.

[6] Brophy P., Craven J., "Web Accessibility", *Library Trends* 55.4 (2007): 950-972.

[7] Chang, S.-J., Rice, R.E., "Browsing: a Multidimensional Framework", *Annual Review of Inf. Sci. and Tech.*, **28** (1993):231–276.

[8] Delort, J.-Y, Buchon-Meunier, B., Rifqi, M., "Enhanced Web Document Summarization Using Hyperlinks", *Proc. of 14th ACM Conf. on Hypertext and Hypermedia*, Nottinghan, UK, Aug. 26-30 (2003): 208-215.

[9] Fox R.,"You've Got Braille", *Comm. of ACM,* **43(**12**)** (2000): 10.

[10] Fuertes, J., Gonzalea, et al., "Here-FFX: a Firefox add-on for semi-automatic web accessibility evaluation" *Proc. of W4A'09*, Madrid, Spain, Apr. 20-21 (2009): 26-35.

[11] Gerber, E. "The benefits of and barriers to computer use for individuals who are visually impaired." *J. of Visual Impairment & Blindness,* **97**, Sept. (2003): 536-550.

[12] Gong, Y., Liu, X., "Generic text summarization using relevance measure and latent semantic analysis". *Procs of 24th Annual Int. ACM SIGIR Conf. on Research and Development in Inform. Retrieval*, ACM Press, (2001):19–25.

[13] Hanson, V., "Age and Web Access: The Next Generation" *Proc. of W4A'09*, Madrid, Spain, Apr. 20-21 (2009): 7-15.

[14] Harper, S., Goble, C., Stevens, R., "Augmenting the mobility of profoundly blind web travellers", *The New Review of Hypermedia and Multimedia*, **11**(1), June (2005):103-128.

[15] Harper S., Patel N., "Gist Summaries For Visually Impaired Surfers", *Proc. of 7th Int. ACM SIGACCESS Conf. on Comp. and Accessibility,* Baltimore, MD, (2005): 90-97.

[16] "*Improving the Nation's Vision Health: A Coordinated Public Health Approach*", Center for Disease Control and Prevention, Report:1-53 http://www.visionandhealth.org/documents/ReportImprovingtheNationsVisionHealth.pdf

[17] Kouroupetroglou C., Salampasis M., and Manitsaris A., "Analysis of navigation behaviour of blind users using Browsing Shortcuts", *New Review of Hypermedia & Multimedia,* **14**(2), Jan. (2008):199-228.

[18] Metzler, D. , Kanungo, T., "Machine Learned Sentence Selection Strategies for Query-Biased Summarization", *SIGIR Workshop on Learning to Rank for Information Retrieval*, (2008):40-47.

[19] Miyashita, H., Sato, D., "Making Multimedia Content Accessible for Screen Reader Users", *Proc. of W4A 2007 - Challenge*, Banff, Canada, May 07-08, (2007): 126-127.

[20] Morkes, J., Nielsen, J., "Concise, scannable, and objective: How to write for the web", (1997) http://www.useit.com/papers/webwriting/writing.html

[21] Radev, D. R., McKeown, K. R., "Generating natural language summaries from multiple on-line sources", *Comput. Linguist.*, 24(3), (1998):470–500.

[22] Rainville-Pitt, S., D'Amour, J., "Using a CMS to create fully accessible websites", *Proc. of W4A 2007 - Challenge, Banff,* Canada, May 07-08, (2007): 130-131.

[23] Sayago, S., Blat, J., "About the relevance of accessibility barriers in the everyday interactions of older people with the web" *Proc. of W4A'09*, Madrid, Spain, Apr. 20-21 (2009): 104-113.

[24] Trombos, A., Sanderson, M., "Advantages of query biased summaries in iformation retrieval", *Proc. 21st Ann. Intl. ACM SIGIR Conf. on Research and Development on Information Retrieval*, (1998):2-10.

[25] "*Web Content Accessibility Guidelines 2.0*", W3C http://www.w3.org/TR/2008/REC-WCAG20-20081211/

[26] White, R., Ruthven, I., Jose, J.M., "Web Document Summarisation: A Task-Oriented Evaluation", *Int. Workshop on Digital Libraries. Proc. of 12th Int. Database and Expert Systems Applications,* (2001): 9-55.

[27] Yesilada, Y., Harper, et al., "Screen Readers Cannot See", *Proc. of ICWE'04*, LNCS-3140, (2004): 445-458.

A Novel Approach for Attention Management in E-learning Systems

Gennaro Costagliola, Mattia De Rosa, Vittorio Fuccella
Department of Mathematics and Informatics
University of Salerno
{gencos,vfuccella}@unisa.it

Nicola Capuano, Pierluigi Ritrovato
DIIMA - University of Salerno
{ncapuano,pritrova}@unisa.it

Abstract

This study presents new approaches for the detection and treatment of the attention of a student by an e-learning system through the use of the information given by the implicit interaction of the student with the system and the data coming from non-invasive devices such as webcams.

Furthermore, the paper proposes two models for the treatment of the attention of students to be applied to an existing e-learning environment, in order to provide personalized content to the students and thus improving their learning experience.

1. Introduction

Over the past ten years e-learning has evolved from early systems to *Intelligent Tutoring Systems* (ITS), *Smart Classrooms* and *Mobile Learning* (e-Learning with mobile devices). Today, e-learning aims to be strongly student-centered, in order to provide a personalized learning experience. Its principal objectives are then not only to foster successful learning but also to involve students in the learning process and maximize their interest [15].

An expert teacher easily captures the emotional state of students and adapts lessons accordingly, in order to maximize their interest and participation. An e-learning system in order to provide a quality educational experience should be able to behave similarly [2].

This need is reflected in the development of systems capable of detecting the attention of a student during e-learning sessions. The use of biometric sensors can provide good information on the student's emotional state [7], but at the same time it can create physical discomfort to the student in addition to possible hardware costs and logistic problems. It is therefore advisable to use non-invasive systems such as log analysis [6] and cheap hardware already present on computers [18] such as a webcam [3, 4].

This document briefly describes the e-learning system *IWT* (*Intelligent Web Teacher*) [9, 5, 14], an ITS that provides personalized content to the student, and *WiSe*, a system that detects the attention of a student through the analysis of the IWT interaction log and of the webcam video of the student during the e-learning sessions.

Finally we present a basic and an extended model for the integration of the student attention detection module with the e-leaning environment, in order to customize the learning content and improve the learning experience.

The contribution of the research presented in this paper lies in the definition of novel models for the application of automatic detection of student attention to e-learning intelligent tutoring systems, and in particular to the IWT system.

This paper is organized as follows. Section 2 discusses the state of the art of attention detection and its use in e-learning environments. Sections 3 and 4 briefly describe the characteristics of *IWT* and *WiSe*. Sections 5 and 6 present the two alternative models for managing the attention in the IWT e-learning environment. Finally, section 7 offers our conclusions and describes further work.

2. Related Work

In literature there are several studies dealing with the detection of the emotional state of computer system users. Only some of them focus on the detection of the user's attention.

In [8] the authors postulated the existence of a direct link between the user's comfort and emotional state. A research team at Purdue University has instead developed a system for the perception of posture [17, 19] through the analysis of the pressure exerted by the body on the chair. For this purpose they used a chair properly equipped with pressure sensors. The system performs real-time analysis using computer vision techniques to map pressure distribution obtained from sensors located on the seat and backrest.

An interesting conceptual study in [11] theorizes that some macro movements on the chair are indicators of the emotion and therefore suggests the possibility of creating an office chair that can adapt itself to the user's emotional state.

The Affective Computing Group of the MIT Media Labs, has carried out an interesting project concerning the perception of the state of interest (high, medium, low, bored, taking a break) in children analyzing their sitting position

[12]. Starting from the results of [17], they exploit a neural network for classification of posture and a Hidden Markov Model for recognizing the state of interest. Through the use of the pressure maps they claim they can obtain an accuracy of 82% for known subjects and 76% for unknown subjects.

In [10] the authors deal with the creation of computing and communication systems that can detect and reason on the human attention by fusing the information received from multiple sources. A probabilistic model combines data from sensors, from user interaction with the system, from the previous pattern of activity and attention in order to estimate the user's attention and then to adapt the system behavior.

In [3] the authors present a neuro-fuzzy approach to infer the attention level of a user in front of a monitor using a simple camera.

In [4] the authors estimate the level of attention/interest of a user who reads the text on the computer using a camera to detect the position and movement of certain points around the eyes and the position of the iris. The system analyzes user behavior and provides a model of six general learners' states (Frustrated/Struggling to read, Tired/Sleepy, Not paying attention, Distracted, Attentive, Full of interest). The authors present a case study where an e-learning system provides modifications to the presentation of the text according to the level of attention detected in children with dyslexia.

In [13], the authors describe the use of physiological signals to improve student interactions with character-based interfaces that adapt themselves to reflect the user's affective state. The paper presents "Emotion Mirror", an example of a system where emotions are sent back to the user and its evolution "Emphatic Companion", an agent (represented by a character) that adapts its behavior according to the emotional state of the user, e.g. giving support and encouragement.

As a part of a learning content recommendation system, the work presented in [15] uses biometric sensors to detect the emotional state of the student in order to adjust the content offered by the system. By comparing sessions implementing emotion detections to sessions without detections, they see that the manual interventions of the student (required when the system does not automatically provide the content that the student needs) are reduced by 91% in the first case.

From the literature we can note that even though there exist studies that treat the automatic detection of attentional/emotional states in e-learning, however they do not show how to apply it to existing e-learning intelligent tutoring systems.

3. Background: The IWT Platform

IWT, [9, 5, 14], is an e-learning intelligent tutoring system whose aim is to customize the learning experience to the real needs and preferences of the student. The innovative features of IWT compared to other e-learning solutions can be summarized as follows:

- possibility of automatic or assisted generation of learning paths from the learning objectives;
- ability of automatically customize courses based on previous knowledge of the individual students and their learning preferences;
- possibility of content management at a high level of abstraction using ontologies;

IWT models the knowledge through Learning Objects (*LOs*), Metadata and Ontologies. The LOs are the basic teaching modules that can be used during learning. The Metadata formally describe the LOs through a standard set of attributes. In particular, IWT adopts the IEEE LOM [1] standard that specifies the description of the LO through 47 items grouped into 9 categories. Ontologies represent the teaching domains and offer knowledge management concept-oriented support at a higher abstraction level.

IWT is able to (automatically) capture the learning preferences and student acquired knowledge during their educational experience. IWT manages this informations using three elements: a Cognitive State, that represents the knowledge possessed by the student by concept-vote pairs, a set of Learning Preferences that refer to fields of the Educational category defined by the IEEE LOM metadata and a set of Evolution Rules used for updating the student's cognitive state at the end of each verification test. In particular, the Educational category fields correspond, among the others, to *Interactivity Type*, *Learning Resource Type*, *Interactivity Level*, *Semantic Density*.

In IWT a course is generated by a set of Objective Concepts defined by the teacher or by the student him/herself. It first builds the best learning path for a given student considering his/her Cognitive State (eliminating already known concepts and adding any missing pre-requisites). From the learning path so constructed, it generates the best presentation for a given student by considering his/her Learning Preferences and choosing, therefore, the LO more congenial to him/her.

In the course, in general, the fruition begins with the first LO and continues until it reaches a verification test (Milestone). When the test ends, adjustments are made to the learning path portion not yet viewed by the student to respond to any weaknesses identified through the insertion of recovery LOs.

4. WiSe attention detection system

In order to obtain information about the student attention, IWT uses the WiSe platform services. The communication between IWT and WiSe follows the request-reply model (IWT requests the student attention when it needs and WiSe replies with the attention level detected in that moment).

WiSe monitors students during learning sessions and es-

timates their attention level by considering:

- the capture of half-length figure of the student with a webcam (video tracking);
- the capture of the computer screen and of keyboard and mouse generated input (input tracking);
- the capture of information about the tasks simultaneously active on the user computer (activity tracking).

Basing on these measurements, WiSe calculates a current attention level that can be one of the following 4 discrete values: high, medium, low, distracted. This value is calculated by combining the results of three types of anal-

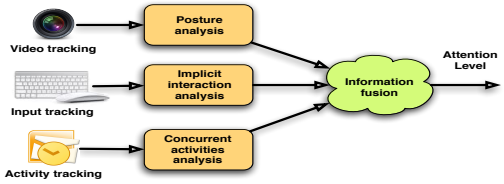


Figure 1. Elaboration process for the calculation of the attention level

4.1. Posture analysis

WiSe determines a first approximation of the attention level through the detection and the analysis of the user sitting posture in front of the monitor. The posture is deduced by analyzing the sequence of images taken by a frontal webcam through head tracking algorithms capable of detecting the position and the orientation of the head in six degrees of freedom.

Through classification algorithms, values representing the position and the orientation of the head are mapped on 9 major and 9 minor poses. Information about the gaze direction (to the screen or elsewhere) is added to the detected pose in order to infer the visual attention focus. The obtained information, in accordance with [12], is analyzed using pattern recognition techniques able to extract an indication of the perceived attention level.

4.2. Implicit interaction analysis

Implicit interactions between the user and the system are a useful indicator of the user's attentional state. WiSe, in particular, measures and processes the following parameters: the TSR (Time Spent for Reading), the TSS (Time Spent for Scrolling), the link clicking/link following time, the downtime, the time spent moving the mouse, the number of clicks, the page resize events and the keyboard typing.

WiSe is able to track and process these parameters, applying a statistical model that relates the implicit interaction events with the attention level. In the elaboration process, the collected data are related to the context information reflecting the specific content received, the environmental conditions and the characteristics of the user obtained from the student model.

4.3. Concurrent activities analysis

For the measurement of attention WiSe also uses an evaluation of the load required by tasks simultaneously active on the user computer. In fact it is widely acknowledged that the voluntary attention, involving cognitive processes, has a finite capacity. The WiSe analysis currently focuses on only two types of events: messages exchanged by e-mail and instant messaging tools.

Through automatic text analysis techniques, WiSe evaluates the marginal significance of information compared to the cost of the interruption to determine whether it is useful information or noise. In particular, incoming messages are classified by measuring the relevance of their content with the content of active LOs and the membership of senders to the same educational context of the user (e.g. teachers, tutors, fellow students).

5. The attention management: the basic model

Depending on the level of attention estimated by WiSe the IWT e-learning platform performs actions aimed at improving student learning.

This section describes our basic model for the management of attention by IWT. In this model, given the current level of attention (as received by WiSe) and the latest recorded levels of attention, IWT decides whether to present the student with a set of seven possible actions in order to raise his/her attention. This model does not require to change the WiSe attention detection system and allows the student to choose how to proceed. This prevents IWT from producing "wrong" actions. Furthermore, since alerts may distract/bother the student (as indicated in [10]), the model tends to minimize its occurrences.

5.1. The attention analysis

In this section we show how to calculate two indices, the *weighted average attention* (waa) and the *attention trend* (at) to represent the attention profile of the students during their learning session. Both of them are calculated taking into account the attention levels returned by WiSe every 20 seconds. This time interval has been defined by analyzing the data of the IWT learning sessions of 37 students. In fact, we observed that the attention used to change every 43.8 seconds on average with a standard deviation of 41.7. Furthermore, we noted that the total time of the attention states with a duration greater than 20 seconds accounted for the 92% of the total duration of the sessions.

The indices are calculated by taking into account the last 15 attention levels and by assigning a numerical value to

each possible attention level, as follows:

high	average	low	distracted
1	0.67	0.33	0

For each of the 15 observations (from the oldest attention level to the newest) we assign an increasing weight calculated according to the function $f(x) = \frac{x}{n}$ ($n = \sum_{i=1}^{15} i = 120$), as shown in the following:

Observation no.	1	2	3	4	5
Weight	0.008	0.017	0.025	0.033	0.042
Observation no.	6	7	8	9	10
Weight	0.050	0.058	0.067	0.075	0.083
Observation no.	11	12	13	14	15
Weight	0.092	0.100	0.108	0.117	0.125

Note that this weight assignment gives more importance to the latest observations.

We then calculate the first index by the formula $waa = \frac{\sum_{i=1}^{15} att_i \times weight_i}{k}$ where $k = \sum_{i=1}^{15} weight_i$, att_i is the attention value in the i-th observation, and $weight_i$ is the weight assigned to the i-th observation in the table. Since $k = 1$, we have $waa = \sum_{i=1}^{15} att_i \times weight_i$. The left chart in Figure 2 shows an example with some sample data.

In order to calculate the attention trend, for each observation, we calculate the difference between the value of the attention and the weighted average attention and then multiply it by the observation weight, that is $a_i = (att_i - waa) \times weight_i$.

Finally, on the values (i, a_i), we first draw the line corresponding to the linear regression (see the right chart in Figure 2) and then we calculate the difference between the y values of the line for $i = 15$ and for $i = 1$. It is possible to prove that this difference ranges between approximately -0.08 and $+0.08$ where a positive value indicates a growing trend of attention, while a negative value a decreasing trend.

In order to have discrete values for the two indices we define two thresholds $0 < t1 < t2 < 1$ for the weighted average attention and two more thresholds $-0.08 < r1 < 0 < r2 < +0.08$ for the attention trend. The thresholds allow us to define the discrete values as shown in the following:

waa	$0 \leq waa < t_1$	$t_1 \leq waa < t_2$	$t_2 \leq waa < 1$
Value:	low	medium	high

at	$-0.08 < at < r_1$	$r_1 \leq at \leq r_2$	$r_2 < at < +0.08$
Value:	decreasing	stable	increasing

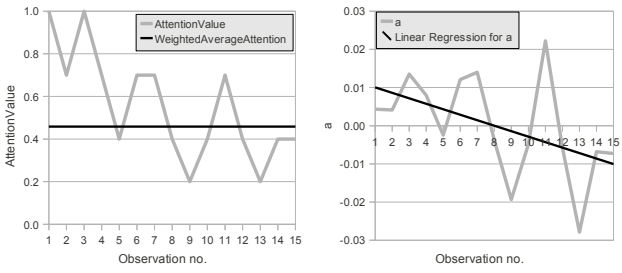


Figure 2. Weighted average of the attention and the attention trend on sample data

5.2. The use of the Attention Profile

Once IWT has calculated the student attention profile it needs to decide how to react in order to improve the student learning experience. In this model IWT will have to decide whether to present the student with a set of possible actions, without being too annoying. The decision is based on the two following conditions:

- the pair *weighted average attention/attention trend* has value medium/decreasing or low/stable or low/decreasing, and
- at least two minutes have passed since IWT has presented the list of actions for the last time.

In the case the student has lately chosen not to do any action then the minutes of waiting will be doubled.

These choices are motivated by the fact that if the attention trend is increasing we do not need to operate, if it is stable we need to operate a change only if the attention is low, and if it is decreasing then we do not need to operate only when the attention is high. Please note that the system will act only in three cases out of nine and only when the student has not been recently helped, and this is done in order to be as less invasive as possible.

In the following we give the list of the IWT possible actions to be selected by the inattentive student:

1. calculate a new learning path to start after the current LO, using different parameters from the original one (the student indicates if the current LO type of interaction is pleasant/unpleasant, if the level of interactivity is too low or too high, if the LO is too easy/difficult, if the semantic density is too high or too low; according to these data, the system obtains new values for the interaction type, interaction level, difficulty level, semantic density);
2. show a LO on the same subject but with different metadata (the student chooses from a list showing the other LOs available and their interactivity type, interactivity level, difficulty, semantic density);
3. insert a test in the learning path after the current LO;
4. display a LO on a correlated topic selected by the system based on its ontology:
5. provide a moment of relax/pause at the end of the current LO;
6. change the LO presentation mode (font size, etc.);
7. do nothing (the student can also indicate whether he was really distracted and why).

The actions will be presented to the student in the form of a simple questionnaire.

6. The attention management: the extended model

As in the basic model the e-learning platform IWT, depending on the level of attention estimated by WiSe, performs actions aimed at improving student learning.

This section describes our extended model for the management of the attention by IWT. Differently from the basic model, in this extension IWT does not always present the student with a list of actions, but it has to perform most of the actions (in order to raise his/her attention) without user interaction, according to the present and past levels of the student's attention. This avoids the risk of distract/bother the student.

This model also requires an extension of the WiSe attention detection system in order not only to discover the attention level but also to detect its cause. This gives more information when choosing the next action to take. The extension requires that the attention state is represented by a pair of values:

- the attention level: a real value between 0 (low) and 1 (high);
- reason of attention: a justification value for an attention level (bored, distracted, sleepy, frustrated (too difficult topic), search on Internet for information related to the course, chat about topics of the course, perform actions not related to the current topics, disturbed by someone in the classroom, idle, absent).

Note that this model involves the risk of performing "wrong" actions, risk that IWT decreases using a statistical analysis on students' previous interactions.

6.1. The attention analysis

As in the basic model IWT requests the attention value every 20 seconds. The attention profile includes the *weighted average attention*, the *attention trend* (calculated similarly to the basic model) and the most relevant *reason of attention*. We also calculate and store the attention trend after the execution of each action, in order to provide useful information to IWT to choose the next actions.

6.2. The use of the Attention Profile

Once IWT has calculated the student attention profile it needs to decide how to react in order to improve the student learning experience. In this model IWT will have to decide whether to perform actions and, if so, which ones by taking into account the logged data from previous IWT decisions.

After an initial training phase aimed at initializing the log, IWT executes the algorithm in figure 3 to calculate the *weighted average attention trend* for each action. The algorithm takes as input the current LO, the current student, the current attention profile and the set of log entries defined by the tuple (*executed action*, *student*, *LO*, *attention profile before the action*, *attention trend after the action*). Moreover it makes use of the two functions *attentionProfileSimilarity* and *LOSimilarity*. Both functions return a value between 0 and 1 but while LOSimilarity calculates the similarity between the metadata of two LOs, attentionProfileSimilarity calculates the similarity between two attention profiles.

for all action a from $actionList$ **do**
  $i \Leftarrow 1$;
  for all log entry le where $le.executedAction = a$ and $le.student = currentStudent$ **do**
    $w[i] \Leftarrow$ (LOSimilarity($currentLO, le.LO$)+ attentionProfileSimilarity($currentAttentionProfile$, $le.attentionProfileBeforeTheAction$))/2;
    $at[i] \Leftarrow le.attentionTrendAfterTheAction$;
    $i \Leftarrow i + 1$;
  end for
  $av1 \Leftarrow \frac{\sum_{n=1}^{i-1} w[n] \times at[n]}{\sum_{n=1}^{i-1} w[i]}$;
  $i \Leftarrow 1$;
  for all log entry le where $le.executedAction = a$ and $le.LO = currentLO$ **do**
    $w[i] \Leftarrow$ attentionProfileSimilarity($currentAttentionProfile$, $le.attentionProfileBeforeTheAction$);
    $at[i] \Leftarrow le.attentionTrendAfterTheAction$;
    $i \Leftarrow i + 1$;
  end for
  $av2 \Leftarrow \frac{\sum_{n=1}^{i-1} w[n] \times at[n]}{\sum_{n=1}^{i-1} w[i]}$;
  $weightedAverageAttentionTrend[a] \Leftarrow \frac{av1+av2}{2}$;
end for

Figure 3. Algorithm that calculates the past attention trend for each possible action

The algorithm takes into account for each type of action:

- the logs entries related to the current student;
- the logs entries related to the current LO.

In the former case the weighted average of the logged attention trend (*av1*) after an action execution is calculated using as weights the similarity between the current LO and the LO indicated in the log and the similarity between the current attention profile and the attention profile stored in the log. In the latter case *av2* is calculated using as weights only the similarity between the current attention profile and the attention profile stored in the log.

The first value provides information on the attention trend observed in the past for the current student with similar attention profile/LO, while the second provide informations on the attention trend observed in the past for other students with similar attention profile on the current LO.

IWT then executes the action with the highest calculated weighted average attention trend if at least two minutes have passed since the latest executed action. It can be noted that the system dynamically adapts itself to the student needs and to the current LO in order to maximize the future attention trend.

In this extended model the possible actions are both selected and executed by IWT. The list of actions are as follows:

- calculate a new learning path to start after the current LO, using different parameters from the original one (e.g. different values for interaction type, interaction level, difficulty level, semantic density);
- insert a test in the learning path after the current LO (e.g. for doubtful cases);
- provide a moment of relaxation/pause at the end of the current LO (e.g. if too tired);
- show an alert message to call a distracted user;
- show an alert message that asks the user whether s/he wants to take a LO on another topic selected by the system (e.g. when the user uses a search engine to search for something on the basic topics of the LO);
- show an alert message that asks the user whether s/he wants to see a LO with different metadata on the same topic instead of the current LO (selectable from a list);
- change the LO presentation mode (font size, etc.);
- do nothing.

7. Conclusions and future research

In this study we described the IWT e-learning environment, an ITS that can customize the learning content according to the profile of the student. We then described the WiSe system for the attention detection of the IWT students inferred from the information granted by the implicit interaction of the student with the system and data coming from non-invasive devices (webcam).

Finally we presented two alternative models for the management of the attention by IWT, in order to dynamically adapt the content presented to the students thus improving the learning experience.

Future research will aim at a more formal validation of the presented models, by organizing e-learning sessions of tests on IWT/WiSe with students from the University of Salerno.

8. Acknowledgement

This work was realized within the WiSe "Wireless Services: an User Centric Approach" project co-financed by the MIUR (D.Lgs. N. 297/1999).

This research has been partially supported by the grant "cofinanziamento per attrezzature scientifiche e di supporto, grandi e medie (2005)" from the University of Salerno.

References

[1] Ieee standard for learning technology-extensible markup language (xml) schema definition language binding for learning object metadata. *IEEE Std 1484.12.3-2005*, pages 0_1–46, 2005.

[2] S. Alexander, A. Sarrafzadeh, and C. Fan. Pay attention! the computer is watching: Affective tutoring systems. In A. Rossett, editor, *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003*, pages 1463–1466, Phoenix, Arizona, USA, 2003. AACE.

[3] S. Asteriadis, K. Karpouzis, and S. Kollias. A neuro-fuzzy approach to user attention recognition. In *ICANN '08: Proceedings of the 18th international conference on Artificial Neural Networks, Part I*, pages 927–936, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias. Estimation of behavioral user state based on eye gaze and head pose–application in an e-learning environment. *Multimedia Tools Appl.*, 41(3):469–493, 2009.

[5] N. Capuano, M. Gaeta, A. Marengo, S. Miranda, F. Orciuoli, and P. Ritrovato. LIA: an intelligent advisor for e-learning. *Interactive Learning Environments*, 17(3):221–239, 2009.

[6] M. Cocea and S. Weibelzahl. Can log file analysis estimate learner's level of motivation? In *Proc. of 14th Workshop on Adaptivity and User Modeling in Interactive*, pages 32–35, 2006.

[7] C. R. C. Conati and H. Mclaren. A study of using biometric sensors for monitoring user emotions in educational games. In *Proc. of Workshop on Modeling User Affect and Actions: Why, When and How*, 2003.

[8] P. A. Fenety, C. Putnam, and J. M. Walker. In-chair movement: validity, reliability and implications for measuring sitting discomfort. *Applied Ergonomics*, 31(4):383–393, 2000.

[9] M. Gaeta, F. Orciuoli, and P. Ritrovato. Advanced ontology management system for personalised e-Learning. *Knowledge-Based Systems*, 22(4):292–301, 2009.

[10] E. Horvitz, C. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communication: from principles to applications. *Commun. ACM*, 46(3):52–59, 2003.

[11] P. V. F. K. C. K. J. Overbeeke. The emotion-aware office chair. In *Proc. International Conference on Affective Human Factors Design, Asean Academic Press, London*, 2001.

[12] S. Mota and R. W. Picard. Automated posture analysis for detecting learner's interest level. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, volume 5, pages 49–49, 16-22 2003.

[13] H. Prendinger, J. Mori, S. Mayer, M. ISHIZUKA, and R. Member. Character-based interfaces adapting to users' autonomic nervous system activity. In *Proceedings of the Joint Agent Workshop (JAWS-03)*, 2003.

[14] E. Sangineto, N. Capuano, M. Gaeta, and A. Micarelli. Adaptive course generation through learning styles representation. *Universal Access in the Information Society*, 7(1):1–23, 2008.

[15] L. W. M. S. R. Shen. Affective e-learning: using "emotional" data to improve learning in pervasive learning environment. *Educational Technology Society*, 12-2:176–189, 2009.

[16] H. Tan, L. Slivovsky, and A. Pentland. A sensing chair using pressure distribution sensors. *Mechatronics, IEEE/ASME Transactions on*, 6(3):261–268, sep 2001.

[17] S. P. Tarzia, R. P. Dick, P. A. Dinda, and G. Memik. Sonar-based measurement of user presence and attention. In *Ubicomp '09: Proceedings of the 11th international conference on Ubiquitous computing*, pages 89–92, New York, NY, USA, 2009. ACM.

[18] M. Zhu, A. M. Martinez, and H. Z. Tan. Template-based recognition of static sitting postures. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, volume 5, pages 50–50, 16-22 2003.

Full Formative Assessment based on Educational Video Games

Salvatore Cuomo
Università di Napoli Federico II
Via Cinthia 26
80137 Napoli
salcuomo@unina.it

Vittorio Fuccella
University of Salerno
Via Ponte Don Melillo
84084 Fisciano (SA), Italy
vfuccella@unisa.it

Aniello Murano
Università di Napoli Federico II
Via Cinthia 26
80137 Napoli
murano@na.infn.it

Abstract

The attention on the use of computer games in education (educational gaming) is quickly growing. By investigating the impact of video games in school, researchers have shown a remarkable improvement in student engagement and motivation. Nevertheless, the effects of educational gaming have rarely been considered on a long running application and, most important, they have never been considered in the whole teaching process as a means for teachers to real-time monitor students' learning. In this paper, we investigate the use of video games associated with formative assessment to pursue the above objective. In particular, we have conceived and implemented a portal, named PlayLearn, aimed at administering online tests based on multiple-choice quizzes, by means of video games. By letting the students play, the portal helps students to learn, it evaluates them and, finally stores their learning parameters which can be used by the teachers to improve in real-time the learning process, all in a virtuous circle.

1 Introduction

Nowadays, the education community is facing a new generation of students: the "*Digital Natives*": born into a world of pervasive digital technologies, they behave differently from the "*Digital Immigrants*" who are teaching them. This difference has led to a very crucial debate on whether it is possible to create a more effective learning process by using these technologies.

Even though, as reported by Bennett et al. [2], claims made on the *digital natives* are often not supported by a scientific evidence, it cannot be denied that the actual generation of students has grown up in a closer contact with digital technologies such as computers, Internet, mobile phones and MP3s, etc., than the past generations. One of the activities in which the new generation is much more involved is playing with video games. As stated by Prensky [6], "today's average college graduate students have spent less than 5,000 hours of their lives reading, but over 10,000 hours playing video games". This aspect was also observed by the authors in a Project (*Innovambiente*) funded by the Italian Ministry of Innovation based on new technologies. In [5] it is shown that in the *Innovambiente Project*, implemented in a secondary first level Italian schools, interactive whiteboard and educational-based games can be used by teachers as useful tools in order to improve the students "emotional intelligence". Many educators think that this attitude to "play" with technologies could be fruitfully used for learning. Indeed, the attention on the use of computer games in education is growing [4, 7, 8, 9]. By investigating the impact of video games on high school students, researchers have shown a remarkable improvement in student engagement and motivation [1]. Nevertheless, the effects of *educational gaming* have rarely been considered on a long running application and, most important, they have never been considered in the whole teaching process as a means for teachers to real-time monitor students' learning.

In this paper, we investigate the use of video games associated with *formative assessment* to pursue the above objective: we have designed and implemented a portal, named *PlayLearn*, which allows the students to access formative online tests while playing an enthralling video game. *Formative* assessment distinguishes itself from *summative* assessment since it occurs during the learning process, gives information on the learning state of the class and allows the teacher to decide the most suitable learning path [3]. Following this principle, the teacher can access the report section of our portal and real time monitor the state of the class. Thus deciding the most appropriate action to improve his/her teaching.

PleayLearn is a web application equipped with a database. Each student has its own credentials to access the system. At any lesson *PleayLearn* works as follows. The teacher generates a test, by randomly selecting the questions from a common base for that lesson. Clearly, each question can be edited by the teacher. Then, the teacher validates the test and opens the game session. Now, the student can log

in *PlayLearn*, play a video game and answer the questions in order to obtain a high score on the game. Each given response will be stored in the database. After giving a response, the student will know the right answer and get a reward only if s/he is right. At any moment, the teacher can access the system and check the number of students who have already played the game, the overall knowledge level of each student on that unit, the most common errors of each student as well as those of the whole class, etc. All these data will help him/her to recover at the next meeting in class the arguments on which most of the students have failed, as well as to better set up the lesson for the future. Also, the system will allow the teacher to access several automatically generated learning indicators on each student as well as on the whole class. Last but not least, the students can access a wiki maintained by the teachers to recover the questions they failed, in order to have a sharper recover of the missing information.

The rest of the paper is organized as follows: the next section contains a brief literature survey on *educational gaming* in *formative assessment*; section 3 describes the online testing system's architecture and interface; lastly, some final remarks and a brief discussion on future work conclude the paper.

2 Related Work

Even though *educational gaming* has been frequently associated to assessment activities, we have not found any work in literature which uses the feedback obtained from the games to improve teaching quality. Rich data collection during game activities is performed in the system presented by Wideman et al. [9]. To elaborate, it can remotely and unobtrusively record screen activity during game play together with synchronized audio of player discussion. The system has been used to gain useful insights concerning the students' behavior during the playing activities, in order to better comprehend their level of engagement and their capacity of cooperating.

By surveying the current literature, we found examples of software games or experiments aimed at evaluating learners or making them learn through online tests. Wang [8] has compared the impact of *formative assessment* through a quiz-like game and through the classical web-based online testing interfaces, obtaining as a result that the game enhances the participation of the students to the learning activities.

Ramani et al. [7] describe a system to administer online tests *clothed* as computer games. Following the principle that users engage not only in playing, but also in the design of games (in fact video game companies encourage *modders*, those users who modify the games), they also allow students to create questions. In this case, their performance is measured through *Item Analysis* and best score is assigned to the authors of the best discriminating items).

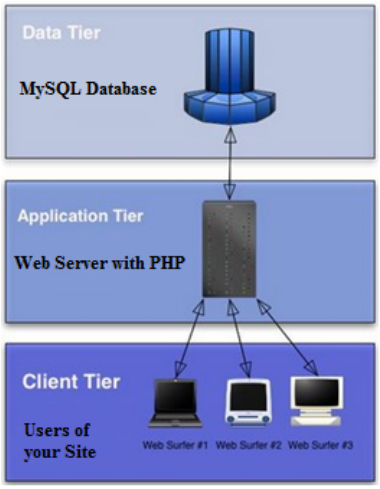


Figure 1. The *PlayLearn*'s system three-tier architecture.

Educational gaming is also spreading in the field of mobile learning. An example of educational video game using test activities accessible through mobile devices is described in [4].

3 The System

PlayLearn is a portal containing a game-based testing system which can be used for training students by creating (the tutor) and taking (the student) online *formative assessment* tests based on multiple choice question items. The system adopts the classical three-tier architecture of the most common PHP-based web applications.

3.1 Architecture

As shown in Figure 1, *PlayLearn* has a tiered architecture. In particular, it is composed of the following three tiers:

- **Client Tier**: this tier is composed of the front-end pages of the teacher's and the student's interfaces. These interfaces are described in detail in the next subsections;
- **Application Tier**: this tier is composed of the PHP modules implementing the business logic of the application. Most of these modules are responsible for processing the user input from the lower tier and for loading and storing data in the upper tier;
- **Data Tier**: this tier is based on a relational database, deployed on a *MySql* database server. *Test* and *Question* entities have a *many-to-many* relationship, meaning that a test is obviously composed of more questions and a question can be presented in more tests.

Macrotags and *Microtags* identify the generic subject of a *Test* and the specific subject of a *Question*, respectively.

The application is fully accessible with a Web Browser. Navigation is facilitated across the simple interfaces based on menus and navigation bars. No browser plug-in installations are needed, except for the Java runtime to run applets. It is worth noting that the system has been tested on recent versions of the most common browsers (i.e., *Internet Explorer*, *Mozilla Firefox* and *Google Chrome*).

3.2 The Student's Front-End

The student's interface has the purpose of administering the formative tests to the student through a simple and captivating video game. After a login interface, the student can directly access the game.

The video game interface requires the student to personify a character, whose objective is to reach a destination located at the end of the game stage. In order to do this, the character has to run and jump through the game stages. Several diamonds are placed on the path to the destination. A diamond can be caught by the student only by answering a question. Once the destination has been reached, the final score obtained by the student is given by the number of the diamonds gathered on the path.

In order to be formative, the questions administered to the student during the progress of the game adopt the following strategies:

- *Options pruning*: if a student gives a wrong response to a question, a second chance to correctly answer the question is given. The previously selected *distractor* cannot be selected anymore. Half the score for the correct response is given in case of right response given at the second chance.
- *Revising*: at the end of the game, a summary of the performance is given. Besides the score and the time needed to complete the game, all of the questions with the given response and the indication of the right option are shown. In the case of a wrong response, a link to the section of the learning material related to the question is shown too.

The difficulty level of the game has been set in order to allow every student to reach the destination without shouldering his/her cognitive and articulatory skills too much. Several game stages have been designed in order to avoid boring repetitive tasks across subsequent game sessions. A screenshot of the interface of the video game is shown in Figure 2. The screenshot shows the game interface while a question is being answered by the student.

At the end of the game, a list of the best five performances is shown, in order to stimulate the learners to improve their outcomes. The limited number of results is due



Figure 2. A Screenshot of the *PlayLearn*'s video game interface.

to the will of not discouraging the less trained students from playing again. Furthermore, a complete history the student's performances, reporting all of the given responses, on his/her previous tests is available, in order to allow the student to revise anytime his/her gaps.

3.3 The Teacher's Front-End

Among other functionalities, the teacher's interface enables the operations of creating questions and tests and of visualizing the reports.

The teacher can manage the lecture material, which is organized in a wiki-based portal. The contents are developed by teachers expert in a given subject, who collaborate to the editing of the articles, basing the content on the textbook. The benefit of using a wiki is twofold: on one hand, the notions to convey to the students are kept down to the strictly necessary to achieve the target for which they have been conceived; On the other hand, it is easy to refer to specific sections of the subjects on which the questions are based, thus correctly advising the students in case of a wrong response.

The report section is rich with information and fit out of charts (see Figure 3) and tables. The teacher can have a complete and deep control over the performance of the class and the learners even on a single subject, and over the effectiveness of the authored resources. To elaborate, the following data are available:

- performance and progress over time of the whole class;
- performance and progress over time of a single student;
- performance and progress over time of the whole class and of a single student on a given subject;

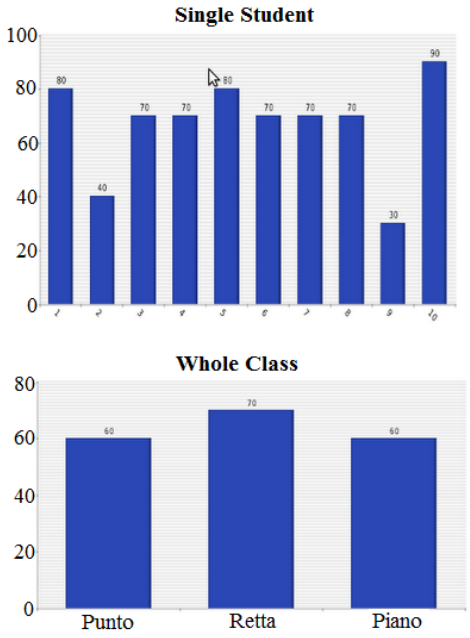


Figure 3. *PlayLearn*'**s Test Report: performance and progress over time of a single student (up) and of the whole class over given subjects (down).**

In *formative* assessment, the main steps of the teaching process are the following:

1. The teacher delivers a lesson on a given topic and schedules a test;
2. the students practice with the formative test;
3. the teacher checks the results and decides the most suitable learning path.

The above methodology requires a relatively long time in monitoring the effectiveness of the teaching due to the large amount of time spent on test preparation, and the correction of its outcomes. *PlayLearn* enables the bringing of the above steps in a real-time continuous process. The benefits are in the following points:

- the tests can be prepared in a relatively short time, since the questions to select for a test are already available in the database;
- test outcomes are immediately available;
- the tests are continuously and regularly taken by the students, since they are motivated by the pleasure of the game.

4 Conclusion

In this paper we have presented *PlayLearn*, a portal intended to exploit *educational gaming* on the long-run as a means for teachers to real-time monitor students' learning, in order to improve the learning process. The usefulness of the system is intrinsic in the necessity of the tutors of checking the effectiveness of their teaching. Unfortunately, this is not possible in the traditional teaching, since the teacher has to wait for the results of a test to have an insight on the level of learning of single students and of the whole class. By exploiting the high level of engagement of the students in *educational gaming* activities, *PlayLearn* allows the tutor to have this kind of feedback in real-time, thus immediately deciding the most appropriate action to improve learning process, all in a virtuous circle.

5 Acknowledgement

This research was partially supported by FARO Project, founded by Universita' degli Studi di Napoli Federico II and Compagnia San Paolo.

References

[1] L. A. Annetta, J. Minogue, S. Y. Holmes, and M.-T. Cheng. Investigating the impact of video games on high school students' engagement and learning about genetics. *Comput. Educ.*, 53(1):74–85, 2009.

[2] S. Bennett, K. Maton, and L. Kervin. The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, 39(5):775–786, September 2008.

[3] P. Frignani and V. Bonazza. *Le prove oggettive di profitto. Strumenti docimologici per l'insegnante*. Carocci, 2003.

[4] P. Lalos, F. Lazarinis, and D. Kanellopoulos. E-snakes and ladders; a hypermedia educational environment for portable devices. *Int. J. Mob. Learn. Organ.*, 3(2):107–127, 2009.

[5] A. Murano, S. Cuomo, and B. D'Aniello. An interdisciplinary project integrating natural science, mathematics and computer science. *MASAUM Journal of Basic and Applied Sciences*, 1(2), 2009.

[6] M. Prensky. Digital natives, digital immigrants. *On the Horizon*, 9(5), 2001.

[7] S. Ramani, V. Sirigiri, N. L. Panigrahi, and S. Sabharwal. Games as skins for online tests. In *DIGITEL '08: Proceedings of the 2008 Second IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, pages 90–92, Washington, DC, USA, 2008. IEEE Computer Society.

[8] T.-H. Wang. Web-based quiz-game-like formative assessment: Development and evaluation. *Comput. Educ.*, 51(3):1247–1263, 2008.

[9] H. H. Wideman, R. D. Owston, C. Brown, A. Kushniruk, F. Ho, and K. C. Pitts. Unpacking the potential of educational gaming: A new tool for gaming research. *Simul. Gaming*, 38(1):10–30, 2007.

Opencast, the Opencast Community and Opencast Matterhorn – A global community-driven project to build a management system for academic video

Todd Jensen[1], Olaf A. Schulte[2]

[1]University of Nebraska-Lincoln, Information Services Instructional Technology
Lincoln, USA
tjensen2@unl.edu

[2]ETH Zürich, ITS-Multimedia Services
Zurich, Switzerland
schulte@id.ethz.ch

Executive Summary

Since its formation in 2007, Opencast has developed to become a truly global community around academic video and all areas of interest related to it. The major endeavour to have emerged from this community is the Opencast Matterhorn Project to be presented here. Matterhorn is a community-driven collaboration to develop an end-to-end, Open Source solution that supports the scheduling, capture, managing, encoding and delivery of educational audio and video content and the engagement of users with that content.

1. Background

Spurred by the development of "podcasting" technology, the increased quantity, quality and use of lecture recording have highlighted video management as a strategic imperative for universities in years to come. Founded in 2007, the Opencast Community is a global community addressing all facets of this domain, thus providing a framework for institutions to look for guidance, best practice and exchange of experience. The Opencast Community also supports a number of projects with the overall goal of facilitating the exchange of audiovisual content.

2. Opencast and the Opencast Community

In 2007, UC Berkeley's Educational Technology Services (ETS) initiated talks with various American and European universities about the status quo in lecture recording. The group soon called itself the Opencast Community to promote the idea of open video in academia - with openness referring to the technology used and the content being created. Today, the community ties approximately 250 institutions and more than 400 individuals.

The exchange of the audiovisual content created at academic institutions is an overall goal the community shares. The community also provides a framework for institutions, individuals, and companies to discuss matters relevant in the context of video production and dissemination, with a focus on lecture recording. Topics typically discussed on the Opencast mailing list are technical issues, organizational approaches and policies, legal questions or didactical aspects of the matter at hand. The project website provides resources for institutions looking to set up or adjust their policy with respect to the management of educational video by displaying best practices and providing a forum for institutions to describe their approach to lecture recording.

The Opencast Community also gave rise to a number of projects: The metadata group discusses ways of harmonizing various metadata schemes in order to increase the searchability of video resources and to foster the exchange of content. However, among these projects, the most important cooperation is certainly the Opencast Matterhorn (Build) Project, a community source project funded by the Mellon foundation and the Hewlett foundation which started in July 2009. Opencast Matterhorn is the system that includes all the components needed to support the full content lifecycle.

3. Opencast Matterhorn – The Build Project

In 2008, the core of the Opencast Community consisted mainly of universities who had already developed solutions for the management of lecture recordings and other audiovisual objects. However, the evaluation of these programs had shown that none of the systems provided by itself the range of functions that the universities required. Some of them incorporated proprietary technology (Podcast Producer), some were limited with respect to the ability to capture and recording technologies, some were unable to scale due to manual intervention needed.

To fill the gap with requirements the community had collected at the same time for a video management and lecture recording system, Opencast Matterhorn was launched out of the Opencast Community as a community source project. Matterhorn is to become an open source management system for academic video, mainly to organize, record, handle and distribute lecture recordings, providing users with tools to engage with the resulting rich medium beyond the mere consumption.

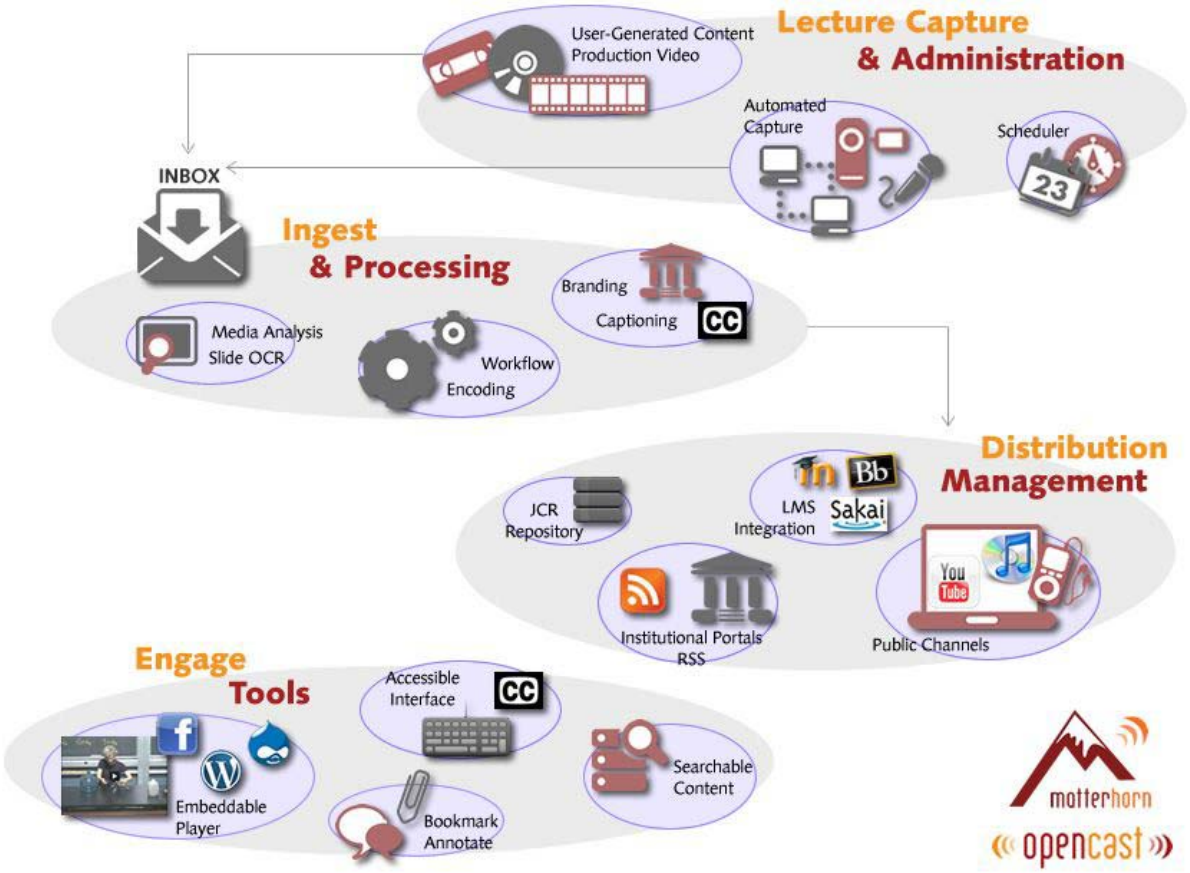


Figure 1: Opencast Matterhorn Architecture

Matterhorn considers the lecture hall as the main source of video. The project partners therefore developed capture agent technology to be integrated into the classroom to record audio, video, and the presenter's VGA signal. These capture agents will be controlled with a scheduler for re-occurring events especially. However, Matterhorn can also ingest video from other sources (customized recordings, films, user-generated content). The processing services allow for branding of videos, generating or adding closed captions with respect to accessibility regulations and of course encoding the video in accordance with the distribution channels selected. In addition, the media from the presentation (VGA) is analyzed and indexed for isochronic metadata. Organized along the video timeline, this allows for pin-point access to the

video based on slide content: For example, students can search that part of the video where "exam" was mentioned on the slide.

Distribution can take many forms with Matterhorn: For certain, the popular channels like YouTube EDU and iTunes (and iTunes U) can be served (or restricted), just like local or institutional portals and repositories. In addition, services are planned to distribute to popular Learning Management Systems (LMS) / Virtual Learning Environments (VLE) with the intention of providing full integration (exchange of metadata, authorization) with these if they prove popular.

Finally, the distributive end also comprises the engage tools. Their function is to engage the user with the video for a more interactive learning experience in an accessible way.

The overall architecture of Matterhorn is service-oriented, allowing users to take advantage of selected features only, should they so wish, thus customizing the product for their needs. Also, the structure allows for easy integration of other services, enabling integration with existing campus infrastructure.

4. International cooperation, inclusiveness and commitment

The Opencast Community, just like Opencast Matterhorn, are inclusive in that we welcome contributions and support from any institution or individual dealing with academic video. These contributions can take many forms, from testing to reviewing code or actually contributing towards the further development of Opencast Matterhorn by aligning existing technologies. The community website at http://www.opencastproject.org/ has all the information on this.

Sakai 3: the Co-evolution of Product and Community

Clay Fenlason
Director of Educational Technology, Center for the Enhancement of Teaching and Learning ,Georgia Institute of Technology, Atlanta Georgia

***Abstract*— The Sakai Community is on the cusp of a new generation of product, Sakai 3. The first early adopter institutions will begin to deploy Sakai 3 as soon as July of 2010, and the first full-featured community release is expected in the middle of 2011. The changes involved are not just new features; rathey, they have broad implications for how the Sakai community will grow and organize itself in the delivery of its collaborative software.**

***Keywords*—Sakai, E-Learning, Distance Learning.**

Introduction

Sakai 3 has been written from the ground up over the course of the last 18 months, and it clearly represents a number of significant changes for the software. To list just a few: breaking down the boundaries of the course site, the exploration of social networking in academic settings, the core 'everything is content' concept, and a new priority placed on the user experience and front-end technologies. At the same time Sakai 3 represents a set of deep changes to how the community of contributors organize to get their work done. Sakai has long insisted that it is not simply a product; it is also a community of practice. In this spirit, it's worth examining Sakai 3 as a co-evolution of product and community.

The Primacy of the User Experience

Among the more striking changes in web applications in recent years has been the explosion of richer and more fluid interfaces, enabled in large part by AJAX technologies. An architectural theme of these technologies is the offloading of much of the processing onto the client (the computer running the browser). A fringe benefit for productivity is that the "front end" (the user interface) can be coded and function almost independently of the "back end" services. By embracing these technologies in Sakai 3 productive work can be accomplished without needing to know the Java programming language. Those with sound

HTML and CSS experience can work effectively, and thus the technical difficulty for doing Sakai work is significantly reduced. This in turn represents an opportunity to open doors onto a wider developer community which has its energies focused on an improved user experience.

The leading contributors to Sakai 3 have explicitly committed themselves to a design-led pattern of software development. It's only with such a process, where a coherent user experience is the driving force for all that supports it, that we can achieve a fully satisfying result. The technical architecture brings in the power to implement sophisticated designs, while also allowing us to mock up, prototype, discard and reinvent new interfaces all the more quickly. Put simply, interfaces are lighter-weight things, as measured by raw effort. The end result should be that the arena of interaction design and front end development should come to rival in the Sakai community the prominence of developing Java services, and by extension we should be more often focused on the user's needs and not just the software engineering problems behind them.

Openness

Openness has long been a strong theme in the Sakai community, but Sakai has nevertheless at times been more focused internally. Most of Sakai's functionality, for example, was written specifically for Sakai 2, partly because there weren't other mature open source offerings that could satisfy our needs at the time: we were froced to make our own. That's no longer the case, and one of the striking things about Sakai 3 is how strongly it depends upon (and participates with) a set of external open source communities - many of which are not in higher education.

This broader partnership with the open source world is not accidental. We've long recognized that many of our needs have a large degree of overlap with more generic functionality (eg. file-sharing, communication, content management), and academic contexts often simply supply them with an extra twist or additional feature. It makes sense for our community to focus its resources on the narrow slice of the problem space that really is specific to academia, our core mission. The Sakai 3 effort makes this focus on our core mission manifest in the code: the Sakai-specific code is a lean fraction of the overall codebase, far less than there is in Sakai 2. The resulting expectation is that we can spend more of our time on what we know best.

Using code from elsewhere (and participating in other open source communities) is one thing, but Sakai 3 also aims for greater openness as a product through interoperation with external systems and services. Widgets backed by services from elsewhere should plug into the same framework, and vice-versa - Sakai data and widgets should be easily represented

in other portals and containers. Our adoption of standards like OpenSocial and IMS LTI encourage and promote this permeability of the application.

Finally, we believe we are opening the doors onto an entire new arena of development. Lowering the difficulty for front end developers is a laudable goal already, but we expect the impact to include the potential to greatly increase the size of our development community. The lower level of required technical skill should allow more of our end-users themselves - teachers and students - to create and use their own widgets that may work against published Sakai services in novel ways. The promise is nothing less than the crowdsourcing of Sakai development.

Innovation as Infrastructure

There have been many innovative educators who have bemoaned the lack of flexibility in LMS offerings, Sakai included. They want the freedom to tailor the functionality to a practice which may vary with the discipline and style of teaching. At the same time a majority of instructors have simple, stock needs, and the technology should not force them to paint each course experience anew from a blank canvas. How can we accommodate innovation without introducing new difficulties and disruption? Even more, how can we capture the successes of innovators and fold them back into an infrastructure everyone can benefit from?

Sakai 3 finds a resolution to this tension in the template. Sakai 3 has a very freeform editing environment that allows for embeddable regions of functionality (widgets) as well as text and embedded media. But it also allows you to save a snapshot of a page or an entire site for later revision or reuse. New forms of workspace, as they prove successful, can be captured.

The community implications for this functionality can be found in the simple but powerful ability to share templates. Templates themselves, like most everything in Sakai 3, are a form of content, and so can be revised, shared and collected to such an extent that patterns for the most successful pages and workspaces can serve as a library of best practice that anyone can use. You don't have to create your own brilliant course site from scratch when you can just use a successful pattern from a colleague. Over time we expect the cream will rise to the top, and today's innovations will organically become tomorrow's infrastructure. Sakai might quite literally become a community of shared practice.

The Sociality of Knowing

The recent advent of social networking software is not best understood as a recent trend. It is in fact technology catching up to where people have always been: interpersonal

relationships are a powerful shaping force in how we discover and understand the world. This is as apparent in academia as elsewhere, and it's been built into practices of mentorship, peer review, and scholarly communities from the beginning. Our users are already networked before we give them a web application to support it. Even more: if social networking weren't already a success on the web, we should be trying to invent it.

But just as there was social networking long before there was a social networking app, so also do we need to pay close attention to the society we're trying to support, and not simply repeat the patterns of a Facebook or Ning uncritically. Academic cultures involve different kinds of relationships and roles, often organized around some shared work rather than the relationship for its own sake, and so there is good reason to go back to basics and design for the fundamental user need in academic contexts.

This return to the user-centered fundamentals of the need we're trying to serve also represents a shift for how the Sakai community pursues its work. It was at one time enough of an advance to cobble together generic file-sharing tools, discussion boards, wikis, etc. under an umbrella of institutional administrative data. We are now at a stage where we need to more often step back from the latest tools on the wider web, think deeply about the relevance to our academic users, design around the demonstrated user need, and the technology should be shaped by all this. That is 'design-led' in a nutshell.

Summary

Dan Behlendorf of the Apache Foundation shared a pithy phrase during his keynote address at an earlier Sakai conference: "good software is an emergent property of a healthy community." In that spirit, we can also view the changes in the Sakai 3 product through the lens of the community.

In fact, we might say that the full implications of Sakai 3 won't be fully understood by looking at the product alone.I'm happy to report that the Sakai community is maturing along with its product, working to support its core mission in a simpler yet deeper and more focused way, while at the same time opening itself up for growth and a wider set of partnerships that an open source license facilitates.

Towards Integration of Data Base Management System for Open-source Course Management Systems

Yuki TERAWAKI[1], Yuichi TAKAHASHI[2], Yuji TOKIWA[1], Yasushi KODAMA[1], Kazuo YANA[1]
[1]Research Center for Computing and Multimedia Studies, Hosei University, Tokyo, Japan
[2]foreach ltd., Tokyo, Japan

Abstract—**This paper describes an integration of different Relational Data Base System (RDBMS) of two Course Management Systems (CMS) called Sakai and CFIVE.**
First, we consider problems in the support of CMS campus-wide, and then two CMSs are reviewed. As a result, these CMS are possible to solve our issues. Furthermore, the integration of different RDBMS on two CMSs is discussed, and the modifications modules for integration are described.

Keywords— **Contents Management System (CMS), Learning Management System (LMS), RDBMS, CFIVE, Sakai.**

I. INTRODUCTION

Needless to say, CMS and Learning Management System (LMS) are used extensively by higher education in Japan. In Hosei University, CMS has been used since 2007 by users campus-wide. We have been operating and supporting the CMS for users campus-wide. Currently twelve thousand courses are registered with the CMS. The CMS are used by almost 90 percent of all students and teachers in HOSEI University. Various issues need to be considered from past experience.

- Large-scale Application
  In HOSEI University, there are many classes that register hundreds of students. In such classes, utilizing CMS is especially effective. Thus, CMS needs a lot of hardware capability and efficiency-enhancing architecture. For example, In order to avoid problem running the CMS using several Application Servers, we have to consider problems such as: - avoid storing a data in a program or decide whether or not to use Session Replication.

- Sudden/ Short-term High-user Demand on CMS
  As deadline approaches, if a lecturer put a time limit on the report, many users simultaneously log in to CMS to submit reports. During the period of examinations, deadlines for submission of reports tend to be concentrated on specific dates. Another problem is simultaneously use students the function of self-registration to register for a course CMS at the beginning of a new term. Therefore, it is essential to consider the problem of sudden short-term high-user demand to CMS not only users of a particular class but also several users from different classes accessing simultaneously.

- Interoperability
  The authentification check of CMS needs to work together with the authentification system constructed by HOSEI University.
  Also when a course is registered in the CMS, course-IDs are imported from the system of academic affairs. Moreover, the data of the attendance management system relate to the CMS. For example, the data of the attendance management system is not only displayed on the CMS but also attendance in the initial class is used to register learners to a course on CMS before determinate registration.

- Customization
  The function of CMS should be redesigned according to a variety of practice in the university. A Functionality requirements for the CMS varies from class to class. For example, required function for Bulletin Board System (BBS) might be the difference to learn either individually or in groups.

- User Support
  To support all users (Lecturers, Learners and Staff), it is preferable that the CMS has a better minimal function than an improperly-matured function. To support thousands of users, our CMS should be kept as simple as possible.
  However even if the functions of CMSs are appropriate, it is difficult to fulfill the needs of faculty staff. Therefore we have to develop tools to satisfy this need. This means that the Application Programming Interface (API) of the CMS should be open. Moreover, it is preferable if the source code of the CMS is also opened.

The current CMS is difficult to redesign due to license and technical restriction because a commercial off-the-shelf system (COTS) CMS is used. It is well known that vendors of COTS try to meet the needs of the marketplace instead of satisfying the requirements of a particular organization [1]. Therefore, to solve these problems, we reviewed a number of open-source CMSs. Consequently, we consider that two open-source CMSs, Sakai [6] and The Common Factory for Inspiration and Value in Education (CFIVE) [5] are the best

for our university. In this paper, in order to use the data of Sakai and CFIVE seamlessly, we will try to integrate different RDBMS on these CMSs.

This paper is organized as follows. Section 2 outlines the candidate CMSs, Sakai and CFIVE. We discuss why we consider it is necessary to combine Sakai and CFIVE. Section 3 describes the developed module for integration of different RDBMSs. In section 4 concludes this work.

II. Candidate CMS

Since the early 2000s, against COTS CMS, open-source CMSs have been developed. In Europe and the United States, obviously, Sakai is one of the leading CMSs.

A. Sakai

Sakai consolidated past practices of systems for education and learning such as Open Knowledge Initiative (O.K.I.) by Massachusetts Institute of Technology (MIT) and CompreHensive collaborativE Framework (CHEF) by the University of Michigan. Sakai is licensed under the terms of the Educational Community License (ECL). The ECL is a minor variant of the Apache License. Sakai uses existing open-source software such as MySQL, Tomcat, and Hibernate. The major running environment of Sakai is shown in Table1.

Several universities in Europe and the United States are involved in the Sakai community. Thus, we know about educational practice in European and American universities.

We need to properly reflect international trends for learning environment. However, Sakai is not a perfect learning environment when Sakai supports Japanese specific educational practice because Sakai was developed mainly in Europe and the United States.

- Japan Specific Educational Practice

  In Japan, a smart card is used to provide the service appropriate to the occasion not for only the purpose of management. For example, the data of the attendance management system is used for interactive lectures. The learners in the classroom are displayed on the profile function of the CMS, and then the lecturer asks learners by name. If the learner has filled out the profile function with his personal information, the lecturer may be give the information related to the learner's profile. This practice via smart card is becoming popular in Japan [2]. There is not an attendance display system among the tools of Sakai, such as one where learners in the classroom are displayed on the CMS using a smart card.

- The Problems of School Year and Order of Names

  Sakai is internationalized in the sense that messages in local languages including Japanese can be shown. However, the FAQs of Sakai are not enough, for beginners to learn how to use Sakai. In addition, the school year of most Japanese universities is different from that assumed in Sakai. For example, in Sakai, academic terms such as 'SPRING 2010', SPRING SUMMER 2010', 'SUMMER 2010' are automatically set up. Unfortunately, the school year of most Japanese universities is divided into two semesters.

  Moreover, names can be tricky, especially when two different cultures come together. In Japan, family names usually come first, but in Europe and the United States, it is the opposite. Unfortunately, these problems remain unsolved [7].

To avoid these problems, we have to look at the availability of CMSs developed in Japan. An open-source CMS has been developed by several Japanese universities. In terms of past results and licensing, CFIVE is the best for our university to complement the issue to adapt Sakai as the campus-wide CMS.

TABLE I
Major running environment of Sakai

Item	OSS
Java	Java (jdk) 1.5
Servlet	Tomcat 5.5.X
RDBMS	MySQL 5.1 or Oracle etc.
Software Project Management Tool	Maven 2.1.0 or later

TABLE II
Major running environment of CFIVE

Item	OSS
Java	Java (jdk) 1.6
Servlet	Tomcat 5.5.X
RDBMS	PostgreSQL 8.3
Software Project Management Tool	Maven 2.0.9 or later

B. CFIVE

CFIVE was developed by the University of Tokyo [3]. Everybody can download source code from the website of CFIVE under the license of General Public License (GPL version 2). CFIVE uses existing open-source software such as PostgreSQL, Tomcat, and Maven. CFIVE is designed with a simplified core, pluggable and extendable architecture. These keep the code proper of CFIVE as small as possible, and the maintenance is made easier. The major running environment of CFIVE is shown in Table 2.

CFIVE has been providing a service to the University of Tokyo campus-wide since 2004. CFIVE provides modules for the system of academic affairs and a smart card reader system [4]. Also, the CFIVE project continues to improve

operationally and is becoming more convenient based on users' requirements. It is thought CFIVE solves the issue of Japan's specific educational practice, school year and other issues. Therefore, to solve the problem of Section 1 and Section 2, we decided to build Sakai and CFIVE as the teaching, learning and academic collaboration platform.

III. Towards Integration of Different RDBMSs

When these CMSs are implemented, the problem of integrity of different RDBMSs is addressed.

As shown by Table 1 and Table 2, Sakai and CFIVE use different RDBMSs. In order to make maintenance easy and to leverage data between Sakai and CFIVE, we decided to integrate the RDBMSs of Sakai and CFIVE.

A. CFIVE-Sakai Integration

A unified running environment is ideal, in order to make maintenance easy. The term "running environment" includes application-container, RDBMS, programming languages and development tools. CFIVE requires Java (jdk) 1.6, Maven 2.0.9 or later, Tomcat5.5.X and PostgreSQL8.3. Sakai requires Java (jdk) 1.5, Maven 2.1.0 or later, Tomcat 5.5.X and multiple RDBMSs. Sakai supports MySQL5.1, Oracle and DB2. We decided to use the environment of CFIVE, and made Sakai integrate with that. Sakai requires jdk1.5. Some of the JSF codes cannot be compiled using jdk1.6. Moreover, jdk1.5 is already obsolete; and Sakai on jdk1.6 is mostly working; so, we ignored the problem in this work.

B. Sakai- PostgreSQL Integration

In order to make Sakai work with a non-supported RDBMS it is necessary to make several modifications: -

1. Add data definition files (e.g. create table/index SQL statements),
2. Add classes to absorb differences among RDBMSs,
3. Modify classes that have mapping between vendor names and classes above,
4. Modify configuration files.

Sakai consists of two types of components. One is a core that is a common framework; the other consists of tools that provide concrete functions. The core components such as sakai-kernel and sakai-entitybroker, are shared in a container, and used by tool components. The tool components are provided as sakai-2.7.x. Sakai mostly accesses database tables using an Object Relational Mapper (Hibernate), and the data definitions are generated by a Hibernate auto-ddl function. However, some foundation parts use a Vanilla JDBC Driver, and the data definitions were written by hand (we call them Legacy Tables). Thus, the access logic and the data definitions should be rewritten in concert with each RDBMS. These previously expressed changes were applied to saka-kernel, sakai-entitybroker and sakai-2.7.x. The list of modified files is shown in figure 1.

C. Data Definitions

PostgreSQL does not have data types, such as tinyint, mediumtext, longtext, datetime and longblob of MySQL. They should be mapped to available types in data definition files. PostgreSQL has different syntaxes of SQL such as an auto increment feature or alter table from other RDBMSs. They also should be suit data definition files. The data type mappings (MySQL to PostgreSQL) are as follows: -

tinyint -> smallint
mediumtext -> text
longtext -> text
datetime -> date
longblob -> blob

There are differences between the syntaxes of MySQL and PostgreSQL. In order to implement the syntaxes of PostgreSQL in Sakai, the following changes are needed: -

INTEGER AUTO_INCREMENT -> serial
CREATE INDEX ... ASC/DESC -> remove ASC/DESC
CURRENT_TIMESTAMP () ->
CURRENT_TIMESTAMP

D. Classes to Absorb Differences among RDBMSs

Classes that implement *ServiceSql interfaces should be created in concert with an RDBMS. They indicate RDBMS specific values such as error codes or notation of Boolean. They also provide SQL statements for accessing Legacy Tables. Some additional classes such as PgSqlSharedPersistenceManager should be created, for same reason.

E. Classes that Have Mapping between Vendor Names and Classes

RepositoryBuilder and BaseDb* classes have a mapping that is an example of a java.util.HashMap class. The keys are vendor names that are used in sakai.properties, and the values are concrete classes. New mapping entries corresponding to PostgerSQL should be added, in order to use the RDBMS specific classes above. Also, the definition files of the Dynamic Injection Container (Spring) should be modified.

F. Configuration Files

The PostgreSQL configuration should be put into a global configuration named sakai.properties that is located in the sakai directory in the application container's root directory (e.g. /usr/java/tomcat/sakai).

PostgreSQL settings – make sure to alter as appropriate

vendor@org.sakaiproject.db.api.SqlService=pgsql
driverClassName@javax.sql.BaseDataSource=org.postgresql.Driver
hibernate.dialect=org.hibernate.dialect.PostgreSQLDialect
url@javax.sql.BaseDataSource=jdbc:postgresql://127.0.0.1:5432/sakai
validationQuery@javax.sql.BaseDataSource=select 1
defaultTransactionIsolationString@javax.sql.BaseDataSource=TRANSACTION_READ_COMMITTED

At this stage, the modifications are still experimental. Thus, they are not ready for real operations. We will improve them, and contribute them to the Sakai community in the future.

Fig. 1. Modification files

```
sakai-kernel
    api
        src/main/java/org/sakaiproject/content/hbm/LockManager.hbm.xml
    kernel-component
        src/main/webapp/WEB-INF/*-components.xml
            (bean definitions of org/sakaiproject/{authz,cluster,content,db,event,site,user}/impl/*SqlPgSql.java)
    kernel-impl
        src/main/java/org/sakaiproject/{authz,cluster,content,db,event,site,user}/impl/*SqlPgSql.java
        src/main/java/org/sakaiproject/jcr/jackrabbit/RepositoryBuilder.java
            (mapping between vendor and subclass of PersistenceManager)
        src/main/java/org/sakaiproject/jcr/jackrabbit/persistance/PgSqlSharedPersistenceManager.java
        src/main/sql/pgsql/*.sql
    kernel-util
        src/main/java/org/sakaiproject/util/BaseDb*.java
        src/main/java/org/sakaiproject/util/*SqlPgSql.java

sakai-entity-broker
    impl
        src/ddl/pgsql/*.sql
    pack
        src/webapp/WEB-INF/spring-jdbc.xml

sakai-2.7.x
    ./assignment/assignment-impl/impl/src/sql/pgsql/sakai_assignment.sql
    ./metaobj/metaobj-api/api/src/sql/pgsql/metaobj.sql
    ./announcement/announcement-impl/impl/src/sql/pgsql/sakai_announcement.sql
    ./content/content-impl-providers/impl/src/sql/pgsql/sakai_event_delay.sql
    ./gradebook/app/business/src/sql/pgsql/category_weight.sql
    ./gradebook/app/business/src/sql/pgsql/sakai_gradebook.sql
    ./citations/citations-impl/impl/src/sql/pgsql/sakai_citation.sql
    ./help/help-component/src/sql/pgsql/sakai_help.sql
    ./mailarchive/mailarchive-impl/impl/src/sql/pgsql/sakai_mailarchive.sql
    ./rwiki/rwiki-impl/impl/src/sql/pgsql/sakai_rwiki.sql
    ./syllabus/syllabus-app/src/sql/pgsql/sakai_syllabus.sql
    ./presence/presence-impl/impl/src/sql/pgsql/sakai_presence.sql
    ./osp/sql/pgsql/osp.sql
    ./calendar/calendar-impl/impl/src/sql/pgsql/sakai_calendar.sql
    ./chat/chat-impl/impl/src/sql/pgsql/sakai_chat.sql
```

IV. CONCLUSION

In this work, based on past experience, we explained what we require of future CMSs. Moreover, we introduced the candidate CMSs, Sakai and CFIVE. Originally, CFIVE was developed using PostgreSQL. Thus, in order to use data of CFIVE under Sakai seamlessly and to make maintenance easy, we tried to implement Sakai using PostgreSQL. Currently, Sakai supports multiple RDBMSs such as MySQL, Oracle and DB2 but PostgreSQL is not supported by Sakai.

The share of open source RDBMSs is as follows; MySQL had 33% and, PostgreSQL had 9% in 2004. These were the global figures. However, in a survey done in Japan in 2008 [8], the share of RDBMS was PostgreSQL - 53.0%, MySQL - 51.4%. Moreover, when the RDBMS was switched from commercial to open-source, in Japan 76.2% of users selected PostgreSQL (the percentage of people who chose MySQL was 45.2%). This survey shows that many engineers use PostgreSQL in Japan. A lot of engineers in Japan are familiar with the SQL statement of PostgreSQL. If PostgreSQL is supported by Sakai, users in Japan will feel more comfortable with Sakai.

In addition, when we modified program files in Sakai, there seems to have been some attempt to support PostgreSQL in some of the program files. The Sakai community also hopes to use PostgreSQL. Our study will bring advantages not only to our university but also the Sakai community.

REFERENCES

[1] Carina Alves and Anthony Finkelstein, "Challenges in COTS Decision-Making: A Goal-Driven Requirements Engineering Perspective", Proceedings of the 14th international conference on Software Engineering and Knowledge Engineering 2002

[2] Y. Tokiwa, Y. Kodama, K. Yana, "Current Status of Open Source Course Management System – Sakai and Trial for its Internationalization and Scalable Adaptation", Journal of Japan e-Learning Association, Vol.8: pp 23-31, 2008 (in Japanese)

[3] T. Sekiya, Y. Terawaki, Y. Onoue, K. Yamaguchi, "CFIVE: Open source learning management system", Journal of Multimedia Aided Education Research, 1(2): pp73-81, 2005. (in Japanese)

[4] T. Sekiya, T. Sato, Y. Terawaki, Y. Onoue, K. Yamaguchi, "Assessments of Continuous Operation and Development: Centered in Open Source LMS, CFIVE", SSS2006, In IPSJ SIG CE, pp. 339 - 346, 2006. (in Japanese)

[5] CFIVE Project, http://cfive.itc.utokyo.ac.jp, 2010

[6] Sakai Project, "http://sakaiproject.org/", 2010.

[7] ja Sakai community, "http://bugs.ja-sakai.org/", 2010. (in Japanese)

[8] Actual Condition Survey of Business Utilizing Open Source Software,"http://www.ipa.go.jp/software/open/ossc/seika_0803.html", 2010. (in Japanese)

Address of the corresponding author:

Yuki TERAWAKI
Research Center for Computing and Multimedia Studies,
Hosei University
3-7-2 Kajino, Koganei City Tokyo 184-8584
Tel: +81-(0) 423-87-6188
Fax: +81-(0) 3-3368-8165
Email: yuki.terawaki.dc@k.hosei.ac.jp

3D virtual space for a quality e-Learning environment in collaborative study

Yoshiyuki HINO[†], Kazuo YANA[†], Hisato KOBAYASHI[†], Kimiyoshi HAYASHI[†], Yasushi KODAMA[†], Billy PHAM[††]

[†]Hosei University Information Technology Research Center, [††]Hosei University Research Institute, California
Email: it@hosei.ac.jp

Abstract : Although e-Learning has been in active use as a tool that offers diverse opportunity to study in recent years, conventional face-to-face lecture still has some edge on it in stimulating students' appetite for study or in offering an environment for collaborative study. This paper introduces our study on 3Di as SNS in a broad sense to establish a new way of supporting research and educational activities.

Key words: CMS, 3-D virtual space, e-Learning, Education method and assessment, Collaborative study

1. Introduction

Hosei University Information Technology (IT) Research Center (hereinafter referred to as "IT Research Center"), in cooperation with several overseas research organizations [1] such as Hosei University Research Institute, California, have successfully realized real-time interactive distance lectures by connecting each overseas base through the Internet, and been working on the development of e-Learning educational tools and study-support systems etc. which aims to design and construct borderless educational and research environment[(1)].

In this paper, we would like to first introduce the past projects which the IT Research Center has taken on, and then talk about how we built and are utilizing the collaborative study environment as a new educational form, which utilizes 3D virtual space such as 'Second Life' (hereinafter referred to as 'SL') etc. as well as our future vision.

2. Past major projects

2.1 IT Research Center

The IT Research Center has carried out internationally distinctive education and research programs such as offering real-time interactive distance lectures connecting several overseas bases or organizing international conferences etc., and worked on the development of e-Learning educational tools and study-support systems which aim to design and construct borderless educational and research environments[(2) (3)]. The practical operation of various educational and research projects has also been carried out in cooperation with several overseas research bases such as Hosei University Research Institute, California.

We have been carrying out the following programs.

(1) Distance classes connecting overseas bases
(2) Development of base system, multilingual system etc.
(3) Holding international conferences
(4) Organizing and carrying out other open courses and symposia

As for system development, we have been making efforts to develop various systems such as an automatic generation system of digital contents [(4)] or an automatic captioning system in order to improve the understanding of learners whose mother tongue is not English. It has also participated in "Sakai" [(5)] which is involved in international cooperative development activities with many top universities in the U.S and other countries. and worked on development activities as well. We have put into practice education making use of the Course Management System (hereinafter referred to as 'CMS') which utilizes the module developed by Sakai and e-Learning course materials (Figure 1).



Figure 1 e-Learning video on Sakai

2.2 Approach toward e-Learning

(1) International Distance Courses

Since the establishment of Hosei University Research Institute, California in 2000, a wide range of international distance courses have been offered connecting two bases (Tokyo, Japan and California, USA). And after establishing two more bases in London (UK) and Zurich (Switzerland) in 2007, many distance courses connecting Japan and Europe have also been offered.

1. Hosei University currently has three overseas bases in the U.S, U.K and Switzerland.

(2) Developing on-demand educational tools

As previously described, we have been working on the development of educational tools and e-Learning materials to realize ubiquitous study environment which was previously difficult to offer due to 'the time difference and limitation of learning spaces to participate in the courses'.

However, it is quite impossible to achieve high educational effects on all students simply by providing such an environment. In order words, the prerequisite for achieving high educational effects in self-learning by on-demand education is that all students must have the same degree of enthusiasm for learning. A strong sense of purpose and self-discipline is required for continuing to study with keeping their motivation for learning throughout the learning period. Without these, the benefit of being able to study at 'any time and any place' will not be appreciated and could end up as "NOT"-studying at 'any time and any place'. In order to prevent this, we have tried to develop so-called hybrid-type courses which combine the utilization of CMS, real-time face-to-face classes and on-demand classes which will be explained in the next section. (Table 1)

	Real-Time Class meetings			On-Demand Learning		
Date	Class	Topic	Lecturer	class	Topic	Lecturer
25-Sep	1	Course overview Chapter 1,2	Hisato Kobayashi	1	Assistive Technologies: Cutting-edge Technologies in nurse, care, rehabilitation Part 1	Hisato Kobayashi
2-Oct	2	Chapter 1,2 wrap-up Quizz	Hisato Kobayashi	2	Assistive Technologies: Cutting-edge Technologies in nurse, care, rehabilitation Part 2	Hisato Kobayashi
9-Oct	3	Chapter 3,4 Quizz	Hisato Kobayashi	3	Assistive Rehabilitation Robots	Machiel Van der Loos
16-Oct	4	Guest Speaker	Guest Speaker	4	Therapy Rehabilitation Robots	Machiel Van der Loos
23-Oct				5	Clinical-Use Rehabilitation Robots	Machiel Van der Loos
30-Oct	5	Chapter 5,6 Quizz	Hisato Kobayashi	6	Virtual Reality in Rehabilitation	Machiel Van der Loos
6-Nov	6	Chapter 3-6 wrap-up	Machiel Van der Loos			
13-Nov				7	Service Robotic Systems for the needed: Overview	Z. Zenn Bien
20-Nov	7	Chapter 7,8 Quizz	Hisato Kobayashi	8	Wheelchair-based Robot System	Z. Zenn Bien
27-Nov				9	Human-Robot Interaction Technique with special attention to Hand Gesture	Z. Zenn Bien
4-Dec	8	Chapter 9,10 Quizz	Hisato Kobayashi	10	Smart High-tech House as a Large-scale Care Robot	Z. Zenn Bien
11-Dec						
18-Dec	9	Chapter 7-10 wrap-up Q&A Session	Z. Zenn Bien	11	Design for Wellbeing Innovations for People	Lennart Karlsson
25-Dec	10	Chapter 11,12 Quizz	Hisato Kobayashi	12	Remote Controlled Robot to Perform Echocardiography on Distance	Kurt Boman & Mona Olofsson
1-Jan				13	Telemedicine and information services for a patient focused process	Lars-Ake Johansson
8-Jan	11	Chapter 13,14 Quizz	Hisato Kobayashi	14	NeedInn	Tobias Larsson
15-Jan	12	Chapter 11-14 wrap-up	Hisato Kobayashi Andreas Larsson			

Table 1 Hybrid-type class schedule of Welfare Engineering

(3) Utilization of "Sakai"

Since 2006, the Hosei University has introduced commercial CMS and been using it throughout the school. Along with this, the IT Research Center has participated in "Sakai Project" as a part of its research projects in order to contribute to communities in the field of internationalization and development of educational tools. At the same time, the IT Research Center has adapted the Sakai CMS which were customized for and by the IT Research Center and conducted demonstration experiments by using it for its supporting courses, and then used the experiment results as feedback for further development activities.

We won't go into details of the "Sakai" development as it is not the main theme of this paper, but the main reason the IT Research Center has adapted "Sakai" is that using an open-source CMS by object oriented scripting language, Java, is desirable from the viewpoint of scalability with the prospect of using it for the entire school in the future. However, as the members of the community are mainly from Europe and North America, their purpose of the system development does not necessarily match the needs of Japanese Universities. So to apply for our use, customization of the system or development of individual tools were needed. More specifically, in order to achieve high educational effects through e-Learning education in today's Japanese universities, more detailed study management and precise analysis are important. Also, attendance management of students who take classes at remote sites where there are no teachers and interaction with those students cannot be neglected. The IT Research Center has developed such study management and support tools, and conducted demonstrative experiments. And with the feedback received from teachers and students, it has been carrying out educational and research activities with the aim to improve educational effects [(6)].

(4) Class assessment

A part of class assessment results done for the subject (Welfare Engineering) offered in the academic year 2008 are shown below. (Figure 2)

About real-time class:

Q. What was your general impression - ambience toward the class where a teacher was giving a lecture from a distant location?

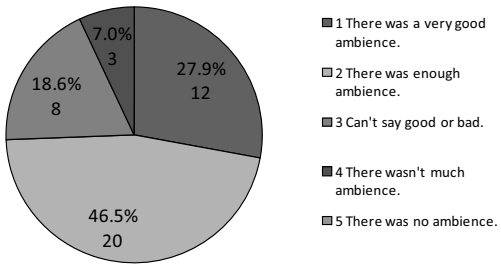


Figure 2.1

Q. Did you get any sense of uneasiness from the class offered in a distance-lecture style?

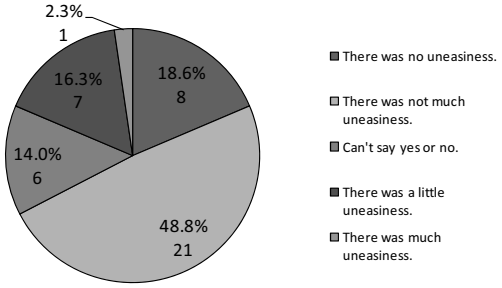


Figure 2.2

About hybrid-type class:

Q. What was your impression of hybrid-type class which combined real-time class and contents class?

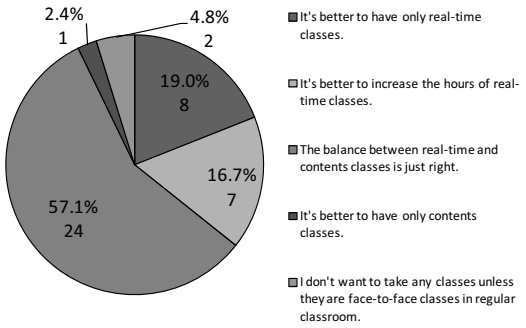


Figure 2.3

About the class in general:

Q. Do you think this class will increase your understanding more compared with regular classroom lessons you have taken before?

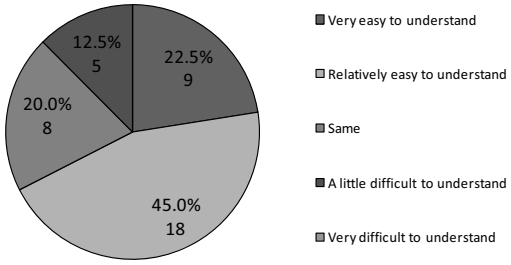


Figure 2.4

Q. Do you think you have enough opportunity for one-to-one Q&A with a teacher compared with regular classroom lessons you have taken before?

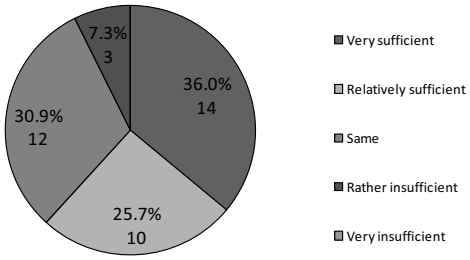


Figure 2.5

The assessment result shows that the hybrid-type classes which combine real-time distance classes and on-demand study have been highly assessed as a whole. This is the result of our repeated trial, error and improvement after conducting numerous numbers of the same style courses since 2005. The assessment result of the survey done in 2008 shows that the balance between "real-time" and "on-demand" classes is just about right, which is the result of having greatly increased the number of real-time classes (about two times of what was offered in 2006) based on the survey result.

As previously described, the IT Research Center has been engaged not only in the development of study-support tools and methods but also in actual educational application. We believe that the success was achieved by a series of efforts aiming to achieve higher educational effect by blending these IT technologies into the field of education.

3. Utilizing 3-D virtual space in educational field

3.1 Background of using 3-D virtual space

Although e-Learning environment has been developed through the introduction of study-support tools and improvement of class style etc. as explained, when compared to conventional face-to-face classes, the improvement of collaborative study such as maintaining an individual learner's enthusiasm for learning or improving interactivity is not yet sufficient. At present, this insufficient area is compensated by increasing the number of face-to-face lectures in hybrid-type courses, but from the viewpoints of original e-Learning or on-demand education, this problem cannot be considered yet as solved. In light of this fact we have focused attention on the utilization of 3-D virtual space such as SL. as a new educational tool to resolve these situations.

We consider the advantage of using 3-D virtual space in the educational field with the following two main points:

(1) Multiple styles of conversation such as one-to-one, one-to-many, become possible, hence interactivity improves.

(2) It is not necessary for learners to gather in a classroom equipped with a large-scale distance lecture facility. Hence, where to take classes is not restricted. And even under this condition, learners can share the learning experience in the same virtual space, which will promote their sense of solidarity and enthusiasm for learning.

3.2 3-D virtual space as SNS

The more activities or mutual communications of participants there are, the more effects can be expected in the 3-D virtual space. As this is nothing but a concept of the Social Networking Service (SNS) which is 'a community-based service to promote and support humane connections', in a broad sense we can interpret 3-D virtual space as SNS. Regarding 3-D virtual space as SNS will open up more efficient methods in e-Learning, which are different form existing study-support tools such as CMS.

3.3 3-D virtual space as CMS and SNS

As for the study-support features of CMS and SNS, it is considered that 'CMS is suitable for formal and direct support, while SNS is for informal and indirect support'.[7] CMS has been used in e-Learning for stimulating the enthusiasm of learners by precisely managing the learning

attitude or progress of their study, but in practical application of education in 3-D virtual space like SL, some facilitation functions such as promoting communication among participants (learners) and constructing collaborative study environment become essential in addition to the above-mentioned management.

Many universities have been openly offering courses to public originally through OCW and recently using YouTube EDU or iTunes U. As a result, formal classes which used to be held at a university campus have become open to the public and lessons which used to be available only to students who paid tuitions have become accessible to those who are not formal students. In this case, the difference between formal students who pay tuitions and other participants are; (1) formal credit earning, and (2) formal guidance given by instructors either directly or through CMS. It is possible to think that universities, by letting external students observe their formal classes in an informal manner, actively disclose the high-quality education they offer and use this opportunity for recruiting future formal students. Similar approaches can be seen with SL. For example, the Harvard Law School offered its formal course subjects on SL. Formal students can take those classes either in real regular classrooms or classrooms on SL. External students can take classes offered on SL together with formal students, but cannot earn credits. Opening formal classes outside of the school with certain restrictions can lead to gaining recognition outside the school and possibly result in indirectly improving educational quality.

We assume the transmission of open knowledge by universities will increase more and more in the future and in such cases, use of CMS and SNS in combination according to the purpose is most important.

4. Previous projects and future views

4.1 Previous projects on 3-D virtual space

In 2007 the IT Research Center built its own campus on SL and has conducted various demonstration experiments towards the possibility of starting courses on the virtual campus in the future.

Since then, the IT Research Center carried out a project which was a further development of the previous projects and which focused more strongly on interaction with the real world. For example the "IEEE RO-MAN 2008" was held in August 2008, with avatars resembling keynote speakers placed in the virtual site on SL and with the audience on SL site communicating interactively with keynote speakers in the real site through text chat and voice chat.

Thanks to the introduction of the avatars, the merging of real and virtual sites and the interactive proceedings of the meeting was made possible [8]. Even so there were still a few problems. When lectures are given to audiences who assemble both in real and virtual sites, the attention of speakers tends to focus more on the audience in the real site. So it is important to have facilitators who connect real sites and virtual sites so that the audience in the virtual site can establish interaction with real-site speakers more naturally. This is similar to what we have experienced in the distance lectures previously conducted using a video conferencing system. In other words, by directly speaking or asking questions to students in remote sites instead of one-way lectures, actual interaction can be achieved, which leads to the stimulation of students' enthusiasm for participating in classes and finally to increasing the level of proficiency.

Previously SL has been mainly used as our demonstration experiment site in 3-D virtual space, but due to the reasons of special viewers being required or of its difficult operation method etc., SL was tough to handle for many people except for existing SL users, hence resulting in a rather low SL use rate. So in 2009, we have tied up with a company to develop a new viewer which merges 2-D and 3-D, aiming to encourage the use of 3-D virtual space in a more natural manner (Figure 3). Thanks to this development, the participants who are only interested in viewing the lecture in a real site can view a 2D Webcasting screen, while those who wish to interact with lecturers or other participants in a virtual site can achieve these interactions through the use of avatars.



Figure 3 Viewer merging 2-D and 3-D

In 2008, we also conducted a project based learning (PBL) course on SL where students simulated the movement of robot arms using a scripting language.

Apart from some technical problems, if we give the assessment of the course which used SL as SNS, the biggest merit of conducting classes on SL can be said that 'informal and indirect' support (instruction) was able to be given as described in 3.3. That is, students can login on SL at anytime and from anywhere they like just like on-demand study and work on their assignments. But the biggest difference from self-learning by the use of one-demand educational tools was that students could consult with a teacher or other students who "happen to" appear on SL and receive advice from them while they were working. We believe that this was highly effective to maintain and improve the motivation of students.

The other noteworthy experimental trial course was the Shanghai Lectures [9] which was lead by Rolf Pfeifer, a Professor of University of Zurich. This is truly a world-scale educational project which utilizes existing distance class style as well as collaborative study environment in 3-D virtual space, with participation by

many universities around the globe.

The IT Research Center hoped to use this participation experience;

- to construct our school's own virtual educational and research environment using open-source 3-D virtual space,
- to build collaborative study environments in international distance education, in the future.

In the ShanghAI Lectures, we utilized 3-D virtual space called 'Project Wonderland' developed by Sun Microsystems, Inc. Open-source 3-D virtual space such as 'Project Wonderland' or 'Open-SIM' has recently prevailed and these 3-D virtual spaces make us possible to;

- build closed environment.
- customize to fit one's needs

The IT Research Center will make use of SL in a more open scenario while working on the virtualization of more closed educational and research environments which is limited to people inside the school.

(2) Transmission of knowledge

In addition, we have also publicly exhibited the precious assets which Hosei University owns and worked on the transmission of information with the use of SL.

The IT Research Center has been actively and widely disclosing to the public the contents of lectures or symposia which they hosted, and recently has been working in cooperation with Hosei University Nogami Memorial Noh Theatre Research Institute on the digitalization of precious informational materials and picture images owned by the Institute, and opened them to the public. In 2008, these projects were further developed to open 3-D virtual museum on SL. The purpose is to further promote Noh Theatre (Noh & Kyogen) by making them accessible to a lot more people in an easier and more friendly manner. As a first stage, we have worked on 3-D virtualization of Kokakanjin Noh Emaki. [2] More specifically, a huge picture scroll was created on 3-D virtual space, and people slip back in time to the world of picture scroll by getting on the "tour ride" attraction.

Also we tried to show the contents in a more easy-to-understand way by displaying 3-D objects, sound, moving images, texts, picture images etc. at specific viewing points. (Figure 4)

With consideration for the fact that most of the residents of SL are English Speakers, this 3-D virtual museum can be appreciated with English explanation in addition to that in Japanese.

One of the important features of the "tour ride" from the viewpoint of SNS is that multiple people can enjoy viewing at the same time. One tour ride accommodates two people and one can talk to the other rider while viewing the picture scroll. Also after finishing the tour ride, one can exchange opinions with different visitors at the lounge space set up along the exhibition space (Figure 5).



Figure 4 Tour Ride display

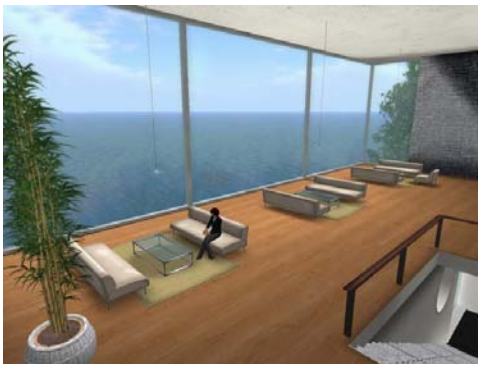


Figure 5 Lounge space set up in the Museum

As described above, construction of these facilities was carefully planned so that informal interactions can occur naturally. Those interactions have been recorded, analyzed, and utilized for the creation of other educational environment in 3-D virtual space.

4.3 Future projects

From the results of many demonstration experiments carried out previously, we have been convinced that 3-D virtual space can be well utilized in various educational activities. However, replacing all the distance learning or all the e-Learning courses offered at present with this kind (which uses 3-D virtual space) would be unrealistic. As we explained in the previous section, we consider it important to create an environment which seamlessly works together with existing systems or study-support tools on 3-D virtual space and to maintain an environment which guarantees interactivity. We could find a proceeding research on the integration of SL and Moodle to create a 3D virtual learning environment [(10) (11) (12)].

At present we have been working on the development of a tool which links attendance records or study histories of students on 3-D virtual space with existing CMS database (Figure 6).

[2] A picture scroll drawing Kanjinn Noh which was performed in Koka 5 (1848) by HOSHO, dayu around Edo Kanda Sujichigai bridge. The writer is presumed to be Gesshin Saito. Original work doesn't exist and only several kinds of copies exist.



Figure 6 A prototype of CMS linked attendance management tool

Hosei University plans to continue offering international collaborative distance courses with even greater enthusiasm in the future. We also expect that in the future students scattered all over the world will participate in classes which are offered on various platforms. To make it possible, we must further develop the tools and create an environment where integrated management of attendance and study situations of those students is possible. In this way, we would like to pursue the goal where can offer opportunities for study not influenced by where and how the lectures are being taken. In addition, by creating such environments where students participating from a remote site can interactively mingle with lecturers and other students, we would like to stimulate their desire for participation in classes as well as their motivation to study.

5. Conclusion

With technologies rapidly advancing, offering e-Learning education will become easier and easier in the future. But when developing tools or creating the collaborative environment, we should never neglect the importance of promoting communications between lecturers and students. We would like to continue our approach in harmony with total education to enhance the quality of educational services without losing sight of these viewpoints and without excessively depending on technologies.

References

(1) Hisato KOBAYASHI, Kazauo YANA, Kimiyoshi HAYASHI, Yoshiyuki HINO "Creating the Educational Environment Across the Space-Time", Journal of Multimedia Aided Education Research, Vol. 2, No. 2, pp.51-58, 2006

(2) Kazuo YANA, Hiroshi TAKEDA, Hiroshi FUKUDA, Takashi IMAHASHI, Toshikazu TAKAO, Reiko TOCHIGI, Kimiyoshi HAYASHI, Yoshiyuki HINO, Shuji KITAO, Billy PHAM, Ivan HO, Elizabeth GREENWOOD, Nicole RANGANATH, Jay MECHLING, Pablo ORTIZ and Dennis DUTSCHKE, "A Pacific Rim E-class Exchange Program," Proceedings of the 7th APRU Distance Learning and the Internet Conference 2006

(3) Ivan HO, Hajime KIYOHARA, Akira SUGIMOTO and Kazuo YANA, "Enhancing Global and Synchronous Distance Learning and Teaching by Using Instant Transcript and Translation," Proc. 2005IEEE International Conference on Cyberworlds, pp. 373-377, 2006

(4) Masami IWATSUKI, Norio TAKEUCHI, Hisato KOBAYASHI, Kazuo YANA, Hiroshi TAKEDA, Hisashi YAGINUMA, Hajime KIYOHARA, "Automatic Digital Content Generation System for Real-Time Distance Lectures", Methods and Applications for Advancing Distance Education Technologies, IGI Global, pp. 1-8, 2009

(5) http://sakaiproject.org/portal

(6) Kiyoshi ASAKAWA, Kazuo YANA, "Applying Flow Theory to the Evaluation of the Quality of Experience in a Summer School Program Involving E-interaction," Proceedings of the 15th International Conference on Distributed Multimedia Systems, pp. 279-284, 2009

(7) Hiroshi WATANABE, Fumihito FURUKAWA, Kumiko TAKAI, "Assistance for Learning Using Course Management System and Social Networking Service", Proceedings of the 33rd Annual Conference of Japanese Society for Information and Systems in Education, pp.2-3, 2008

(8) Billy PHAM, Ivan HH, Yoshiyuki HINO, Yasushi KODAMA, Hisato KOBAYASHI and Kazuo YANA, "A 3D Real-time Interactive Web-cast Environment for E-Collaboration in Academia and Education," Proceedings of the 15th International Conference on Distributed Multimedia Systems, pp. 275-278, 2009

(9) http://shanghailectures.org/

(10) Daniel Livingstone, Jeremy Kemp, Edmund Edgar, "From Multi-User Virtual Environment to 3D Virtual Learning Environment", Association for Learning Technology (ALT) Journal, Volume 16, Issue 3, pp. 139 – 150, 2008

(11) Andrea De Lucia, Rita Francese, Ignazio Passero, Genoveffa Tortora, "Development and evaluation of a virtual campus on Second Life: The case of SecondDMI", Computers & Education 52(1), pp. 220-233, 2009

(12) Andrea De Lucia, Rita Francese, Ignazio Passero, Genoveffa Tortor, "Development and evaluation of a system enhancing Second Life to support synchronous role-based collaborative learning", Software Practice and Experience 39(12), pp. 1025-1054, 2009

Proceedings
International Workshop on
Visual Languages Computing
(VLC 2010)

Editor

Giuseppe Santucci, *University of Rome, Italy*

Visualizing the Commonalities Between Hierarchically Structured Data Queries

Chad Jones
University of California, Davis
Email: cejjones@ucdavis.edu

Ryan Armstrong
University of California, Davis
Email: rnarmstrong@ucdavis.edu

Kwan-Liu Ma
University of California, Davis
Email: ma@cs.ucdavis.edu

***Abstract*—In this paper we present an information visualization approach for visualizing the intersection between multiple feature hierarchies. In cases where a dataset has multiple complex features of interest, especially ones that have been hierarchically clustered, it is often very difficult to grasp the commonalities between them. Understanding the overlap between features can help researchers discover correlations or hot spots in their data. Our method steps back from actual data features, simplifying their representation, and presenting them in a fashion which facilitates interpretation of their inter-feature relationships. We call this new representation a *knowledge graph*, and we show how it can be used to investigate the overlap of multiple features in a photo database and a social network dataset.**

I. Introduction

As data continues to increase in both size and complexity, extracting a single feature from an entire collection of information becomes a useful path on the road toward discovery. Smaller subsets of data are easier to understand and process, using methods such as numerical analysis, hierarchical clustering or direct visualization. Instances arise, however, where several different features are required to adequately study a problem. Besides their individual importance, the separate features of information could potentially contain correlating values that help support or deny a hypothesis. Exploring and understanding feature correlations is a task that can benefit greatly from interactive visualizations.

What a feature represents depends on the underlying data and how the feature is extracted. For social networks, a data query might return communities of friends with a common interest, but for an image database, a feature could be a collection of photos sharing a specific tag. Individually, each query may result in thousands of data items being returned, so when the filtered data becomes overwhelmingly complex, a common approach is to hierarchically cluster results into separate categories that show more detail as one expands each level. Once all the results are collected and processed, a researcher may then want to see how the various resulting feature hierarchies are related. In order to assist in the exploration process and provide a meaningful basis for feature comparison, we propose a representation that can be applied to all types of problem sets

Our approach creates a high-level abstraction of feature information in the form of a *knowledge graph* visualization. In a knowledge graph, hierarchically structured features and their intersections are shown in a multi-level compound hypergraph, allowing researchers to focus on higher level connections between features and providing interactive ways for hiding and revealing details. Unlike recent work in the area of a set visualization, we focus on visualizing sets that are clustered into one or more levels of abstraction using interactive techniques provided by node-link diagrams. The classification of features within a feature hierarchy provides new ways of understanding how different features correlate on multiple levels.

II. Background

The term knowledge graph is chosen for this feature-space visualization approach because the final graph represents the intersection of the user's results that were generated during the exploration process. The knowledge graph is a collection of what the user has discovered and how those features overlap. Our approach is built from areas of visualization meant to show relationships and correlations between result sets.

A. Visualization of Set Relations

The visualization of overlapping features can be traced all the way back to the 19th century when John Venn invented his set diagrams. Euler and Venn diagrams are the most common ways of visualizing the overlap of multiple sets, and they effectively show regions representing inclusion, exclusion, and containment relationships. The simplest representation utilizes overlapping circles. However, this approach is limited by the fact that it cannot accurately show intersections of four or more sets. To solve this problem, there has been considerable research to expand Venn diagrams to handle more possible set intersections [7], [17]. Recently, many algorithms have been developed to generate such diagrams automatically [16], [4].

Verroust and Viaud provided a new constructive method for creating extended Euler diagrams, which relax and extend the strict conditions on the traditional form of the diagrams [21]. With this new method, up to 8 sets can be visualized simultaneously. Using a planar intersection graph as the underlying layout, extended Euler diagrams do not require that regions be convex or completely filled, which allows for more control over the drawing procedure. A visualization system called VennMaster was developed by Kestler et al. to show multiple overlapping sets of results from genetic databases [12]. The authors provide a genetic algorithm to generate and position area-proportional sets and their overlap. The inconsistencies are listed in a separate table that is linked to the visualization.

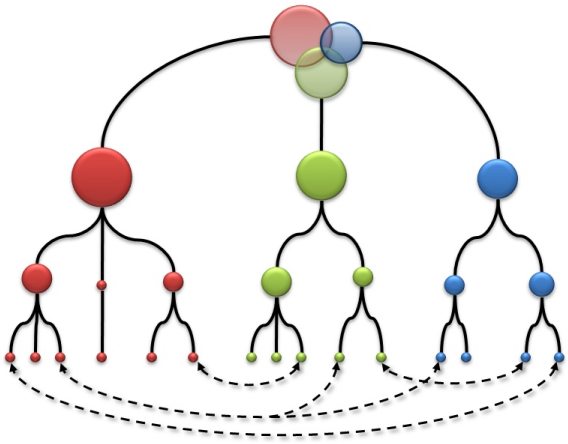


Fig. 1. Each set of data items is hierarchically structured. On every level of the hierarchy, there are adjacency (hyper)edges, shown here with dotted lines, connecting nodes containing items in common. For simplification purposes, only the lowest level adjacency edges are shown in this example.

Since not all set relationships must be visualized using standard Euler diagrams, other techniques have been introduced to extend or complement traditional methods. One such approach, developed by Simonetto and Auber, relies on the underlying intersection graphs for drawing Euler-like diagrams as flexible wrappers [18]. In an intersection graph, all nodes represent regions to be drawn and edges represent zones that are adjacent. The boundary of the sets are then drawn with the intersection graph as a guide, creating blob like structures that overlap where appropriate. The InfoCrystal layout, created by Anselm Spoerri, uses a specialized arrangement of shaped nodes to show every possible set relationship at once [19]. Every data item is assigned to the node that represents the region it belongs to, such as the intersection of sets A and B. Chiara and Fish developed a system called EulerView to replace certain tasks that usually rely on tree views to control and navigate hierarchies [2]. It utilizes a look and feel similar to tree views so users can build and manage their own set relationships. Other systems, such as Mirage [8], leave the data correlation task to the user by providing a suite of various information visualization tools.

Our graph representation shares commonalities with different portions of the previous work in the field. However, we focus on an edge-standard hypergraph representation, which is rarely used for directly visualizing set relations. Unlike a standard binary graph, hypergraphs allow edges that can connect any number of nodes together using a hyperedge. Johnson and Pollack were the first to utilize the commonality between Venn diagrams and hypergraphs, and they showed that determining the planarity of a hypergraph is NP-complete [11]. Harel used a combination of Venn diagrams and hypergraphs to create higraphs, which have been widely used for diagramming concurrent states [6].

Hypergraphs offer many advantages when dealing with the overlap of hierarchically structured sets due to their flexibility and interactivity. Ambiguities resulting from misleading overlap are not a problem in a hypergraph as intersections are discreet edges, and the overwhelming amount of information can be handled in an interactive focus+context manner.

B. Graph-Based Relational Visualization

Graphs are classified into a wide-array of types depending on the kind of relational information being conveyed [14]. The most relevant type of graph for our research is the compound graph, which combines an intersection graph for visualizing graph hierarchy and an undirected or directed graph for showing adjacency relations. Visualizations that show adjacency relations on top of hierarchies include TreeMaps, ArcTrees [13], polyarchies [15], compound digraphs [20], and balloon graphs. Holten and van Wijk improved on these visualizations by introducing hierarchical edge bundling [9]. These represent the common basis for creating visualizations that can represent intersections between two or more distinct hierarchical graphs.

To solve the problem of visualizing node correlation between independent graphs, VisLink expands on the idea of coordinated multiple views by connecting the views with adjacency edges, essentially creating heterogeneous compound graphs [1]. While the visualization only deals with binary connections, it does an efficient job of revealing connections between pairs of visualization results. Along a similar line, Holten and van Wijk created a visualization for showing the relations between a pair of distinct hierarchies [10]. Each hierarchy is laid out separately, and edges are used to connect hierarchies that are related, similar to the way VisLinks connects related nodes between views.

The flexibility and interactivity of these visualizations is what led us to explore the possibility of representing complex set correlations using hypergraphs. We utilize many of the principles found in the compound graph visualizations discussed here and make our own adjustments and additions to meet the needs of the problem.

III. Knowledge Graphs

Our system, KnowHow, utilizes a *knowledge graph* to represent the feature hierarchies generated from a collection of data. In the following sections we will describe the basic concepts that KnowHow is built from and then cover the visual representation and our implementation.

A. Feature Hierarchy

One important part of data analysis and exploration is separating meaningful subsets of data from the entire collection. This process can result in numerous features, depending on the type of data and the extraction method. We define a set to be a collection of data that results from some type of search query, such as an SQL select statement. Unlike previous set visualizations, we focus on sets that are hierarchically categorized.

We define a feature to be a collection of data items, and a feature hierarchy is defined as a tree where features are divided into smaller and smaller groups. The root level is a single feature that combines all of the data from all search queries,

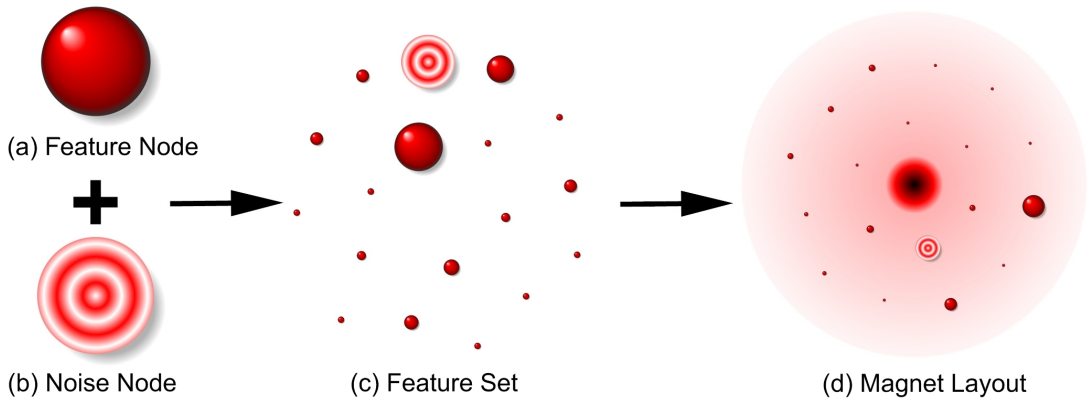


Fig. 2. The nodes of the knowledge graph represent features are colored according to the parent set. A regular feature node is shown in (a) and a specialized noise node is shown in (b). The noise node aggregates small features together. Together, nodes from the same level in the hierarchy form a feature set (c). The nodes of the feature set are positioned with a force directed layout and are influenced by its movable feature magnet, which is the center icon of (d).

the first level represents the original high level features, where each query has a single node. The leaves represent degenerate features, where only a single data item is represented by that node. The middle levels contain features generated by hierarchically categorizing the resulting data. We refer to each collection of features on a given hierarchy level as a feature set. Figure 1 shows three query sets that each have one hierarchical level of division.

The basis of our knowledge graph representation, however, is to show the overlap between features from multiple search queries. Since degenerate features, or the leaves, may be found in more than one search query, we connect duplicate leaves with an adjacency edge or *hyperedge*, which are shown as dotted lines at the bottom of Figure 1. These adjacency edges are propagated up the tree so that each feature in the tree is connected to other features with shared items. The structure is similar to the one visualized by Holten with hierarchical edge bundles [9], with the addition of hyperedges.

The grouping of data items into feature hierarchies depends on the what the user deems important, and each type of data can have different attributes used for dividing up the results. For a social network, a set would contain a collection of individuals who fall into a certain search query, such as users from the same university. Calculating community structure is a common method for dividing the results into a feature set. In such an example, an overlap of features exists if two community clusters contain the same person. Other types of simple categorization could hierarchically divide data by geographical location or by date and time.

B. Hypergraph Representation

Given that each feature hierarchy constitutes a collection of clustered results, the relationships between those results at a given tree level can be represented by a hypergraph. In a knowledge graph, each node represents a clustered feature and the edges convey the intersections that exists between the features.

1) Nodes: At the first level of a feature hierarchy, each result subset is represented by a single node. We assign an individual color to each of these nodes, such that all features originating from it will share that color. This high level node is split into multiple feature nodes as the user descends the tree, revealing new feature sets at each level. In addition to coloring, we collect a family of features from the same result subset into a cohesive magnet cloud. The size of a node is proportional to number of data items contained within that clustered feature.

We also allow for a special type of node, called a noise node, to help reduce clutter in the graph. A noise node, shown in Figure 2(b), is an aggregate of features that contain only a small number of data items. The noise node aggregation is performed on a per feature set basis and is colored by the parent feature set. The noise nodes are given a special type of rendering so the user does not mistake it for a normal feature.

2) Edges: The edges of the knowledge graph represent the intersection of data items between features. In terms of two different photo database searches, intersection means that the features contain at least one photo in common. For standard partial intersection, i.e. where $A \cap B \neq \emptyset$ and $A \cap B \neq A$ for $|A| \leq |B|$, we use a straight edge between nodes, as shown in in Figure 3(a). The edge size describes the number of items in the intersection and the color and brightness describe what percentage of each node is contained in that intersection.

Since part of the knowledge graph's power is the ability to correlate multiple result subsets simultaneously, we also generate hyperedges, such as the one shown in Figure 3(c). A hyperedge shows the multi-intersection among a group of nodes in the same fashion as a binary edge. Whenever a group of three or more feature nodes share data items then there exists a subset of data contained in all those features. If this exists, then a hyperedge can be drawn as a star structure by having an outgoing edge from every node meet at a centralized crux, which we define as the centroid of the intersecting nodes. The same color scheme is used here as with binary edges, and thus, the hyperedge shows the overlap of many sets at a time.

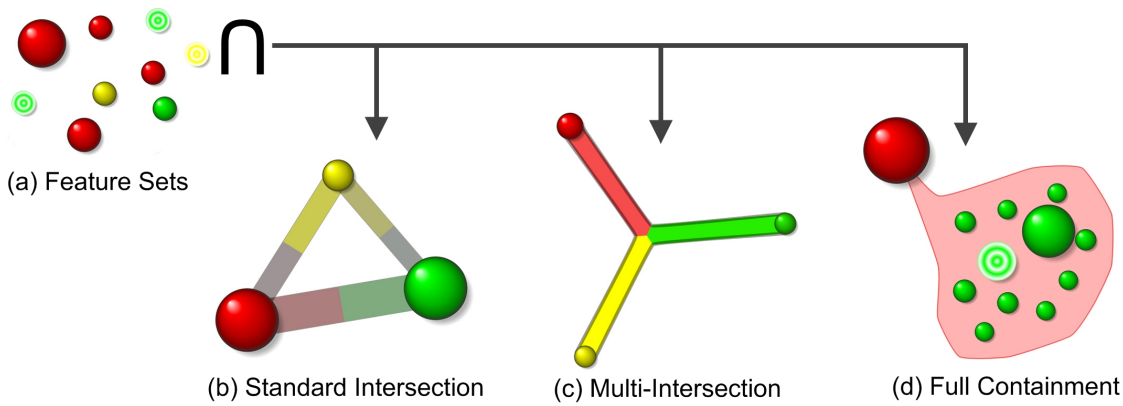


Fig. 3. When intersecting multiple features (a), partial intersections are shown as straight edges, (b) where the color shows the percentage of containment for each node. The brighter the color, the more of that node is contained in that intersection. When multiple features simultaneously share data, a multi-edge is used as in (c). When features are subsets of another feature, they are aggregated into a one-way bubble edge (d).

In the case of total containment, i.e. where $A \cap B = A$ for $|A| \leq |B|$, we use a specialized type of edge, called an edge bubble, that groups all such nodes into a single collection of contained nodes. This helps to reduce visual edge clutter in the knowledge graph by aggregating nodes with a commonality and eliminating unnecessary edge rendering. For every node A_i that is completely contained in a node B, the nodes A_i are clustered together and surrounded by a polygonal boundary, as shown in in Figure 3(d). This polygon boundary then connects to the node B just like an edge and is filled with the color of node B's set.

3) Hypergraph Layout: In order to better utilize screen space, the graph is positioned using a force directed layout [5]. During the layout phase of the visualization, however, the algorithm takes certain hidden forces into consideration. To properly handle the contents of an edge bubble, the nodes inside the bubble are conceptually tied together and attached to the larger feature node. This keeps the bubble tightly bundled, which reduces clutter and improves visual appeal. Hyperedges do not influence the layout of nodes in the graph; only the the binary edges, or simple intersections, are used. Since a complete binary subgraph exists for every hyperedge's incident nodes, it is unnecessary to use the hyperedge in the force calculations, regardless of its visibility.

A significant modification to the force directed layout is the addition of interactive magnets, which is similar to a technique that has been applied to multivariate data [22]. The purpose of magnets is to organize a scattered collection of nodes and edges into more meaningful sectors. A magnet exists for each result subset, i.e. for every search query, and can be repositioned by the user. The magnet stays anchored to its set location and attracts all of its corresponding feature nodes toward it, as is shown in Figure 2(d). Since all magnets pull on their respective nodes, the end result is a separation of connection types, e.g. unconnected nodes float around the magnet while nodes with a standard edge to another node will be positioned in the middle of the two feature magnets.

C. Interactive Exploration and Modification

While the force directed, magnet layout reduces overlap, other tools are required for exploring the graph in detail and reducing clutter. One of the simplest ways of modifying the knowledge graph is by moving through the hierarchy levels, revealing more detailed feature sets at lower and lower levels. In order to help the user keep track of information between transitions, such as adding a new result set or changing the hierarchy level, we provide animations that lead the user from one stage to the next. When adding a new set, the nodes are faded into the existing knowledge graph gradually with the layout only being modified slightly from the previous one. These types of animated transitions are used throughout to provide visual feedback of changes being made, similar to the work done by [3].

Besides being able to move from one hierarchy level to another to see each feature set, it is occasionally advantageous for a user to see a single node's children without completely changing the view. In this case, a user may select a node from the current feature set and expand its contents much like a balloon graph. The children nodes become surrounded by an expanded circle to show that they are on a lower level than the rest of the feature set. The edges of the children become visible as they connect to higher level nodes outside of the balloon.

Additional interface options are provided for filtering nodes and edges. Cardinality sliders can hide nodes or edges that do not fall within the specified range. The layout updates interactively as items are hidden. We also allow individual edge types to be selectively made invisible without disrupting the layout positioning. For example, the user may select only tri-edges, which represent the intersection of three sets, or quad-edges to be rendered while hiding all others. This allows some fine grain control to generate the exact image the user would like to see.

A more powerful interactive tool is the edge toggler. When viewing the knowledge graph, it may not be useful to see intersections between certain features sets and not others.

Thus, we provide a simple user interface that allows the user to turn on or off edges between any pair of feature sets. The unwanted edges become invisible, and the force directed layout will adjust node positions accordingly, as if those nodes no longer attract. This would be analogous to a Euler diagram choosing not to overlap circles between certain sets.

The system also allows a user to see the original data represented by each feature node through mouse selection. When a node or edge is picked, the original data is displayed in a side window. The default view is a table of entries and specialized views, such as a photo viewer for flickr data, can be written for specific types of data.

IV. Results

Now we will show the overlap of search results from Flickr data and social network data. While the method of obtaining and clustering the results is not a topic of this paper, we created a simple test application that uses a SQL database for queries and simple categorization that is coded specifically for each different dataset. It is possible to find all possible types of intersections manually using SQL queries, but knowledge graphs present a complete view of all intersections simultaneously and provide a suite of interactive methods that let the user explore the hierarchy in context while creating a custom view of the data.

A. Flickr

Our Flickr database contains over 13 million photo details pulled from Flickr prior to 2009. Each photo entry contains 31 values, including user, timestamp, view count, location, title, and technical specifications, plus any number of possible tags. Since the last images we collected were at the end of the 2008 U.S. presidential elections, we performed our case study to examine the popularity of the candidates over time based on their appearance in photo tags.

In Figure 4 we begin with five tag searches, which include the president and vice-president candidates and the tag 'election'. The first level is, of course, one node per result. We then create the feature hierarchy by dividing nodes by date of photo, with the second level being year, the third level being month, etc. Level 1 of Figure 4 shows how some of the interactive tools are used to discover high-level aspects of the data. First, the nodes appear fully connected, and with edge hiding, we show a hyperedge representing photos containing all five tags. Second, by ballooning out the nodes and using the edge toggler, we can see just how the tag 'election' relates to every other tag based on year. In this example, 2008 is the largest year for all of the feature sets, and it is the only year that the the vice-president candidates appear with the tag. In addition, there are some photos from the years 2006 and 2007 that reference either John McCain or Barack Obama.

We transition to the next hierarchy level, which reveals features sets separated by year. We remove the 'election' tag to focus on just the candidates. In 2008, Barack Obama is the most popular tag of the group, which can be seen by comparing feature node sizes, so we decide to expand that

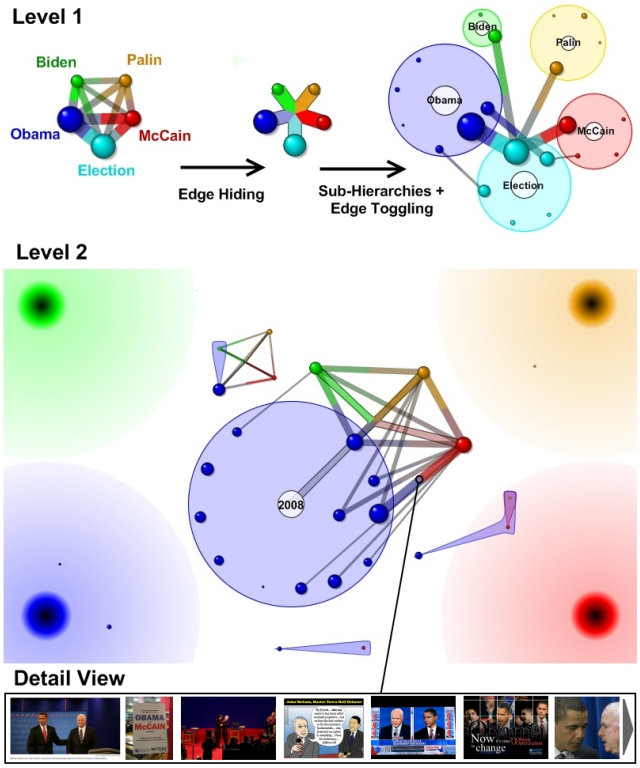


Fig. 4. The above example shows feature hierarchies from a Flickr database. Each hierarchy represents a tag pertaining to the 2008 US presidential elections clustered by time. Using several interaction techniques, the first level reveals 2008 was the most popular for photographing all candidates, and the second level shows how Barack Obama became more prominent as the November election approached. A detailed view of one of the edges between Obama and McCain is shown at the bottom.

node to reveal it's sub-hierarchy. If we mouse over the nodes to reveal the cluster labels, we see that the most popular month is November, the same month as the election. Also, Obama and McCain began appearing together before either of the vice-president candidates came into the picture. A detail view of images is seen below the graph by selecting the edge connecting Obama and McCain during the month of October. Overall, this exploration of the knowledge graph reveals the rise of important figures in the media during the long election process. Without knowledge graphs, the same conclusions might require multiple searches or visualizations of the data in order the represent the same hierarchy and overlaps.

B. Orkut Network

Our social network dataset, which comes from the Orkut network (a Google product), contains over 15,000 anonymized user profiles and over 60,000 friend connections. The feature hierarchies in this case are generated using community structure clustering methods commonly found in social network studies. Therefore, each node represents a cluster of closely connected friends. In the example shown in Figure 5, instead of taking queries intended to generate highly correlated groups of people, the queries were chosen randomly, without considering the others. The feature hierarchies are as follows: people

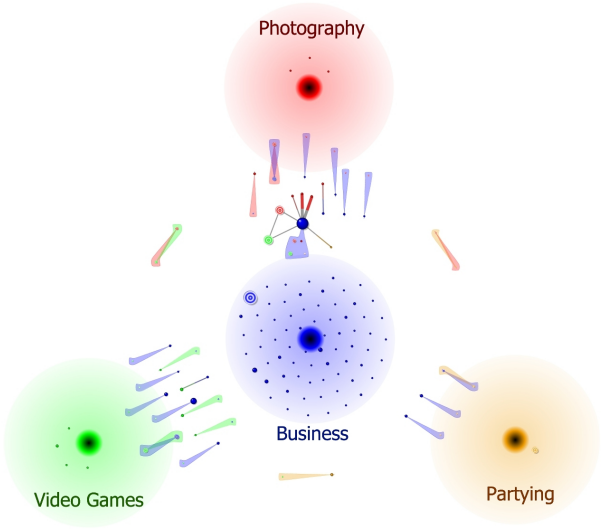


Fig. 5. In this figure, four unrelated search queries are combined into a knowledge graph. The sets represent the interests of Orkut users, divided into photography, video games, business, and partying. Even though no overlap was intended, some users were found having eclectic listings of activities.

interested in cameras or photography (red), people interested in video or computer games (green), people over the age of 20 on Orkut for business (blue), and people interested in partying or hanging out (orange).

The knowledge graph here shows where these very different feature hierarchies have commonalities. Between each feature set, there exists a small amount of intersection that indicates a combination of those interests. In the center, a large community of business minded people can be seen sharing interests from each of the other feature sets. Since these are not full containment or subgraph cliques, these are minorities from the larger feature node. A social network researcher may be able to discover individuals who appear in many popular, unrelated searches, which could then lead to an understanding of some small outlying connection among them all.

V. Conclusion

This paper has introduced knowledge graphs, a visualization of overlapping feature sets. Using a modified node-link diagram representation where nodes are features and hyperedges are intersections, our system, KnowHow, provides a high level view of feature hierarchies from a wide range of different datasets. Unlike other set relation diagrams, knowledge graphs focus on sets that have been hierarchically clustered. KnowHow utilizes various aggregation, abstraction, and interaction techniques to help reduce the complexity of the feature space and allow for feature exploration. We have applied this system to a photo database and a social network. Without modification, KnowHow is able to abstract different types of data into the same feature space, which shows that the knowledge graph representation is versatile enough to be applied to many areas.

It is clear from our evaluation of the system that being able to comprehend and navigate the knowledge graph requires the user to become familiar with the new abstract view of their features. Linking the feature nodes directly to the original data helps in this regard, though, as it maps the abstract features back to the actual features. As our work on feature space visualization continues, we are prepared to perform user studies to test the ability of users in using knowledge graphs to answer questions about features and its effectiveness as an interface for exploring the original data. In addition, we plan to study enhanced visual representations that may be more intuitive at conveying the same information or are capable of simplifying the information.

References

[1] S. Carpendale and C. Collins. Vislink: Revealing relationships amongst visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1192–1199, 2007.

[2] R. D. Chiara and A. Fish. Eulerview: a non-hierarchical visualization component. *Symposium on Visual Languages and Human Centric Computing*, pages 145–152, 2007.

[3] P. Eades and M. L. Huang. Navigating clustered graphs using force-directed methods. *Journal of Graph Algorithms and Applications*, 4(3):157–181, 2000.

[4] J. Flower, A. Fish, and J. Howse. Euler diagram generation. *Journal of Visual Languages and Computing*, 19(6):675–694, 2008.

[5] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA*, 99(12):7821–7826, June 2002.

[6] D. Harel. On visual formalisms. *Commun. ACM*, 31(5):514–530, 1988.

[7] D. W. Henderson. Venn diagrams for more than four classes. *American Mathematical Monthly*, 70:424–426, 1963.

[8] T. K. Ho. Interactive tools for pattern discovery. In *International Conference on Pattern Recognition*, pages 509–512, 2004.

[9] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, September 2006.

[10] D. Holten and J. J. van Wijk. Visual comparison of hierarchically organized data. *Computer Graphics Forum*, 27(3):1192–1199, 2008.

[11] D. S. Johnson and H. O. Pollak. Hypergraph planarity and the complexity of drawing venn diagrams. *Journal of Graph Theory*, 11(3):309–325, 1987.

[12] H. Kestler, A. Muller, J. Kraus, M. Buchholz, T. Gress, H. Liu, D. Kane, B. Zeeberg, and J. Weinstein. Vennmaster: Area-proportional euler diagrams for functional go analysis of microarrays. *BMC Bioinformatics*, 9(1), 2008.

[13] P. Neumann, S. Schlechtweg, and M. S. T. Carpendale. Arctrees: Visualizing relations in hierarchical data. In *EuroVis*, pages 53–60, 2005.

[14] H. Omote and K. Sugiyama. Method for visualizing complicated structures based on unified simplification strategy. *IEICE - Transactions on Information and Systems*, E90-D(10):1649–1656, 2007.

[15] G. Robertson, K. Cameron, M. Czerwinski, and D. Robbins. Polyarchy visualization: visualizing multiple intersecting hierarchies. In *Conference on Human Factors in Computing Systems*, pages 423–430, 2002.

[16] P. Rodgers, P. Mutton, and J. Flower. Dynamic euler diagram drawing. In *Symposium on Visual Languages and Human Centric Computing*, pages 147–156, 2004.

[17] F. Ruskey, C. D. Savage, and S. Wagon. The search for simple symmetric venn diagrams. *Notices of the American Mathematical Society*, 53(11):1304–1311, 2006.

[18] P. Simonetto and D. Auber. Visualise undrawable euler diagrams. In *Intl. Conference on Information Visualisation*, pages 594–599, 2008.

[19] A. Spoerri. Infocrystal: a visual tool for information retrieval. In *Visualization Conference*, pages 150–157, 1993.

[20] K. Sugiyama and K. Misue. Visualization of structural information: Automatic drawing of compound digraphs. In *IEEE Transactions on Systems, Man and Cybernetics*, volume 21, pages 876–892, 1991.

[21] A. Verroust and M.-L. Viaud. Ensuring the drawability of extended euler diagrams for up to 8 sets. In *Diagrams*, pages 128–141, 2004.

[22] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

Towards Security-aware Program Visualization for Analyzing In-lined Reference Monitors

Aditi Patwardhan, Kevin W. Hamlen, and Kendra Cooper
Department of Computer Science
The University of Texas at Dallas
Email: aditi.patwardhan@student.utdallas.edu, hamlen@utdallas.edu, kcooper@utdallas.edu

***Abstract*—In-lined Reference Monitoring frameworks are an emerging technology for enforcing security policies over untrusted, mobile, binary code. However, formulating correct policy specifications for such frameworks to enforce remains a daunting undertaking with few supporting tools. A visualization approach is proposed to aid in this task; preliminary results are presented in this short paper. In contrast to existing approaches, which typically involve tedious and error-prone manual inspection of complex binary code, the proposed framework provides automatically generated, security-aware visual models that follow the UML specification. This facilitates formulation and testing of prototype security policy specifications in a faster and more reliable manner than is possible with existing manual approaches.**

I. Introduction

Software security is becoming increasingly important with the growth of the Internet and mobile code technologies like Java. Mobile code technologies generate software components for environments in which *code-consumers* receive code from separate *code-producers*. These components are mainly distributed as binary executable files that are downloaded from web pages or as email attachments. In many realistic settings, not all code-producers are fully trusted; for example, web pages may be served from untrusted servers or emails may arrive from untrusted senders. Security is an obvious concern in such an environment. Violations can range from information leakage to access control violations and data corruption.

Secure mobile code environments constrain the behavior of untrusted code by enforcing liveness or safety properties. Many practical policies can be formulated as safety policies. A classic example is the confidentiality policy that prohibits network-send operations after the process has read from a confidential file. This prevents the untrusted process from divulging the file's content over the network. These *history-based* (i.e., stateful) safety policies can be encoded formally as *security automata* [1]. A security automaton is a finite state automaton that accepts policy-permitted sequences of security-relevant events. Bisimulation of the security automaton with the untrusted program is used as a mechanism to enforce the underlying security policies. When an impending violation is detected by the automaton, the offending process is terminated.

In-lined reference monitors (IRM's) [1] have been proposed as a mechanism to insert security automata into untrusted binary code by modifying it (e.g., [2]–[4]) prior to execution. This results in *self-monitoring code* that is guaranteed to self-terminate before policy violations occur. IRM's provide a powerful means to enforce application-specific security policies, but identifying and defining a good security policy, particularly at the bytecode level, is challenging. For example, to prohibit network-send operations, one must be able to rigorously define what constitutes a network-send operation. This may involve considering hundreds of primitive instructions that constitute potentially security-relevant operations, each of which must be identified in a complete, correct specification of the policy. In general, much manual analysis and inspection of malicious and non-malicious binary code may be required in order to formulate a policy that prohibits all undesired program behaviors without curtailing desired behaviors.

A visualization approach, with tool support, is needed to aid the analysis of the untrusted binary code and to facilitate faster and more reliable discovery and prototyping of application-specific security policies. In this short paper, we present a visualization framework that generates security-aware UML-based visual models for the low-level Java bytecode; preliminary results are reported. The framework supports static and dynamic models. The static visual model represents the underlying class structures; the dynamic visual model represents the possible execution sequences (control-flows) in the application. Each execution sequence is mapped to the set of corresponding state transitions of the security automaton (given by the security policy); a user-defined *color code* is provided to visually map the control-flow blocks to the corresponding security automaton states and indicate possible security violations. The approach and tool support have been validated using a classic confidentiality policy and two test applications that violate this policy by leaking information over the network.

The remainder of this paper is organized as follows. Section II discusses the existing tool support for analysis of bytecode and other related work. Section III describes our visualization framework proposed for the security analysis, and the prototype proof of concept that we have developed. Section IV summarizes and proposes future work.

II. Related Work

Practical IRM systems constitute a growing body of past work (c.f., [5]). The Java-MOP system [3] combines runtime monitoring with aspect-oriented programming. It specifies the desired security properties, along with the code to execute if a security violation is detected. The specification is then

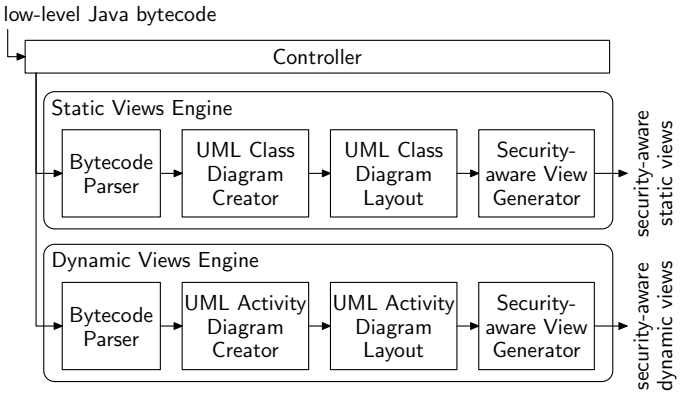


Fig. 1. A security-aware visualization framework

translated to AspectJ code and integrated into the application program using an aspect weaver. The SPoX (Security Policy XML) system [4] provides a purely declarative policy specification language in which security-relevant events are designated via AspectJ pointcuts, and policies over these events are specified as security state transitions. We developed our visualization tool based in part on bytecode analysis libraries provided by the SPoX toolkit.

Traditional text-based, code-level visualization is supported by many established tools including decompilers (e.g., [6], [7]), debuggers, and various low-level libraries for static code analysis. Eclipse [8] provides an integrated debugger for Java source code that allows execution of the source code interactively by stepping through each line of code. Some open source byte code debuggers are also available for finding the trace of execution at a binary code level. Reverse engineering tools (e.g., [9], [10]) and context-independent analyses (e.g., [11]) augment such analyses with support for abstraction visualization and binary format discovery.

However, all of these tools are general purpose; they do not provide specialized support for security. For example, policy-violating control-flows are not automatically identified for the user; they must be manually discovered and extracted. The modeling tools (e.g., [9]) use UML as the modeling notation, which is a powerful and established standard for the graphical modeling of object oriented software design and analysis. We take the approach of catering to security-specific requirements while utilizing the powerful features of a visual modeling notation like UML and using the concept of abstraction for a better representation of the underlying bytecode.

The BCEL API (Byte Code Engineering Library) [12] provides a programmatic foundation for analyzing Java bytecode. Our visualization framework uses this API to extract low-level Java bytecode information.

Graphical models of low-level code provide easier, faster, and more reliable analysis of an application's structure and its possible control-flows than their text-only counterparts. We therefore adopt a graphical approach. Desirable quality attributes and functional requirements for general-purpose code visualization tools have been well-studied [13], but there has been no similar study of security-aware tool functionality to our knowledge.

III. Security-aware Visualization Framework

Figure 1 depicts the structure of our visualization framework, which transforms low-level bytecode into UML-based visual models [14], [15]. It is composed of a controller and separate view engines for the generation of static and dynamic visual models. The *static views engine* generates the view for the static structure of the application modeled as a UML class diagram. The *dynamic views engine* generates the dynamic views for the detailed control-flow diagrams for each class method within the application code, modeled as a UML activity diagram. These UML specifications are used to provide the security-aware views of the application. We choose the Unified Modeling Language (UML) specification for our visual model, as it has become the de facto standard for visual modeling of software applications. We represent the class structures and their relationships using *UML class diagrams*, and the control-flows using *UML activity diagrams*.

The UML class diagrams can be used to compare the untrusted bytecode with the self-monitoring, rewritten code obtained from an IRM framework. The UML activity diagrams can be used to map the prototype security policy to the underlying control structure and to visually identify the security events in the control-flows.

Our visualization framework is built atop the Eclipse plug-in architecture [8]. It can be launched within the Eclipse environment or used as a standalone rich client desktop application. The plug-in architecture allows for easy extensibility to add further custom diagrams related to the security specifications. Further, it does not require any special setup; the only main requirement is a standard Java runtime environment. The prototype tool developed is available at the URL http://utdallas.edu/~aap085000/VisualizationTool.

A. Security-Aware Static View

To effectively prototype and analyze real IRM's and the policies they enforce, it is important to be able to easily visualize and compare the class structure of original and IRM-modified Java bytecode applications. For example, most practical policies constrain usage of certain security-relevant system classes by untrusted applications. The IRM must therefore track the security state of these security-relevant objects at runtime to enforce the policy. The IRM typically accomplishes this by injecting new wrapper classes that inherit from and extend the system classes with extra security state fields maintained by the IRM [3], [4]. Thus, visualizing the class structure of original and IRM-modified applications reveals much about the potential effects of the policy-enforcement upon the untrusted application, including undesired side-effects and potential runtime overhead.

The UML class diagrams [14] represent the static structure of the original and/or IRM-modified system as a structure of classes and relationships between them. They are generated via four pipelined components: a *bytecode parser*, *UML class diagram creator*, *UML class diagram layout*, and *security-aware view generator*.

The bytecode parser identifies the classes, their data attributes and methods, the visibility options of the data attributes and methods, and the relationships between the classes. The UML relationships supported include generalization (based on class inheritance) and association (based on class object reference).

The diagram creator generates the entities of the UML class diagram that represent the information extracted by the parser. Classes are mapped to UML class elements, and class relationships are mapped as generalization or association relationships of the UML class diagram. Supported UML class diagram elements include classes, data attributes, operations, visibility attributes, generalization relationships, and association relationships. We construct the class diagram with inheritance relationships up to one level into the system libraries. Since the class hierarchy is a tree rooted at `java.lang.Object`, this yields a strong inheritance-based structure.

These UML diagram entities are then input as graph elements to the layout algorithm to generate the visual diagram. We use an inheritance-based structure similar to the one in [16] for the automatic layout generation of the class diagram with minimal edge cross-overs. The visualizer allows users to manually adjust the generated layout via a select, drag, and drop functionality for the class elements of the diagram.

The framework can additionally render a visual comparison between the original and the rewritten, IRM-modified application bytecode. This is extremely useful for analyzing changes in the static structure that result from enforcement of a given policy by an IRM. Separate UML class diagrams are generated for the original and IRM-modified application bytecode. Visual color-coded cues are provided to the changes made to the original bytecode during the rewriting process for enforcing the policy. Currently our prototype supports this at the class level granularity; any new classes added to the bytecode by the IRM are highlighted in the class diagram.

The framework's static model visualization has been validated using a test application that divulges a confidential file by sending its content over the network. We used the SPoX IRM system [4] to enforce a policy that prohibits write-access to the Java `Socket` library once a confidential file has been accessed. The class structures of the original, unsafe application bytecode and the SPoX-modified bytecode were then compared using the visualization framework. The new security class injected by the IRM was identified and highlighted in the visual model. Due to space constraints, the screen shots for the before and after class diagrams are not included here; they are available in [17].

B. Security-Aware Dynamic View

UML activity diagrams are used to illustrate the control-flows in a system [14], [15]. Activities typically consist of a network of nodes and edges that represent the flow within the activity; we therefore use them to represent the control-flows. The dynamic views engine consists of four pipelined components: a *bytecode parser*, a *UML activity diagram creator*, a *UML activity diagram layout* and a *security-aware view generator*.

The bytecode parser extracts instruction sequences that are partitioned into *basic blocks*—subsequences of consecutive instructions for which all control-flows (other than exception flows) enter at the beginning and leave at the end without intermediate branching. (Exceptions may result in premature exit from a basic block, and therefore receive special treatment described below.)

The UML activity diagram creator maps these basic blocks as the *call action nodes* of the activity diagrams. Call action nodes define the units of work that are atomic within the activity [15], and are therefore well-suited to basic blocks. Each basic block that ends in a conditional branch introduces a decision node to the diagram. Basic blocks that could throw exceptions caught by a local handler are visualized via control-flow edges from the basic block to the entry-point of the local exception handler. For visual clarity, the control-flow edges to exception handler blocks are shown as dashed arrows to differentiate from the normal control-flow edges.

The UML activity diagram constructs include call action nodes, control-flow edges, and *control nodes*, including an *initial node*, *final nodes*, and *decision nodes*. We use a dataflow analysis technique to identify all possible control-flows in the control structure of the activity diagram by traversing them statically. The UML diagram layout generates a flowchart-like layout that contains call action nodes ordered by underlying bytecode offsets, and the control-flows between them.

The security-aware view generator further provides security-aware views of the activity diagram generated. It takes an input security policy from the user and maps the policy to the control-flows depicted by the activity diagram. We detect all possible policy violations in the control-flows by computing a function $f : Q \to 2^S$ that maps each node $q \in Q$ in the control-flow graph to a conservative approximation of the set of security automaton states $s \in S$ that the node could assume at runtime. The computation involves obtaining the least fixed point of the functional $F : (Q \times S) \to (Q \times S)$ defined by

$$F(g) = g \cup \{(q_0, s_0)\} \cup \{(q', \delta(q, s)) \mid (q, q') \in E, s \in g(q)\}$$

where q_0 and s_0 are the start states of the control-flow graph and security automaton (respectively), $E \subseteq Q \times Q$ is the transition relation for the control flow graph, and $\delta : (Q \times S) \rightharpoonup S$ is the transition function for the security automaton, which defines how each basic block modifies the security state when executed. Our current implementation is intra-procedural; an inter-procedural extension is left for future work.

The visualizer then identifies sites of potential policy violations by identifying the control-flow graph nodes $q \in Q$ for which there exists a security state $s \in (\mathit{fix}(F))(q)$ such that $(q, s) \notin \delta^{\leftarrow}$. These are the states for which the security automaton has no transition, and that therefore might exhibit a policy violation at runtime. These nodes q are therefore the sites where an IRM will typically in-line runtime security checks to detect and prevent potential violations. The visualizer renders these nodes in a unique, user-specified color to

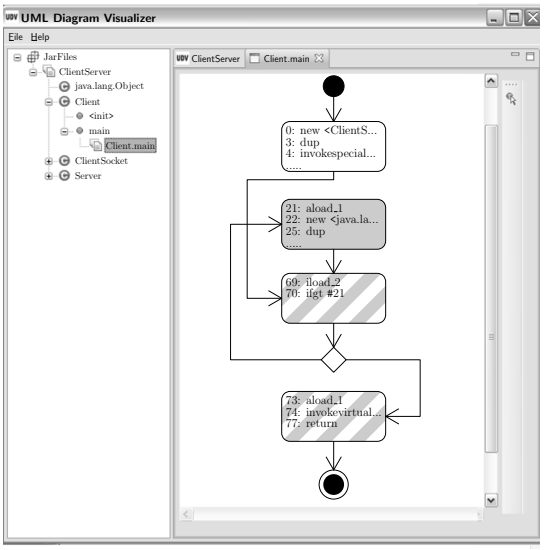


Fig. 2. A UML activity diagram with control-flow nodes color-coded by possible security state

bring them to the attention of the user.

We take a conservative approach in which the detection of policy violations may include some false positives. For example, a potential violation might be identified within an execution branch that is never actually traversed at runtime due to the value of some trusted input variable. However, false negatives will not occur; all potential violations are conservatively identified.

The framework's dynamic model visualization has been validated using a test client-server application (Figure 2). The screen shot illustrates the main method of the client, which contains a loop with a send and a read operation. The security policy prohibits sends after reads. The visualizer detects that this could result in a policy violation for control-flows that exhibit two or more iterations of the loop body. The generated high-level view of the control structure maps the possible security states of the automaton that each basic block could enter, thereby identifying those where a security violation could result. In this case the white node indicates that the security automaton is in an initial state, the striped nodes indicate that the security automaton may be in various different policy-adherent states on various different runs, and the filled node indicates a possible policy violation.

IV. Conclusions and Future Work

We have introduced a security-aware, bytecode visualization framework that facilitates fast and easy prototyping and analysis of IRM security policies and their implementations. This is the first approach to address this concern. Experiments show that the framework, implemented as a prototype tool, can generate visual diagrams and identify code points where possible security violations could occur at runtime. Without an automated tool, such analyses are extremely difficult and time-consuming even for experts, since they involve manually understanding an application's binary structure, its possible control-flows, and its potentially security-relevant operations. Rapid development and prototyping of candidate security policies is therefore typically impractical without such a tool.

A number of limitations have been identified in the visualization framework. With respect to the UML diagrams created, for example, the framework does not support reverse engineering to the complete, standard definitions of UML class and activity diagrams. The current subset has been adequate for the example applications used in the validation, but may need to be extended in the future. In addition, the security-aware color coding scheme used has not been rigorously defined as a UML extension, such as a profile or a stereotype. The graph based algorithms to create the diagrams also need to be systematically defined and analyzed.

The approach taken to identify the possible security violations may include false positives, wherein some unreachable control-flows are identified as possible sources of policy violations. However, the approach does not suffer from false negatives; it conservatively identifies all violations.

The framework can be easily extended to provide further support for visualization with respect to security, including extended debugging functionality and support for visual modeling. These are avenues we intend to explore in future work.

References

[1] F. B. Schneider, "Enforceable security policies," *ACM Transactions on Information and System Security*, vol. 3, no. 1, pp. 30–50, 2000.
[2] Ú. Erlingsson and F. B. Schneider, "SASI enforcement of security policies: A retrospective," in *Proc. New Security Paradigms Workshop*, September 1999, pp. 87–95.
[3] F. Chen and G. Roşu, "Java-MOP: A monitoring oriented programming environment for Java," *Lecture Notes in Computer Science*, vol. 3440, pp. 546–550, 2005.
[4] K. W. Hamlen and M. Jones, "Aspect-oriented in-lined reference monitors," in *Proc. ACM Workshop on Programming Languages and Analysis for Security*, 2008.
[5] J. Ligatti, L. Bauer, and D. Walker, "Run-time enforcement of non-safety policies," *ACM Transactions on Information and System Security*, vol. 12, no. 3, January 2009.
[6] P. Kouznetsov, "JAD: The fast Java decompiler," http://www.kpdus.com/jad.html.
[7] Ahpah Software, Inc., "The SourceAgain decompiler," http://www.ahpah.com/products.html.
[8] "The Eclipse platform," http://www.eclipse.org.
[9] International Business Machines, "IBM software architect," http://www-01.ibm.com/software/awdtools/swarchitect.
[10] H. M. Kienle and H. A. Müller, "Rigi: An environment for software reverse engineering, exploration, visualization, and redocumentation," *Science of Computer Programming*, vol. 75, no. 4, pp. 247–263, 2010.
[11] G. Conti, E. Dean, M. Sinda, and B. Sangster, "Visual reverse engineering of binary and data files," *Lecture Notes in Computer Science*, vol. 5210, pp. 1–17, 2008.
[12] "Byte code engineering library," http://jakarta.apache.org/bcel/.
[13] H. M. Kienle and H. A. Müller, "Requirements of software visualization tools: A literature survey," in *Proc. 4th IEEE International Workshop on Visualizing Software for Understanding and Analysis*, June 2007.
[14] B. Gruegge and A. H. Dutoit, *Object-Oriented Software Engineering: Using UML, Patterns and Java*, 2nd ed. Prentice Hall, 2004.
[15] J. Arlow and I. Neustadt, *UML 2 and the Unified Process: Practical Object-oriented Analysis and Design*, 2nd ed. Addison-Wesley, 2005.
[16] J. Seemann, "Extending the Sugiyama algorithm for drawing UML class diagrams: Towards automatic layout of object-oriented software diagrams," *Lecture Notes in Computer Science*, vol. 1353, pp. 415–424, 1997.
[17] A. A. Patwardhan, "Security-aware program visualization for analyzing in-lined reference monitors," Master's thesis, University of Texas at Dallas, June 2010.

Immersive Gesture-based Interface for Effective 3D Geographic Data Manipulation

Andrea F. Abate, Vincenzo Del Fatto, Michele Nappi, Stefano Ricciardi, Genoveffa Tortora

Università degli Studi di Salerno, Dipartimento di Matematica e Informatica
84084 Fisciano (Salerno), Italy
{abate, vdelfatt, mnappi, sricciardi, tortora}@unisa.it

ABSTRACT - *This paper presents a framework providing a collection of techniques to enhance reliability, accuracy and overall effectiveness of gesture-based interaction applied to the Geographic Information Systems context. We propose, indeed, an interaction framework featuring a gestural interface, operated by two-hand gestures, for an enhanced manipulation of 3D geographic data immersively visualized by means of a Head Mounted Display (HMD). The interaction paradigm, exploits one-hand, two-hand and time-dependent gesture patterns to allow the user to perform inherently 3D tasks, like arbitrary object selection and manipulation, or measurements of relevant features, in a more intuitive yet accurate way. As the HMD orientation is tracked, the user is able to see a much greater amount of geographic information by simply turning its head. The preliminary experiments conducted so far shows a measurable and user-wise perceptible improvement in performing 3D interactive tasks, like the selection of particular geographic locations even on a complex 3D surface or the distance measurement between two geographic features.*

1. INTRODUCTION

Gesture based interfaces have been often regarded by human-computer-interaction literature as a very natural way to navigate through and interact with virtual reality environments [1]. Thanks to more recent progress and to the overall cost reduction of related hardware, a broader and growing spectrum of applicative fields, like geovisualization may benefit of this technologies [2, 3]. Today, sophisticated Graphical User Interfaces (GUIs) have replaced basic commands of first GIS applications, while the rapid growth of computer graphics tools enables the visualization of complex 3D datasets [4], as witnessed by many research works combining GIS and VR particularly in the field of urban/environmental planning [5]. The convergence of various fields, such as computer science, HCI and geovisualization may indeed result in an important progress in GIS interfaces. In this study, we propose a framework aimed to enhance reliability, accuracy and overall effectiveness of gesture-based interaction applied to the interactive manipulation of 3D geographic data within a GIS context. The main goal pursued is to effectively combine a small set of simple (user-wise) one hand and two-hand gestures with a context adaptive virtual interface to intuitively and accurately perform tasks like 3D navigation and features selection. According to this aim, geographic data are visualized in immersive modality via a stereoscopic see-through Head Mounted Display (HMD, while the interface's layout is overlayed close-at-hand within the user's field of view. The HMD's binocular design is exploited to provide the user with a stereoscopic visualization of the 3D geographic data while minimizing the confusion between the user interface plane and the background thanks to the depth perception enabled. As the HMD orientation/position are tracked, the user is able to see a much greater amount of geographic data by simply turning its head. The interaction paradigm proposed, concurrently exploits both hands to perform precise manipulation of geographic data reducing the interaction effort, allowing both complex actions on 3D features or even more conventional operations (i.e. a keyboard can be visualized within the virtual space and finger tracking allows virtual typing).

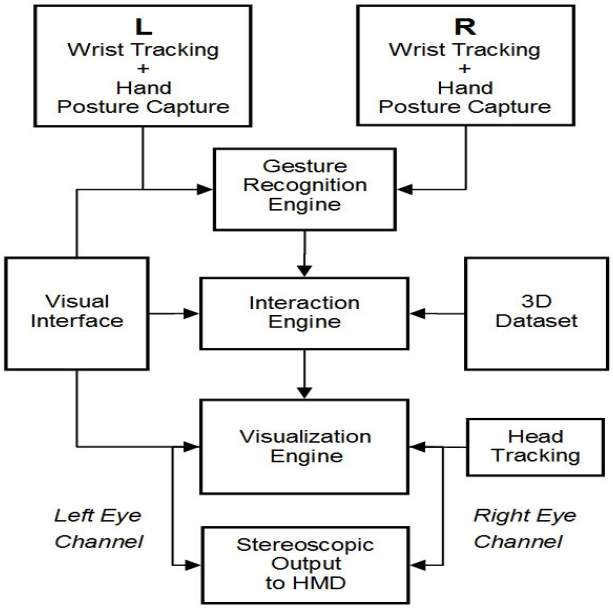


Figure 1. Schematic view of system's architecture.

2. SYSTEM'S ARCHITECTURE

The suggested framework exploits head/wrists tracking plus gesture recognition, in order to provide a more intuitive way of managing visual data and features. This kind of (virtual) direct manipulation, is expected to improve global user performance and increase applications effectiveness. Of course, a careful design is required to obtain this goal. A schema of the overall architecture is shown in Fig. 1. The main hardware and software components of system architecture are described in the following subsections.

2.1 Hands and Head motion capture

The user's wrists/head tracking is performed by means of a wireless ultrasonic solution combined with instrumented gloves for gesture capturing. Such choice provides almost unconstrained hands movements and simplifies the gesture recognition stage respect to an image based system, since in this case inter-hands and inter-fingers occlusions are not an issue: each single finger has individual sensors for flexion and abduction, which are unaffected by any other finger. In our case, four fingers are considered excluding the thumb. Gesture acquisition is performed via a couple of wireless 5DT Data-glove 14 ultra, featuring fourteen channels for finger flexion and abduction measurement each with 12 bit (0 to 4095) of sampling resolution. Additionally a binary (open/closed finger) value, resulting from the comparison of normalized joint flexion values to a threshold, leads to 24 different combinations or gestures of the four tracked fingers. It would be possible to access an even much more accurate finger status representation via each sensor's raw values. However, the adopted classification is more appropriate for the addressed settings, as it simplifies both user training (partially flexed fingers do not compromise gesture recognition) and the design of the GRE module. The four hand gestures used in this study (fist, index finger point, not index finger point and flat hand) are shown in Fig. 3a. They have been selected as they are the simplest to perform for most users, and among the most used in natural interaction. As data-gloves do not provide any spatial information, the system relies on Intersense IS-900/VET ultrasonic motion tracking hardware with six degrees-of-freedom, to detect head and wrists position in 3D space, as well as their rotation on three axes (yaw, pitch and roll).

2.2 Gesture Recognition Engine (GRE)

GRE checks for particular predefined posture tokens, which trigger associated interaction activities. The performed analysis is based on timed automata, able to detect one-hand and two-hand timed posture patterns, which are associated to manipulation functions. Timed automata are labeled transition systems used to model the behavior over time of single components in real-time systems. Classical state-transition graphs are further annotated with timing constraints. Accordingly, a timed automaton performs time-passage actions, in addition to ordinary input, output and internal actions. In more detail, a timed automaton is a standard finite-state automaton extended with a finite collection of real-valued clocks. The transitions of a timed automaton are labelled with a guard (a condition on clocks), an action, and a clock reset (a subset of clocks to be reset). Intuitively, a timed automaton starts execution with all clocks set to zero. Clocks increase uniformly with time while the automaton is within a node. A transition can be executed if the clocks fulfill the guard. By carrying out the transition, all clocks in the clock reset will be set to zero, while the remaining keep their values. Thus transitions occur instantaneously. Semantically, a state of an automaton is a pair composed by a control node and a clock assignment, i.e. the current setting of the clocks. Transitions are either labeled with an action (if it is an instantaneous switch from the current node to another) or a positive real number, i.e. a time delay (if the automaton stays within a node letting time pass). A timed automaton accepts timed words, i.e. infinite sequences in which a real-valued time of occurrence is associated with each symbol. In this way, timed automata provide a feature that classical finite automata do not address in any way, namely timing. Embedding time allows to change the status of involved entities according to time-based events; this enhances the quality of user-system interaction when a feedback is required in a reasonable time, or when the time elapsed between elementary actions can influence the interpretation of their composition. The aim here is to augment the basic one-hand postures through timed patterns, or via a combination of left and right hands for a simple yet more powerful interaction. The use of timed automata offers a further key benefit for the proposed architecture, as it enables the framework designer to formally verify the interaction model by means of well established model checking procedures. In our framework, eight gestures are used as shown in Fig. 2. Four of them are defined as basic postures (Fig.2a), two are defined through a two hand combination of basic postures (Fig. 2b), while the last two (Fig. 2c) are defined by a timed sequence of basic postures (for instance, fist-flat_hand-fist, or double pointing). Recognized gestures are represented by a vector including gesture index, first hand x-y-z spatial coordinates, first hand yaw-pitch-roll angles, second hand x-y-z spatial coordinates, second hand yaw-pitch-roll angles.

2.3 Interaction Engine (IE)

Each time a valid gesture is fully recognized by the GRE, the corresponding vector is provided to the Interaction Engine (IE), which deploys a similar architecture. It is also based on timed automata and is responsible for any visual interaction allowed by the system, by translating gestures into actions. Gestures are evaluated according to the current interaction status, so that the same gesture may trigger different actions in different operative contexts (rotation, zoom, pan, measurements, features selection, etc). Operational modes and data management functions are

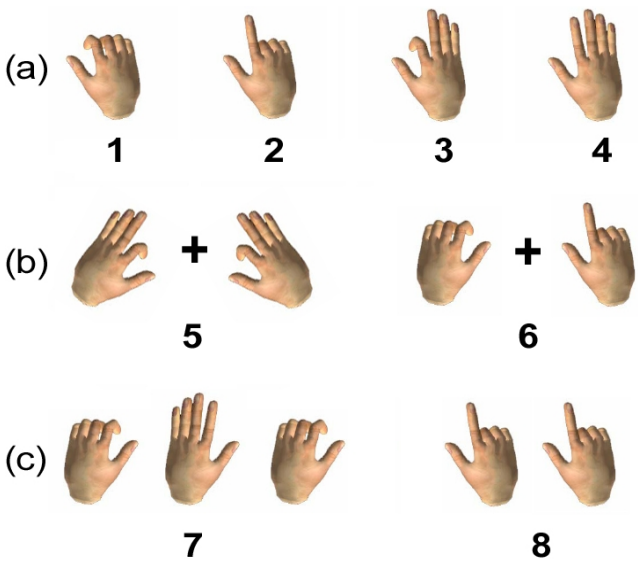


Figure 2. One-hand (a), two-hands (b) and (c) time-based gestures.

selected via a virtual interface. The virtual interface's is superimposed to the 3D geographic data, and may include both graphics and textual information related to the ongoing operations (numerical values for coordinates and angles, distances, etc.). Thanks to the stereoscopic rendering, the interface is visually perceived as a floating object, separated from the actual data. It is displayed in a close-at-hand position along the depth of the visual field, in agreement with the results of a calibration procedure. During such procedure, the user touches with the finger a sequence of small targets positioned at various depths, thus allowing an adaptation of the parameters that regulate the stereo effect. One of the main aims of the interface design is to minimize the number of gestures required to operate it. For this reason a gesture adaptation of the classical point-and-click interaction paradigm is adopted: selection is triggered hitting an active area by index fingertip (see Fig. 2, gesture #2), an action or a confirmation is triggered by double hitting (see Fig. 2, gesture #8), a cancel/escape command is triggered by a fist-flat_hand-fist sequence (see Fig. 2, gesture #7). Though only one-hand gestures are used to operate the interface, the design would allow an experienced user to exploit both hands to operate in a faster and more comfortable way (for instance, typing characters in a text field by both hands results in a much faster operation than via a mouse-like character-by-character selection, and would not require a physical keyboard). Visual and acoustical feedbacks are provided to confirm the "pressure" of a key or the acknowledgment of a particular command, thus reducing wrong operations. If required, interface layout can be hidden at any time via a gesture toggle. At present, only a basic set of 3D actions has been implemented, allowing to pan/rotate the geographic data displayed, to select features and to take distance measurements between them. Panning is conventionally achieved with any of the two hands. Data rotations in 3D space are the actions that fully highlight the advantage of two-hand gestures. They are addressed by associating dataset rotation over three axes (though a user may lock one or two of them) with the rotation of a vector connecting the two points of contact between index finger and thumb on each hand (see gesture #5 in Fig. 2). This approach is much more precise than the typical one-hand based interaction often seen, for instance, in other typical VR applications. As a matter of fact, it allows a greater control of rotation (two hands better visualize the overall rotation), a more comfortable operation (it does not matter how each wrist rotates during interaction, as only the vector connecting the two hands is relevant), and a less jerky interaction, yet without losing responsiveness (a weighted average of both hands spatial information improves tracking). Once the rotation operation has been selected, the user can set the rotation handle for each hand (the anchor points used to interact with the model) by gesture #3 in Fig. 2 by simply moving the fingers along the dataset surface. A valid handle location, i.e. one that lies within a valid region, is highlighted by a colored spot. If the gesture #5 (Fig. 2) is recognized, then the vector connecting the two handles is visualized and the interactive rotation is performed until this condition is true. Feature selection may be accomplished according to two different operation patterns. In the first case, a rotation of the dataset is performed as explained before, to expose the location of interest, and then a feature is selected by double hitting (gesture #8 in Fig. 2). A more intuitive, though less precise, operation pattern requires to perform the rotation of the object by a single hand, grasping the object (see gesture #1 in Fig. 2) in a position which acts as the pivot point, and double pointing the location on which the landmark is to be assigned. For any task involving positioning in 3D space, a precise calculation of the actual finger positions is performed through a combination of a forward-kinematics approach applied to a 3D parametric hand model, which is adapted to the real user's hand measures during a calibration session. In this case the raw flexion values are exploited for each finger. This setup is performed only once, and may be saved and retrieved during system start up. In some situations the need arises to perform classical keyboard-and-mouse supported operations, such as selecting an object from a menu of models or entering parameters for a complex operation. On one hand, switching to a "real" device would break off the interaction flow in a disturbing way. In other words, the user might need to move or change position to reach a different place or equipment, so interrupting the task operation flow. On the other hand, we must consider the way the user is accustomed to perform such operations. Consistency with familiar interaction patterns is a very important guideline, aiming at reducing the time and mental efforts needed to learn a new application. For this reason, a virtual keyboard is projected onto the actual environment, so that the user

can comfortably switch from manipulation operations to system interaction operations without physically shifting his/her locus of attention/operation. Moreover, the familiar point-and-click pattern is re-proposed through fingers movements, to extend the range of available functions while keeping the number of basic gestures to a minimum.

2.4 Visualization Engine (VE)

3D objects, as well as the virtual interface, are processed by the Visualization Engine (VE), which is responsible for real time transformation and stereo rendering of 3D scenes. It manages all the graphic tasks typically required to convert 3D data in a sequence of rendered frames. Similarly to the previously described engines in subsections 3.2 and 3.3, the VE is built on the Quest3D graphics programming environment and on the underlying DirectX API. Quest3D is a commercial well known and very flexible authoring environment for real-time 3D applications. Its most peculiar features are the graphical nature of the software environment and the "edit-while-executing" philosophy. The main advantage is that no compilation step or script interpretation is necessary, because the channels consist of compiled code which is therefore performance-wise comparable to conventional compiled programs, as opposed to interpreted scripts or to applications executed on a virtual machine.

3. First Experiments

In order to validate the visualization/interaction approach proposed, a first experiment has been conducted by integrating the framework described in the previous section with a real world GIS application concerning environmental protection issues. The resulting prototype application has been developed on the base of the Geographic Resources Analysis Support System (Grass) [6], an open source GIS. Figure 3 shows a sample screen of the application.



Figure 3. The user's virtual hands visualized within the GIS 3D dataset while a feature selection is performed.

The test-bed hardware used for system's evaluation is based on a workstation including two quad-core Xeon Processors coupled with a Nvidia Quadro FX-5600 graphics board featuring 128 parallel cores and 1,5 GB of VRAM and outputting to a Cybermind Visette HMD. During experiments, this hardware setup has been capable to render in stereo complex 3D dataset at an output resolution of 2x1280x1024 pixels, with a frame rate always above 30 fps. As user testing is crucial to evaluate the performance of the proposed system, we prepared a user questionnaire to asses the perceived quality of the interaction after performing a number of tasks. The first evaluation sessions involved eleven users, nevertheless the obtained results are encouraging. All subjects reported an advantage in either achieving the right landmark placement, or in the features selection, with an overall good confidence feeling during interaction.

4. Concluding Remarks

In this paper, we presented a framework for gesture-based interaction aimed at 3D geographic data manipulation in a virtual reality environment. The proposed architecture exploits easy to perform one-hand, two-hand and timed interaction patterns, combined with a virtual floating interface, enabling a wider set of commands and functions. The obtained results, although on a small group of users, are encouraging. All subjects reported an advantage in either achieving the right feature selection, or in navigating the 3D dataset, with an overall good confidence during interaction. As we are still in an early stage of this research, we are working on an extension of the basic gesture operated functions implemented so far, as well as on a more objective and in depth interface evaluation.

5. References

[1] Brooks, F. (1988). Grasping reality through illusion - Interactive graphics serving science. ACM CHI p. 1-11.

[2] Menno Jen Kraak. Visualization viewpoints: Beyond geovisualization. Computer Graphics and Applications, IEEE, 26(4):6{9, 2006.

[3] Alan M. MacEachren, Mark Gahegan, William Pike, Isaac Brewer, Guoray Cai, Eugene Lengerich, and Frank Hardisty. Geovisualization for knowledge construction and decision support. IEEE Comput. Graph. Appl., 24(1):13-17, 2004.

[4] Norbert Haala, (2005) Toward Virtual Reality GIS, Photogrammetric Week 05, Dieter Fritsch Ed.

[5] Huang B.; Jiang B.; Li H, An integration of GIS, virtual reality and the Internet for visualization, analysis and exploration of spatial data. International Journal of Geographical Information Science, Volume 15, Number 5, 1 July 2001 , pp. 439-456(18)

[6] Geographic Resources Analysis Support System (Grass), website, http://grass.itc.it/index.php

Augmented Map navigation through Customizable Mobile Interfaces

D. De Chiara, L. Paolino, M. Romano, M. Sebillo, G. Tortora, G. Vitiello

Dipartimento di Matematica e Informatica
Università degli Studi di Salerno
Salerno, Italy
{ddechiara, lpaolino, marromano, msebillo, tortora, gvitiello}@unisa.it

***Abstract*— In this paper we present Framy-AR, an integrated solution for mobile devices which combines the potential of augmented reality with the Framy ability to visually analyze and present spatial data. The rationale behind Framy-AR is to enhance mobile users' capabilities by providing them with information about the surrounding environment in terms of features distribution and pathways. In particular, Framy-AR improves map contextual information by taking into account user's preferences. Moreover, it constantly guides the user during navigation by linking static and dynamic data presented through an advanced multimodal interface. The resulting prototype reveal enhanced properties also in terms of support for visual or motor impaired people, who can benefit from both the customized calculation of the best route and the way such information is delivered through the mobile device, e.g. audio, visual and tactile modality.**

Keywords: Geovisualization; mobile devices; geographic data; geographic information systems;

I. INTRODUCTION

The pervasive nature of mobile devices has stimulated developers to realize applications useful for human daily activities, whose interests are focused on supporting users and improving their life style through immediate and light solutions. In particular, the availability of pervasive integrated tools capable to send, receive and handle information has encouraged the development of advanced solutions in several application domains, ranging from cultural heritage to environmental protection, to emergency management to entertainment and edutainment.

The interest towards such a wide collection of applications is leading researchers towards new challenges, embracing the adoption of suitable technology, the identification of efficient and effective algorithmic solutions and the design of usable and powerful user interfaces.

Within this scenario, the goal of the research we are carrying out is meant to design and develop advanced user interfaces to support users' exploration activities. In particular, we aim at providing users with services where mobility comes into play. This is the case for user's navigation services, where advanced functionalities are devoted to capture a qualitative global view of a scenario, by simplifying the detection of relevant elements.

In previous work in the area of mobile interaction we have introduced an innovative visualization technique, named *Framy*, which exploits an interaction metaphor for picture frames to provide hints about off-screen objects [6-7]. The idea is to visualize a frame along the border of the device screen, where the color intensity of each frame portion represents a visual summary of data located besides the visualized map area. With Framy an appropriate tradeoff is achieved between the zoom level needed to visualize the required features on a map and the amount of information which can be provided through a mobile application.

The system we present here, *Framy-AugmentedReality* (*Framy-AR*, for short), suitably combines the benefits of the Framy visualization technique with those of augmented reality (AR) to enhance user's navigation activities with mobile devices. In the new combined visualization method a semi-transparent colored strip is used to summarize information on the points of interest located inside the camera's visive cone, and augmented reality techniques are used to emphasize those points in the current field-of-view. A customizable interface characterizes the system, thanks to the definition of a new technique for computing the best user-adaptive pedestrian route. In fact, the support to the pedestrian navigation is achieved by considering specific features of the pathway under exploration weighting them against user's profile. Moreover, in order to fully exploit the communication potentials of current mobile technology, we have designed a multimodal interface, including visual, audio and tactile interaction modes.

Previous experience in the area of mobile services by social networks as proven the efficacy of the adopted technological solution. In [1] we present an integrated system for mobile devices, Link2U, which combines the benefits coming from augmented reality with the ability to "communicate" of social networks.

The main advantage that users may gain from Framy-AR is that they can immediately obtain hints about objects located around them. Moreover, weighting elements along a pathway allows to detect the routes which best satisfy requirements derived from user's profile. This latter characteristics makes the resulting application suitable also for impaired people, who may benefit from the customized calculation of the best route and from the way such information is delivered through the mobile device, i.e., audio, visual and tactile modalities.

The remainder of the paper is organized as follows. In Section 2 some related work is discussed. Section 3 briefly recalls salient aspects of the original Framy visualization technique and shows the new Framy-AR multimodal

technique. In Section 4 the customization model of Framy-AR is described. Section 5 presents a working example which illustrates the new Framy-AR functionality. Some final remarks are made in Section 6.

II. RELATED WORK

A valuable work in the area of mobile GIS accessibility has been carried out by Yairi and Igi, who followed a universal design approach to develop their Mobility Support GIS meant for a variety of pedestrians including elderly, disabled, pregnant, and infant users [11]. The routes in their system are represented by lines and nodes. Lines are associated with attribute data of barrier/barrier-free features often found on the sidewalks while nodes are points indicating a nexus with lines carrying different attributes. The gained know-how is proposed in the form of useful guidelines for the development of pedestrian oriented accessible GIS.

An interesting application targeted at physically disabled users was presented in [10]. The authors describe a Web-GIS providing impaired people with pedestrian routes. Each route is divided into segments, so that each segment has homogeneous characteristics. The overall characteristics of a route are determined taking into account the difficulties of each obstacle as well as psychological factors related to the specific user. Obstacles are mapped onto a database and a weight is assigned to each of them, so that users can plan the route and analyze the obstacles carefully. We found this very appropriate for a mobility support GIS and decided to apply a similar technique also in the customization model of Framy-AR. Also, the distinction between discrete and continuous obstacles has been retained in the model. However, in Framy-AR weights not only depend on user's ability, but also on user's interest in the target destination and on user's degree of disability. Furthermore, the addition of the tactile and auditory interaction modalities makes our system also suitable for visually impaired users.

Recently, growing attention is being devoted by researchers to the adoption of augmented reality techniques for mobile devices, as a means to guide users or to assist them in the recognition of objects [2,8,4,3].

The AR technique we present here is based on user's position and on sensors providing orientation on the mobile device. It was partially inspired by the pedestrian navigation system presented in [2]. Based on augmented reality techniques, the user is capable to see the POIs superimposed on the captured video image, he can recognize their location outside the view limits of the camera. The same technique is adopted in Framy-AR, better targeting the use of the original Framy orientation technique to all kind of mobile pedestrian users.

In [8] the authors exploit technologies related to augmented reality and geo-localization to provide visually impaired users with pedestrian routes. The equipment used for the purpose consists of a wearable computer, a microphone and a headphone. Outdoor user's position is provided by a GPS receiver. Indoor position, instead, is retrieved through the ultrasonic receivers. Similarly, our prototype uses a GPS receiver in outdoor areas but it uses a wi-fi hotspot triangulation technique for indoor areas. Finally, the system is smart-phone based.

Also inspired to the benefits of augmented reality, it is worth mentioning the system presented in [4]. The authors propose a walking guide on mobile devices presented as a sequence of "augmented photographs". Geotagged photographs are retrieved from a database and are automatically augmented with navigational instructions.

Finally, a domain-specific augmented reality application is described in [3]. The authors propose a software for mobile devices capable of recognizing cataloged objects in a museum. The technique is based on user's position and on image recognition algorithms.

III. THE *FRAMY-AR* MULTIMODAL TECHNIQUE

The Framy visualization approach is explicitly targeted to enhance geographic information visualization on small-sized displays. The rationale behind Framy is to display a semi-transparent colored frame along the border of the device screen to provide information clues about different sectors inside or outside the screen. For a given POI distribution, the color intensity of each frame portion is proportional to the number of resulting objects located in the corresponding map. Thus, the frame may indicate both the distance and the direction of specific Point Of Interests (POIs). It may also represent the amount of POIs located towards a specific direction. In general, the frame portion color intensity represents a summary of data located besides the screen border. Moreover, the number of frame portions can be interactively increased for better indicating, e.g., the direction where certain objects can be found. A further improvement is to visualize numerical values over each portion of the strip so as users may have a quantitative view of the distribution. In the following of this paper, we avoid to use this increased view in order to highlight the usability of the visual strip.

Without providing more details about Framy, which is largely discussed in some previous work [6], in the following of this section we focus our attention on the new augmented modality of Framy-AR.

Framy-AR consists in a project which merges the concepts introduced with Framy into an augmented reality environment to improve the perception of the objects or objects properties on a mobile application characterized by a strong geographic component.

Recent advances in technology have been urging IT researchers to investigate innovative multimodal interfaces and interaction paradigms, able to exploit the increased technological power as explained in [9]. In Framy-AR, users may benefit from two different visualization modalities, namely Basic and Augmented. The former corresponds to the basic two-dimensional Framy view while the latter exploits augmented reality features to improve users' sensory perception of the context. Users can switch between the Basic mode and the Augmented mode by laying down and drawing up the device.(see Fig.1)

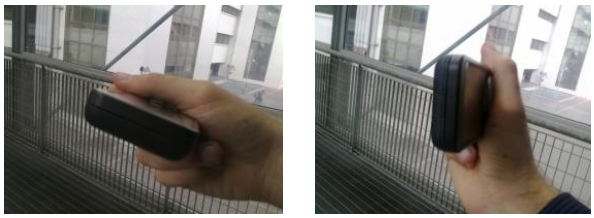
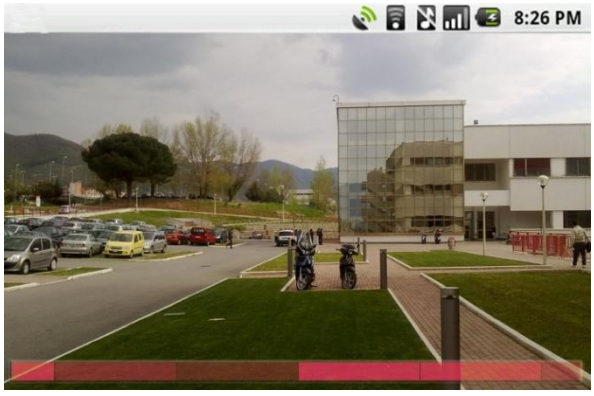
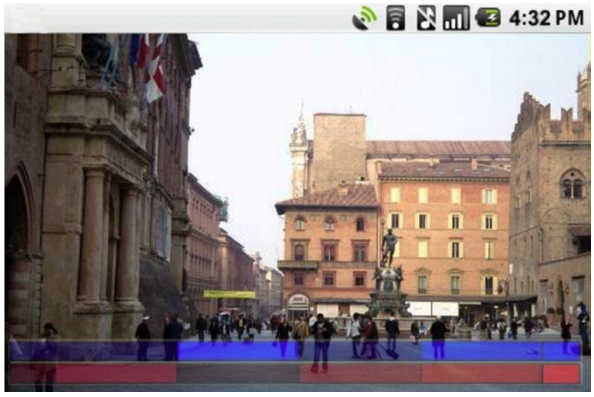


Figure 1. Device positions for Basic Mode and Augmented Mode

With respect to the Basic mode where information is presented by physically inscribing a colored frame all around the border of the screen, in the Augmented mode information is described and visualized by means of a strip which is depicted on the bottom of the display (see Fig.2).



(a)



(b)

Figure 2. Framy-AR in Augmented Mode.

The idea behind Framy-AR is to divide the space around the user into n sectors S1,…, Sn of equal width, 360°/n. Figure 3 gives an idea of this subdivision using n = 16. The surrounding physical environment is augmented by a semi-transparent colored strip which is divided into the corresponding n portions C1,…, Cn. Once the user has a POI distribution, the color intensity[1] of each Ci, is modified on the basis of the value of an aggregate function such as sum, count, etc., applied on a property of the objects located in Si. Thus, the color intensity may be expressed as

$$intensity(Ci) = f(g((Si)) \text{ for each } i \in \{1, \ldots, n\} \quad (1)$$

where f is a monotonic function and g is a function which aggregates a numeric value starting from a set of spatial data.

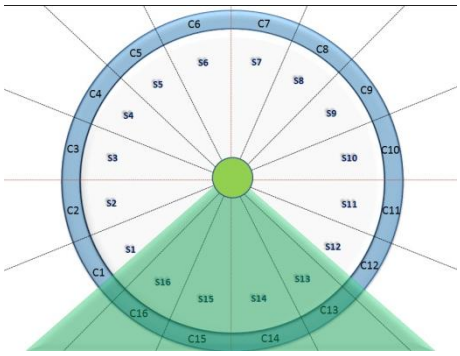


Figure 3. An example of screen subdivision accomplished by Framy and visual cone of user

When in Augmented mode, users will see only the sectors of the surrounding environment which are being captured within the camera visual cone, and the projections of the corresponding strip portions on the vertical plane of the display, to convey information about objects lying in the focused sectors. His view will dynamically change as he turns the camera around.

In Figure 2.a, we applied the sum aggregate function on the POIs located in the University of Salerno (Italy). It is possible to notice that most POIs are located on the front right hand side of the user.

When different aggregates of POIs are needed (e.g., hotels and monuments), Framy-AR can visualize the colored strips representing each of them separately so that the user may summarize information from the displayed strips and make suitable decisions, e.g., about the best direction to the target destination. As an example, in Fig. 2.b the aggregates of hotels and of monuments are considered. Besides taking a glance at the map, the user may want to understand in which direction he should go in order to find the easiest-to-reach area where both hotels and monuments lie. In order to derive this summary information, two strips are visualized – red for monuments and blue for hotels, respectively. The user will notice that the lower red strip has the highest color intensity on the right-hand side of the map, and that the upper blue strip has the highest intensity on the same side. This indicates that, in response to the request, the user may expect many POIs moving the map focus to the right.

In order to extend the capability of Framy-AR to transmit information about the surrounding spatial data, the semitransparent strip can be used to convey information clues both about POIs lying inside the camera's field-of-view cone and those outside it.

As an example, in order to enable a comprehensive view of all surrounding objects, the whole surrounding strip may be visualized on a logarithmic scale (see Fig.4). The scale was chosen to give a priority to the elements inside the camera's

[1] the intensity corresponds to the saturation index in the Hue-Saturation-Brightness (HSB) model.

cone. So, a scale of measurement that uses the logarithm of a physical quantity instead of the quantity itself is visualized.

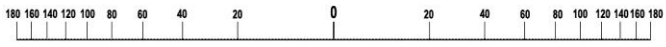


Figure 4. Logarithmic scale for 360° view

As shown in Figure 4, degrees are displayed on a logarithmic scale from 0 to 180, both right and left, summarizing the most significant information coming from the front by more extended intervals.

The trend of intervals can be customized by the user changing the base of the logarithm that modifies the layout and size of intervals. In the figure the default value 10 was left for of logarithm base. A user may choose either to represent the intervals using a logarithmic scale, or to display only equally sized portions of the circular strip that correspond to the captured sectors.



Figure 5. Augmented Mode with Logarithmic scale

In order to allow the adoption of Framy-AR by individuals with hearing or sight impairments, we decided to associate sounds and vibrations as additional tips of the device. The idea is that a user can use a device as a detector that identifies objects. When a user directs the device in the suggested direction, he will perceive an increasing frequency of a beep or of the vibration.

Given a data distribution, users can direct the device towards the suggested direction and hear a sound or feel a vibration whose pitch is synchronized with the color intensity of the corresponding frame portion. Each frame is mapped onto a different signal (sound or vibration) and each frame portion onto a different signal intensity, computed on the basis of the corresponding color intensity thus conveying the same information as the visual counterpart.

IV. THE CUSTOMIZATION MODEL

In the area of pedestrian routes calculation, differently from car route, in order to reconstruct a good pedestrian route it is necessary to take into account possible obstacles to the user based on his physical abilities [8, 10, 11]. Thus, in order to adapt the system to the specific requirements of possibly impaired users, we enriched the system with a customization model which allows to highlight those sectors which better respond to user's accessibility needs and preferences. To this aim, we take into account the accessibility score of each POI *a* belonging to a sector *S* computed on the basis of the conditions of the roads conducting the user *u* to that POI and the parameters characterizing *u*. Formally,

$$Score(S,u) = \sum_{a \in S} Score(a,u)$$

where *Score(a, u)* is the score that the model computes for a POI *a*. Basically, our customization model measures the minimum effort that any user has to experience for reaching a given destination. The basis of this model is the length of the paths leading from the user's location to the destination. These values are successively increased on the basis of the path conditions or decreased on the basis of the user's preferences:

- we penalize a path in a way proportional to the difficulties that users would face by walking along it, while
- we reward the path on the basis of the specific interests of the user in the destination POI.

From a general point of view, penalizations are specifically due to the presence of obstacles that users find along the paths while rewards are special motivations for the POIs to be reached.

Thus, the score of each POI may be expressed as the difference of two components:

$$Score(a,u) = w_1 P(a,u) - w_2 R(a,u)$$

where *P(a,u)* is the least penalization to reach *a*, *R(a,u)* is the total rewards, while w_1 and w_2 are the corresponding weights whose values belong to [0, 1] summing up to 1. Each weight specifies the importance of a component in the formula. Basically, if either w_i is set to 0 or 1 the corresponding component should be not considered or it is the only feature considered for the final score computation. In order to balance the significance of *P(a)* and *R(a)*, weights may be set to 0,5.

In particular, $P(a,u)$ may be calculated by the following formula:

$$P(a,u) = \min_{i=1}^{t}(d(p_i) + P(p_i,u))$$

where $\{p_1, p_2, \ldots, p_t\}$ are the paths the user may use to reach the POI *a* starting from the his location, $P(p_i,u)$ are the total penalization along the path p_i and $d(p_i)$ is its length.

The final computation of the penalization may be, in turn, divided into two components on the basis of the obstacles typologies that users have to face. From now on, we will distinguish between discrete and continuous obstacles. The former represent obstacles whose degree of difficulty is not dependent on the path length (e.g., stairs or holes) while the latter are dependent on the path lengths (e.g., dirty or steep roads). In general, we defined a pair of matrices (*Disability* x *Obstacle*) where the weight associated to each cell corresponds to the total penalty in case of a discrete obstacle and the unitary penalty in case of a continuous obstacle. Generally, these values may be extracted by exploiting a learning step where the difficulty to overcome the different obstacles are registered by the system when faced.

TABLE I. Weights associated to (a) discrete obstacles, and (b) continuous obstacles.

	D_1	D_2	…	D_n
O_1	W_{11}	W_{12}	…	W_{1n}
O_2	W_{21}	…	…	…
…	…	…	…	…
O_p	W_{p1}	…	…	W_{pn}

(a)

	D_1	D_2	…	D_n
O'_1	W'_{11}	W'_{12}	…	W'_{1n}
O'_2	W'_{21}	…	…	…
…	…	…	…	…
O'_s	W'_{s1}	…	…	W'_{sn}

(b)

More formally, let us suppose that:

- the path *p* is divided into *m* segments $\{s_1, s_2, \ldots, s_m\}$ so that each segment contains an instance of either only discrete $\{o_1, o_2, \ldots, o_k\}$ or only continuous $\{o'_1, o'_2, \ldots, o'_l\}$ obstacles, with $m = l+k$.
- the lengths of the segments containing continuous obstacles are $\{d'_1, d'_2, \ldots, d'_l\}$
- the weights of the obstacles along the path, as defined in Table I, are $\{w_1, w_2, \ldots, w_k\}$ and $\{w'_1, w'_2, \ldots, w'_l\}$, and
- the disabilities percentages of the user *u* are $\{u_1, u_2, \ldots, u_k\}$ for the discrete obstacles and $\{u'_1, u'_2, \ldots, u'_l\}$ for the continuous obstacles, then

$$P(p,u) = \sum_{i=1}^{k} u_i w_i + \sum_{i=1}^{l} u'_i w'^{i} d'_i$$

On the other hand, the interest in POI *a* could be considered as a further motivation for reaching it. Basically, it could be seen as a reward which should be set against the penalties. Rewards are based on two values, which consider a classification of the POI domain *I(a,u)* and the POI characteristics mapped onto user's general interests *Is(a,u)*, respectively:

$$R(a,u) = I(a,u) + Is(a,u)$$

Initially, when building a user profile, the system stores the score the user assigns to each domain as well as to its derived subcategories. For our purpose, we have considered a classification of 7 POI categories. As an example, the domain museum has the item subcategories archaeology, sculpture, painting, etc. Each category is then described by means of a set of terms which are successively used for calculating the similarity with the POI description. When a user *u* initially registers to the services, she/he is asked to assign an interest value to each general category, and an interest value to each of its subcategories. For every domain, information concerning interests assigned by user to the categories and their specific item subcategories are stored as vectors, as shown in Table II.

TABLE II. User-specified interests for the (a) categories, and for (b,c) their subcategories.

Cats	C_1	…	C_m
u	i_1	…	i_m

(a)

C_1	$S_{1,1}$	…	$S_{1,n1}$
u	$i_{1,1}$	…	$i_{1,n1}$

(b)

…C_m	$S_{m,1}$	…	$S_{m,nm}$
u	$i_{m,1}$	…	$i_{m,nm}$

(c)

As each POI has a pre-assigned subcategory, selection with respect to this reference is immediate. Each POI a is assigned the score associated with the corresponding specific category in the corresponding user profile. Thus, the interest in a POI *a* of a user *u*, classified as belonging to subcategory $S_{i,j}$, corresponds to the interest assigned to the subcategory $S_{i,j}$ by the user *u*, namely:

$$I(a,u) = S_{i,j}(u)$$

In order to calculate the second part of the reward formula, a similarity value $sim(a, C_i)$ is then computed between the description of the POI *a* and the general categories C_i, by exploiting the cosine similarity formula for the vector space model. In particular, in this work we used a simple approach where each term weight of the POI description is set to 1 if the normalized term frequency is higher than a given threshold. Consequently, the relevance between a POI *a* and all the general categories of a user model *u* is computed using the following formula:

$$Is(a,u) = \frac{\sum_{i=1}^{7} \mathrm{sim}(a, C_i) i_i(u)}{\sum_{i=1}^{7} i_i(u)}$$

V. A WORKING EXAMPLE

Let us suppose that two users u_1 and u_2 have to reach the city town hall from the train station. User u_1 has a partial visual impairment while u_2 has a motor impairment and uses a wheelchair.(Table III) Basically, three paths, p_1, p_2 and p_3, are available from the starting point, the length of p_1 is 100m, p_2 is 160m and p_3 is 200m (Figure 6). Two discrete obstacles are located along path p1, a continuous obstacle is located along p2 for the entire length and, finally, there is no obstacle in p3.

TABLE III. Visual impairments for u1 and u2.

	O_1	O_2	O_3	$\%u_1$	$\%u_2$
visual	30	40	0.2	1	0
motor	50	60	0.5	0.1	0.7

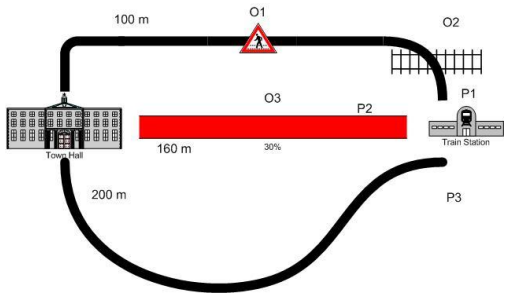


Figure 6 A schematic example of the obstacles located along the paths for reaching the town hall from the train station.

$$P(Town\ Hall, u_1) = min_{t=1}^{3} P(p_i, u_1)$$

Then,

$$P(p_1, u_1) = 30 * 1 + 40 * 1 + 30 * 0.1 + 40 * 0.1 = 77$$

$$P(p_2, u_1) = 0.2*50 + 1* 0.2 = 11$$

$$P(p_3, u_1) = 0$$

and,

$$P(p_1, u_2) = 30 * 0.7 + 40 * 0.7 = 49$$

$$P(p_2, u_2) = = 7$$

$$P(p_3, u_2) = 0$$

Finally,

$$P(Town\ Hall, u_1) = \min\{100 + 77, 160 + 11, 200 + 0\} = 171$$

$$P(Town\ Hall, u_2) = \min\{100 + 49, 160 + 7, 200 + 0\} = 149$$

As a matter of fact, it is possible to notice that even if the shortest path between the train station and the town hall is p_1, the topology of the obstacles along the paths changes the resulting scores. In our example, the impairment of u_1 with the obstacles along p_1 increases its distance of 77 meters which is much more than the 11m calculated along p_2. In the same way, the impairment of u_2 modifies the real distances of p_1 and p_2 by adding the increments due to the obstacles, in this case, 49 for p_1 and 7 for p_2. Given this obstacle and impairment configuration, the path with the shortest calculated distance for u_1 becomes p_2 and the differences between any pair of calculated distances (177, 171, 200) are smaller than the distances without obstacles (100, 160, 200)). On the other hand, p_1 continues to be the best desirable path for u_2 As a matter of fact, the penalization calculated along p_1 is not sufficient to make p_2 the best path.

VI. FINAL REMARKS

Our previous work in the context of mobile interaction has led to the introduction of the innovative visualization technique, named *Framy* to provide hints about off-screen objects by displaying semi-transparent colored frames along the border of the device screen. In [5] Itten's theory of colors was investigated to enhance the quality of the information conveyed to the user. Indeed, we used this method in order to generate the more contrasted colors for helping visual impaired people. In [7] some interactive "sonification" capabilities were exploited in order to make the solely-visual version of *Framy*, useful when users are in awkward visibility situations, such as strong reflection in the screen. The achieved evolution of the initial prototype has laid the basis for the design of the system presented in this paper, Framy-AR. The further enhanced multimodal interface allows to provide support also to visually impaired users, and to enhance their interaction with mobile devices, especially in disadvantaged circumstances. Finally, the adoption of a customization model extends the use of the system to all kinds of disabilities.

As future work, we plan to analyze the results of a longitudinal study we are carrying out with different categories of users (including elderly people, children, impaired people, etc.), meant to evaluate Framy-AR against usability and accessibility factors. Moreover, we plan to extend the system functionalities so that expert users may formulate appropriate requests to the underlying geographic information system and receive rapid feedback from the multimodal augmented interface characterizing Framy-AR .

REFERENCES

[1] De Chiara D., Romano M., Sebillo M. , Tortora G. , Vitiello G., "Link2U: An augmented social network on mobile devices". *Proc. of the Iadis Int. Conf. Applied Computing* Rome, Italy, vol. 1, p. 283-290, Nov. 2009.

[2] Braun M., Spring R., "Enkin". Tech. Rep., University of Koblenz-Landau,Website: http://www.bassignana.net/mediawiki/images/6/64/Enkin.pdf

[3] Bruns E., Bimber O., "Adaptive training of video sets for image recognition on mobile phones", J. *Personal and Ubiquitous Computing*, vol. 13(2), Springer-Verlag London, UK, p. 165 – 178, Feb. 2009.

[4] Hile H., Vedantham R., Cuellar G., Liu A.L., Gelfand N., Grzeszczuk R., Borriello G., "Landmark-based pedestrian navigation from collections of geotagged photos". *In Proc. the 7th Int. Conf. on Mobile and Ubiquitous Multimedia, MUM '08*. ACM, New York, 145–152 , 2008.

[5] Itten, J.; *The Art of Color*, New York, J. Wiley & Sons, Inc. 1961.

[6] Paolino L., Sebillo M., Tortora G.,Vitiello G., "Framy – Visualising geographic data on mobile interfaces", *J. of Location Based Services* Volume 2(3), p. 236-252, Sep. 2008.

[7] Paolino L., Romano M. , Sebillo M. , Tortora G., Vitiello G., "Audio-Visual Information Clues about Geographic Data on Mobile Interfaces". *Proceedings of PCM 2009*. IEEE, pp. 1156-1161, 2009

[8] Ran L., Helal S., Moore S., "Drishti: an integrated indoor/outdoor blind navigation system and service," *2nd IEEE Conf. on Pervasive Computing and Communications. PerCom 2004. Proc.the*, vol., no., pp. 23- 30, 14-17 March 2004.

[9] Sebillo M., Vitiello G., De Marsico M., "Multimodal Interfaces". *Encyclopedia of Database Systems*. LING Liu And M. Tamer Zsu (Eds.), pp. 1838-1843, Berlin Heidelberg: Springer Verlag ISBN/ISSN:.978-0-387-35544-3, 978-0-387-3.

[10] Izumi S., Kobayashi G., Yoshida T., "Route navigation method for disabled access GIS in consideration of abilities and psychologies," *Proc. 2nd International Conference on Digital Information Management*, pp.519-525, 28-31 Oct. 2007.

[11] Yairi I.E., Igi S., "Mobility Support GIS with Universal-designed Data of Barrier/Barrier-free Terrains and Facilities for All Pedestrians Including the Elderly and the Disabled," *IEEE Int. Conf. on Systems, Man and Cybernetics, 2006. SMC '06.*, pp.2909-2914, 8-11 Oct. 2006.

A design methodology for the development of GIS applications

Rosanna Cassino, Monica Sebillo, Maurizio Tucci
Dipartimento di Matematica e Informatica
Università di Salerno
84084 Fisciano, Salerno – ITALY
{*rcassino, msebillo, mtucci*}*@unisa.it*

***Abstract*— The interest in new methods to easily and efficiently design and implement graphical user interfaces is increasing due to a growing need of rapid prototyping tools to construct quality computer-based application systems quickly. In the GIS domain the evolution of technology and the consequent availability of powerful and inexpensive personal and mobile computers poses a series of new challenges for user interface design and development: user interfaces have to exploit the capabilities of various access devices and be suitable for different contexts of use, while preserving consistency and usability. In this paper, we present a methodology to design and implement graphical user interfaces for GIS environments in a guided and totally visual manner. We also describe a system prototype that implements the proposed approach. It derives from the customization of TAGIVE, a visual toolkit to design and develop general purpose interactive visual applications [1].**

Keywords - GUI development tool; GIS interfaces.

I. INTRODUCTION

The evolution of tools to support the development of interactive visual applications enables software systems to manage larger and larger amounts of information and relationships among data, models and documents. The availability of large amounts of information does not guarantee that all the information is effectively accessible and usable to the user, which must often face with not homogeneous environments and useless complexity of data. To experiment a new form of modeling and implementing multimodal visual applications, in this work we propose a methodology that integrates the advantages of a diagram design technique with a visual construction of applications using a component assembly mechanisms. We describe the approach using GIS interfaces development as an example of interactive visual application domain.

In literature there are several graphical toolkits to implement graphical user interfaces [3, 4]. These systems have become popular with programmers, but they provide only partial support for consistency. Often the specification of the interactive runtime behavior of the system is a cumbersome and tedious task. As a matter of fact, programming language fragments have to be associated with the GUI interactive components, to handle the occurrence of related events.

The proposed approach overcomes these limitations allowing a designer to define the GUI of a software application by combining diagrams, whose nodes are containers for components and whose edges represent the interaction paths from a component to another. Consequently, the application interface is describes in terms of a set of *interactive scenes* and a set of *interaction paths*.

GIS systems are an application field where the need of guidelines to develop usable interfaces is very relevant.

Several application domains, related to the study and the management of the territory, find in the GIS the possibility to integrate data represented in very different format, scale and system of reference and to elaborate and to analyze them according to determined cognitive objectives [16].

The evolution of technology and the consequent availability of powerful and inexpensive personal computer have widely contributed to the dissemination of GISs. Recently, Internet has contributed to it, thanks to its capillary ramification and its facility in the use. In fact, geographical data distribution (Geo-Data Server), maps visualization (Map Server) and also Web GIS processing applications are largely available [2]. The wide range of different devices with varying capabilities and interaction modalities as well as changing user context in nomadic applications, poses critical challenges to the design of user interfaces.

Examples of GIS applications are: management of the architectural, rural, historical, artistic, archaeological properties. GIS allow to address in an effective and innovative way the strategies of management of the Cultural Properties in the context of the territorial planning. It allows to monitor zootechnical companies and to realize GIService to establish buffer zones related to the spread of specific diseases [15]. In this perspective, the present research concerns the development of the elaboration environments characterized by a user-friendly interface that allows an expert users of the domain to perform and/or customize applications and geo-processing tasks. The advantages are the possibility to follow the several phases of the development process of the GIS application and to easily modify them by reusing components previously realized. The use of open source software to implement several functionalities (Java, MySQL, MapServer, ecc…) reduces the licence costs both for the use and the development.

In this paper, we present a system to support a (non-programmer) domain-expert user in modeling an interactive application in a guided and a totally visual manner. As a designer, the user is in charge of defining objectives, analysing

data and functionalities, specifying the features of the graphical interface in terms of objects, tasks and visual interactions.

The tool derives from an extension and customization of the TAGIVE system, an environment to specify, design, and implement two-dimensional interactive visual applications [1]. In particular, starting from a top-level design of the visual environment, a usable form-based interface of the tool allows detailing the graphics aspects of each component. A similar visual technique allows designers to manage the interactive mechanisms, so that all the features related to an event (i.e., the source object, the action performed on it, and its effect on the application itself) can be easily specified.

The rest of the paper is organized as follows. In Section 2 we introduce the features of a GIS system and the importance of the development of a usable interface to interact with it. Section 3 describes some existing tools and methodologies to develop GIS interfaces. In section 4 we present the proposed methodology to design and implement GIS interfaces in a guided and totally visual manner. In section 5 we briefly describe an example of using the implemented prototype to develop a map viewer. Section 6 presents some conclusions and further researches.

II. GIS SYSTEMS

Geographical Information Systems are suitable tools to manage geographical and georeferential data. GISs offer a better comprehension of the process and the factor related to an area connecting several information levels of it. The number of levels depends on the purpose of the application. There are many application domains where GISs are applied, for example environmental management, safeguard of cultural patrimony, technological nets management, satellitar navigators, GPS software, interactive map gateway etc... Cartography plays a fundamental role in every GIS application; as a matter of fact the nature of maps influences notably the quality of a system. The map is the fundamental tool to represent territorial information. Using a system of symbols, it is possible to represent a model of the real world (basic information) and the phenomena located (thematic information). On the base of this classification we define basic or thematic maps [2]. The map is not only a possible input, but it is often also the output of an informative geographical system. The system of reference is an important characteristic of a map. It allows to individualize a univocal correspondence between the geographical and the planar coordinates. The *datum* defines the system of reference. It identifies the mathematic model used to compute the geographical coordinates of the single point [10].

The datum and the type of *projection* are the main properties to define a map. In fact, to visualize geographical objects in a map involves representing the three-dimensional features of the terrestrial surface in a bi-dimensional view. This operation takes the name of projection, that is a conversion from a system of geographical reference to a Cartesian one. Each projection involves an inevitable distortion, then, on the base of the preserved properties of the elements, we have the following classification of projections:

- Equivalent: the area of the represented objects remains unmodified.
- Equidistant: the distance between objects remains unmodified.
- Conform: the shape remains unmodified.

The actual projections, obtained by algorithmical procedures, are classified on the base of the light position, of the location and shape of the plane and of the center size [2].

The GIS functionalities could not be used without an easy and adequate human-computer interaction system. There are several GIS platforms with procedures accessible using interfaces related to the specific systems and to the mental model of the end user that uses them. Nevertheless, in the GIS development process an expert user of the domain may not have knowledge about design and implementation; again a designer could not be experienced of the application to realize. Besides, the main GIS platforms are characterized by commercial tools that involve expensive resources and not customizable functionalities.

In this perspective, in this work we present an approach that intends to support novice programmers in the development of GIS applications while reducing expert designers' workload. The GIS are developed using open source technologies so to adapt the systems to the different contexts. The approach integrates the advantages of diagram design with a visual construction of software applications using a component assembly mechanism. In particular, starting from a top-level design of the visual environment, a usable form-based interface of the tool allows detailing the graphic aspects of each *scene*. A similar visual technique allows the designer to manage the interactive mechanisms, so that all the features related to an event (i.e., the source object, the action performed on it, and its effect on the application itself) can be easily specified.

We have exploited such technique enhancing the visual approach to develop the graphic and interactive features of an application interface presented in TAGIVE (Tool for the Aided Generation of Interactive Visual Environments) [1]. The top-down development approach prevents possible incorrectness and incompleteness errors. The modularity of the realized environments guarantees reusability of software components. The visual development model for both static and dynamic aspects makes more intuitive the design and the implementation.

III. THE PROPOSED VISUAL DEVELOPMENT METHODOLOGY

The proposed development methodology is based on the definition of an interactive visual application in terms of a set of scenes and a set of external components, related to each other on the basis of the kind of interactions performed. A scene is made up of elementary components representing interface widgets and arranged on the scene according to a certain layout. Each elementary component appearing in a scene is said to be dynamic, if some event is associated with it or static otherwise. External components are instead any file that can be invoked by some dynamic component of a scene (e.g., image files, text files, video clips, etc…). This model allows a two-level approach in the construction of an application. The former level is focused on the individuation of the main elements of the application and on the interactions between scenes or between scene-to-external components. For each scene, the latter level details its graphical properties and

allows to implement actions and events related to its dynamic components, with the possibility to switch from one level to the other at any time during the development process. Then, the proposed design methodology integrates the advantages of a diagram design with the benefits coming from a visual construction of applications using a component assembly mechanism. The integration supports the development of interactive visual environments by directly relating the design phase to the implementation phase. In fact, a diagram design approach is used to build the "application map", in terms of a top-level transition diagram representing all the possible interaction paths. Indeed, oriented and connected graphs are a very good means to design and examine the number of scenes or external files and the actions performing o*n an interactive visual applica*tion. The idea is that the application map directly guides the development of the interactive environment. A connected component assembly technique characterizes the implementation phase of different scenes and *a visual event-handling mecha*nism is used to implement interaction processes.

A. GIS interface design

A GIS design is articulated according to a logical model divided in several levels [16]:

- Definition of objectives.
- Database implementation.
- Data analysis.
- Presentation of results.

The definition of objectives guides the modelling of the system. In this perspective, at first it is necessary to understand the problem to solve and the type of application that should be released. Then, it is necessary to understand the typology of the end user which will use these systems and if the solution is realizable using general purpose tools or if GIS functionalities are needed. Later, the possibility to reuse application components to realize the desired systems should be considered.

The structuring of the database follows the definition of the requirements and it is a fundamental point in the creation of the GIS. As a matter of fact, the quality of the analysis process and of the query results depend on the content and the conceptual organization of the database. In this phase it is necessary to identify the spatial data to solve the problem, to perform quality controls of the data, to define the data insertion way in the GIS, to identify the alphanumeric properties to link to the spatial data and to arrange the tables, to organize the data on the several thematic layers.

The maps can be connected to the tables, graphics, multimedia representations, animations or other type of external files.

A GIS shows all its potentiality by the spatial analysis functionalities. "Find" or "Info" queries can be performed on the textual or geographic data also in integrated way, on a single layer or on a combination of several layers so to obtain additional informative layers.

On the base of maps, the spatial data and the database produced by the experts of the domain, in the present work we propose an approach and then a system to support the developer to design and implement GIS systems in a guided and totally visual manner.

The released environment integrates the analysis spatial tasks typical of the GIS systems with external components connected to the main application if necessary. An user friendly interface allows to the designer to reuse and then to customize the design and/or the several components of other applications previously realized and to visual implement several GIS functionalities. The functions that can be managed are:

- Map visualization actions. Actions to expand (Zoom in) or to reduce (Zoom out) an image or part of it on the base of a factor. Panning function allows to visualize not visible section of an image selecting a change of position.
- Layers actions. The functionality allows to manage the layers which compose the map (it is possible to insert or to remove a layer, change the order of their visualization).
- Geo data actions. They include all the function to retrieve the information stored in the database connected to the map.
- Spatial Analysis actions. These actions are related to the management of the buffering and overlay functionalities that involve the creation of a zone around of a specified point, line or polygonal area. The result is a new layer, which can be used in any query to determine which entities occur either within or outside the specified buffer zone.

Easy-to-use tools and forms allow a designer to define the graphical features of the GIS scenes to show the results of the queries and of the spatial analysis processes.

In the next section we give details of the proposed development process by the implementation of a simple map viewer. In particular, in the proposed approach we consider ESRI *shape files* as spatial data format [12]. A shape file consists in a main file (with *.shp* extension), an index file (with *.shx* extension) and a dBase table (a file *.dbf*). The main file is to direct access and it contains records of variable dimension, each record describes a shape with a list of its vertices. In the index file, each record contains the offset of the corresponding main file record from the beginning of the main file. The dBASE table contains feature attributes with one record. The one-to-one relationship between geometry and attributes is based on record number. Attribute records in the dBASE file must be in the same order as records in the main file [12]. As described in the next section, in the proposed system the layer handler involves the designer at development phases and the end user (if the possibility to insert new layers is allowed) selects only the path of the main file, instead the index file and the dBASE table are loaded automatically.

To manage spatial data we have extended the original tool TAGIVE with the OpenMap Java OpenSource library. The package presents several useful components to implement applications witch manage geographical information. OpenMap was created in the half of the years 80s [8]. Demonstrations of his validity derives from the fact that it is inside of the Open GIS Consortium from the 1995, and because it is currently used in numerous universities around the world, but also in many important firms and institutes, as well the Department of the Defense of the USA and of Australia.

In particular, we have used the OpenMap libraries to perform the following functionalities:

- zoom and pan management;

- layers management;
- buffer management.

In the following section we describe the architecture of the tool that implements the proposed methodology to design and implement GIS environments.

IV. THE IMPLEMENTED SYSTEM

In this section, we describe the adopted strategies to customize the design and the implementation technique of the standard interactive visual environments so to realize GIS applications maintaining a homogeneous visual development process.

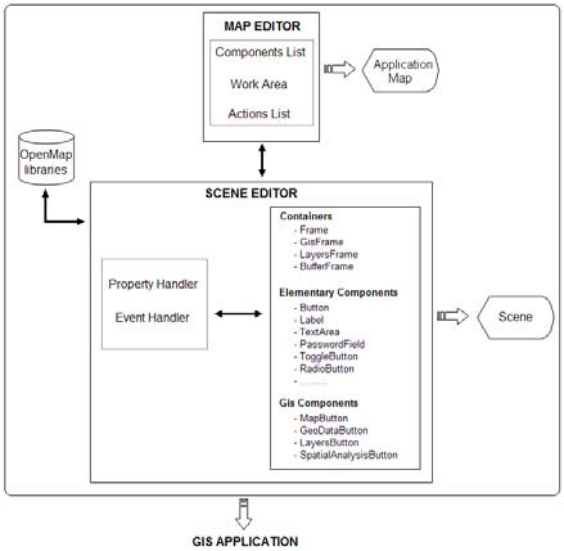


Figure 1. Schema of the implemented system[1].

We have improved the TAGIVE tool [1] by means of the Open Map package. Figure 1 presents the structure of the TAGIVE to develop GIS interfaces.

A. The Map Editor

The Map Editor module represents the top level of the development process. In this subsystem, a Work Area contains the diagram that represents the application map. The nodes identify the components expected in the environment (i.e., frames, panels, *.bmp*, *.txt* and web or multimedia files). The edges represent the possible interactions paths that the end-user can to perform in the visual environment.

It is well known that the efficiency of a diagram approach to software design may decrease when the number of nodes rises, as the diagrams become too complex to manage.

A proper navigation technique is useful to manage maps with a limited number of nodes [4], while clustering is suitable to model more complex applications [3]. As a matter of fact, by the navigation technique the designer can only manage a small subsystem of nodes and edges at a time. On the other hand, a clustered diagram is a diagram G where each node represents a vertex of an recursive partition. Each "sub diagram" is a cluster. The purpose is to preserve the global structure of the underlying graph, clustering more and more large sub-graphs and designs them like single nodes or regions (see figure 2). The advantage to group the vertexes in clusters is that the graph can be divided in levels of decreasing details that could be visited in top-down manner [3].

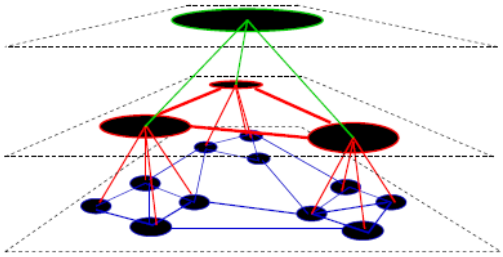


Figure 2. A multilevel drawing.

In fact, in our approach we implement the ability to manage the modeling of an application by navigation, and we are developing capabilities to design the map application gradually according to the methodology of graphical clustering.

In the proposed system the edges of the application graph are labelled with the possible actions that allow to perform an interaction. The considered actions are: mouse single left click, mouse double left click, mouse right click, *map actions*, *geo data actions*, *layers actions*, *spatial analysis actions*. The last four actions are typical of GIS events. In particular, a *map action* labels a loop transition on a GIS frame because it can define an interaction on the GIS map so to perform zooming and unzooming (it is possible to customize the zoom or unzoom level), previous or next visualization, navigation (to move by the arrows) and panning (to move by map dragging). A *geo data action* can perform the opening of the global database table related to the GIS map, the table with the textual information related to a "find", "selection" or "info" query. The interaction to retrieve the data of a particular site on the map is implemented by a "find" action; by a "selection" action the information of a selected area; by an "info" action the end user inserts a textual query key in a defined visual form, than, the GIS highlights the corresponding point on the map. A *layer action* opens the defined frame to visualize the layers available of the GIS map. A *spatial analysis action* can open a defined frame to manage the buffer operations, the retrieve of point at a certain required distance.

The Application Map is the output released by the Map Editor.

B. The Scene Editor

The Scene Editor implements the next step of the development process. The module provides the designer with tools to visually implement any scene of the application to realize. In particular, it provides *Containers*, *Elementary Components* and *GIS Components* Palettes. The *Containers* palette allows to select a general frame or panel that represents a scene, and in this context it is extended by the introduction of a *GIS frame*, a *layers frame* and a *buffer frame*. A *GIS frame* is a customizable frame where the GIS map loads at runtime and where the design specifies and arranges

[1] The shape ▭ indicates a tool module. The shape ⬭ indicates a graphical visualization unit of an interactive visual application. The arrow → denotes the interconnection between tool modules whereas the arrow ⇨ denotes the production of an output.

elementary components and/or buttons related to several actions. The *layers frame* represents the layers legenda where the designer defines the layers available and the eventual buttons that allow to the end user to move a location of the layer label in the list and/or to add or to remove any layer. Similarly, the *buffer frame* is a customizable frame where the designer defines how the end user can manage each buffer functionality.

The *Elementary Components* palette contains the widgets related to several objects (i.e., label, text field, button, text area, password field, radio button, toggle button, ecc…), that the designer can insert in a panel. Similarly, the *GIS Components* palette contains objects specific of the GIS environment. They are: *map button*, *geo data button*, *layers button* and *spatial analysis button*. These components allows to perform the related action specified in the Action List (section 5 presents a detailed description of these aspects).

In the Scene Editor the Property Handler module allows to the designer to define the graphics properties of a scene in totally visual manner. Similarly, the Event Handler offers visual forms to specify the events related to the actions expected in the map application, without the need to write source code manually.

In the following section we briefly illustrate the described characteristics by an example of using the proposed system to develop a simple map viewer application.

The integration of the TAGIVE with the Open Map libraries and its extension is necessary to manage the geographic information related to the GIS applications.

C. The Open Map subsystem

OpenMap [8] allows to integrate different data sources to compose a traditional overlay-based GIS where the map visualization is based on the overlapping of many layers (see figure 3).

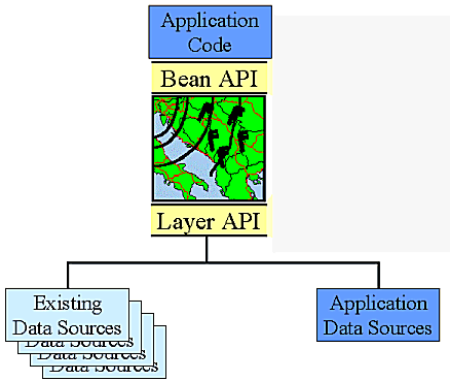


Figure 3. OpenMap data interfacing.

The OpenMap Java Toolkit is made up of Java Bean components, of which the MapHandler, MapBean, and Layer/PlugIn components play a central role.

MapHandler is the basic component that manages and connects OpenMap components in a very flexible manner. MapHandler can be thought of as a conceptual map, which can contain the MapBean and other components that manage the layers, mouse events, and projection controls (see figure 4). MapBean is a drawing canvas that derives from the Swing JComponent class. Since MapBean is a Swing component, it can be added to a Java window hierarchy like any other Swing windowing component. MapBean manages a hierarchy of layers which can paint themselves to the canvas and a projection object to manage the view.

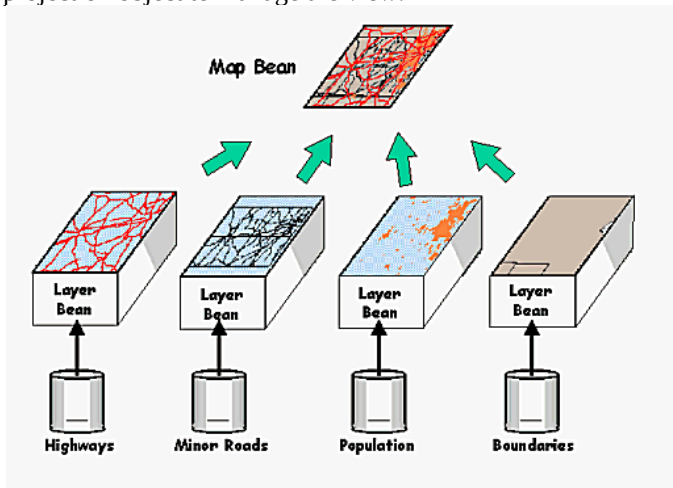


Figure 4. The MapBean class.

BufferedMapBean extends MapBean by forcing its layers to paint their graphics into a buffer. This drawing buffer is then rendered whenever the AWT causes a redraw. This dramatically increases the performance of window showing since it avoids the (potentially expensive) layer painting process. Of course, if a layer causes a redraw, we regenerate the drawing buffer with the graphics and render the new image. Therefore, an OpenMap Viewer application uses BufferedMapBean instead of MapBean precisely to increase the performances.

Layers are components that get added to MapBean in a hierarchical stacking order. The map is drawn by painting the graphics of each layer starting from the bottom one and proceeding up in the hierarchy. Successive layers render their graphics on the top of the graphics of lower ones.

PlugIns are components that are used by PlugInLayer to fetch data and to prepare graphics for the map.

Layers and PlugIns are responsible for acquiring, constructing, and rendering their own graphical data. The OMGraphics package provides a simple way to construct vector and raster graphics out of geo-spatial and XY data. They can also be interactive, by reGIStering for mouse events and constructing their very own GUI widget controls. When a layer is added to MapBean, it automatically becomes a ProjectionListener of MapBean. This means that the layer receives notification when the Projection (or view) of the map changes. To reiterate, the standard capabilities of a Layer are:

- Create, manage, and render graphical objects.
- Receive and respond to mouse events over the map.
- Create GUI controls for additional configuration.

Using these functionalities we have customized the TAGIVE tool so the design and implement a simple map viewer.

V. MAP VIEWER IMPLEMENTATION

In this section, we present an example of the realization of a map viewer by the proposed approach. For short, we omit the screenshots of the tool, a demo of this is visible at link http://lab8.dmi.unisa.it/video/

The first stage of the development process is the modelling of the application map in terms of a graph. In the map the nodes labelled "Scene1", "Scene2", "Scene3" represent frames of the application opened to show respectively the main panel which contains the GIS map, the legend of the available layers and the frame to perform buffer functionalities. The "Database" icon represents the opening of a table which contains the textual data related to the pictorial information on the GIS map. The labels on the edges identify the actions by which to perform an interaction. In particular, the loop transition labelled "Map Action" highlight a possible event on the GIS map as a zoom in, zoom out or panning visualization. The "GeoDataAction" outgoing from a GIS frame indicates an event that presents the textual information related to the points on the GIS map. The "Layers Action" leads to the opening of a new scene related to the legenda of the layers that the end user can to visualize on the GIS map; similarly, the "Spatial Analysis Action" leads at the opening of a new scene related to the buffering functionalities. The "Single_Left_Click" outgoing from the "Scene2" and from the "Scene3" indicate that on these frames will be there dynamic components related to an event witch modifies the GIS map visualization on the "Scene1". At the lower level of the development, the Scene Editor is used to visual implement each scene and the several events related to the foreseen actions in the application map. It is a main frame of the application containing the application GIS map. As a matter of fact, the designer selects the GISFrame widget from the Containers palette and then, he/she details the several layout features by the related Property Handler module and he/she inserts the desired GIS Components from the corresponding palette (i.e., MapButton, GeoDataButton, LayersButton, SpAnalysisButton). The designer defines the "Scene2", related to the frame which will contain the layers legend and the eventually buttons by witch the end user can to manage their visualization (i.e., "Add New Layer", "Remove Layer", etc…). By the BufferFrame widget from the Containers palette and selecting the Elementary Components, the designer defines the visual forms which will guide the end user in the management of the buffer operations. For each scene the Event Handler module allows to visual describe the events related to the inserted dynamic components.

VI. CONCLUSION AND FURTHER WORK

We have presented a methodology to modelling and implement GIS applications. The approach integrates the advantages of diagram design with a visual construction of software applications using a component assembly mechanism. In particular, starting from a top-level design of the visual environment, a usable form-based interface of the tool allows detailing the graphic aspects of each scene. A similar visual technique allows the designer to manage the interactive mechanisms, so that all the features related to an event (i.e., the source object, the action performed on it, and its effect on the application itself) can be easily specified. We have exploited such technique enhancing the visual approach to develop the graphic and interactive features of an application interface presented in TAGIVE (Tool for the Aided Generation of Interactive Visual Environments) [1]. The top-down development approach prevents possible incorrectness and incompleteness errors. In particular, the first level of the development provides an overview of a complex application allowing a structural analysis and a rapid prototyping.

The modularity of the realized environments guarantees reusability of software components. The visual development model for both static and dynamic aspects makes design and implementation more intuitive.

We are currently implementing the system ability to visually define other spatial analysis tasks and we are testing the system to develop GIS applications related to several domains, as well as geology, natural resource management, archaeology.

Further works concerns the improvement of the tool to design Web-GIS interfaces and to implement a SVG markup language supported by several communication devices.

REFERENCES

[1] Cassino, R, Tortora, G, Tucci, M. and Vitiello G. (2006): A Methodology for Computer Supported Development of Interactive Visual Applications - WSEAS Transactions On Information Science and Applications Journal.

[2] M. N. DeMers - Fundamentals of Geographic Information Systems – Wiley.

[3] P. Eades, Q.-W. Feng, "Multilevel visualization of clustered graphs", in: S. North (Ed.), Proc. 4th Intemat. Symp. on Graph Drawing (GD'96), Lecture Notes in Computer Science, Vol. 1190, Springer, Berlin, 1997, pp. 101-112.

[4] Herman, Melanon, G., and Marshall, M. S., "Graph visualization and navigation in information visualization: A survey", IEEE Transactions on Visualization and Computer Graphics, Vol. 6 (1), IEEE 2000, pp. 24--43.

[5] http://www.abacogroup.com/ita/dbm/index.htm

[6] Intergraph, Geomedia, http://www.intergraph.com/GIS/geomedia, 2002.

[7] Oracle, Spatial, http://otn.oracle.com/docs/products/spatial/content.html , 2001

[8] OpenMap Architecture http://openmap.bbn.com/doc/ openmap-arch.html

[9] Postgresql, Documentation, http://www.postgresql.org, 2001

[10] http://www.colorado.edu/geography/gcraft/notes/

[11] http://www.dbase.com/

[12] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[13] http://www.mediadigitali.polimi.it/ddd/ddd_004/art_pdf/tonelli.pdf

[14] http://www.postGIS.org/

[15] http://www.innova.campania.it/servizi

[16] http://www.labgis.net/sapere/gis.shtml

Visual Spatio-Temporal Programming with a 3D Region Connection Calculus

Julia Albath[1], Jennifer L. Leopold[2], Chaman L. Sabharwal[3]
Department of Computer Science
Missouri University of Science and Technology
Rolla, MO, USA
[1]jgadkc@mst.edu, [2]leopoldj@mst.edu, [3]chaman@mst.edu

***Abstract*—The ability to perform qualitative spatial reasoning over a collection of 3D objects would be useful for a variety of problem domains, including biomedical analyses, geographic information systems, and mechanical modeling. In addition to the logical consistency checking that would be required for such a system, clearly there would be a need for a graphical interface that would allow the user to view and manipulate the objects and the corresponding spatial relationships. Additionally, consideration should be given to representing how those spatial relations can change over time. Herein we present a visual programming environment that facilitates spatio-temporal representation of and reasoning over a collection of 3D objects. This system effectively allows the user to create visual "programs", utilizing a region connection calculus to identify and enforce the spatial constraints that logically must hold between the objects over a series of abstract time periods.**

Keywords-qualitative spatial reasoning; constraint logic programming; region connection calculus

I. INTRODUCTION

Spatial reasoning is an important human skill that is employed in many tasks such as biomedical diagnostics, and architectural and mechanical design. In such applications, a collection of 3D objects may be available for the analysis, possibly organized as a time series, wherein the spatial changes that occur from one time period to the next are of particular interest.

Automated spatial reasoning systems typically use logical consistency checking to detect inconsistencies and enforce constraints. Given the graphical nature of spatial data, clearly a visual user interface to such a reasoning system would be beneficial. To a certain degree, CAD/CAM applications have addressed this need. However, the reasoning capabilities of such systems are very limited. The constraints that are enforced are often domain-specific and there is no underlying support for knowledge discovery.

Herein we introduce a visual programming environment, VRCC-3D, which allows the user to create visual "programs" that represent a collection of 3D objects and the (generalized) spatial relations that hold between the objects. A region connection calculus, RCC-3D, is utilized to identify and enforce spatial constraints that logically must hold between the objects over a series of successive time periods.

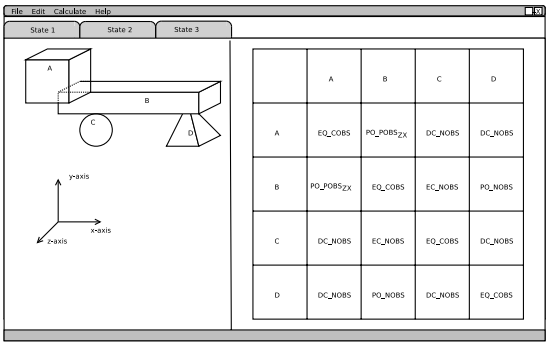


Figure 1. VRCC-3D Interface Example

As shown in Fig. 1, the VRCC-3D user interface has two panes: on the left hand side is a graphical representation of a collection of 3D objects, and on the right hand side is a textual representation of the RCC-3D relations that hold between the various objects. Any changes made to the textual representation will be reflected in the graphical representation, and vice versa. The initial display can be created in two different ways. A file containing the 3D object data can be loaded, and the RCC-3D spatial relations determined [1] and displayed textually. Alternately, a file specifying RCC-3D relations can be loaded, and the system will display a graphical representation of a corresponding collection of objects, produced by prompting the user with a series of questions about each object (e.g., to select from a standard set of shapes like cubes and spheres, to choose the size and orientation, etc.).

Both of these "scene initialization" modes further allow the user to specify an abstract time period (e.g., t_1, t_2, t_3, …) to be associated with a particular set of objects and spatial relations. These different "states" then can be analyzed to determine if there exist RCC-3D relations that could hold between the objects at successive time periods, effectively supporting (or refuting) a spatial transformation of the objects from one state to another state.

In the next section, we provide a very brief overview of related work in qualitative spatial reasoning, and visual constraint logic and spatio-temporal programming systems.

We then discuss how VRCC-3D utilizes a conceptual neighborhood and a composition table to facilitate consistency checking and spatial transformation determination based on RCC-3D.

II. Related Work

A. Constraint Logic Programming

Logic programs which have constraints in the clauses are known as constraint logic programs (CLP). Constraints (specified statically or generated dynamically) are tested in order to control execution of the program. A survey introduction to CLP is provided by Jaffer and Maher [10].

VRCC-3D employs a similar approach, whereby each program "state" represents a set of constraints on the spatial arrangement of the objects at an abstract time period. A number of visual logic programming systems ([11], [12], and [16]) and visual constraint programming systems ([2] and [15]) have been proposed. However, no system has previously been developed to support visual constraint logic programming based on a formal region connection calculus.

B. Spatio-Temporal Programming

In general, spatio-temporal programming facilitates the specification of spatial relationships that must hold at particular time periods. Related work includes spatio-temporal reasoning [6], spatio-temporal constraint logic programming [13], and spatio-temporal simulation [7]. VRCC-3D would best be categorized as spatio-temporal visual constraint logic programming. However, as will be discussed later in this paper, VRCC-3D surpasses the capabilities of many spatio-temporal programming languages; namely, it allows the user to not only specify, but also to discover, spatio-temporal knowledge.

C. Qualitative Spatial Reasoning

Qualitative Spatial Reasoning (QSR) concerns the issues related to automating human spatial analytical processes [8]. A region connection calculus was introduced in [14], and has been one of the foundational theories for QSR.

One focus of QSR is the modeling of and reasoning about *changes* in topological relationships. For example, Egenhofer and Al-Taha introduced the concept of topological distance, and used it to create a conceptual neighborhood graph [5], a visual representation of the changes that transform the spatial relation between a pair of objects into another relation. A similar approach is employed in VRCC-3D to effectively answer questions such as "is object B a possible topological transformation of object A."

III. RCC-3D

Before discussing the specific functionality of VRCC-3D, we provide details about the underlying region connection calculus, RCC-3D, and two associated tools (i.e., a composition table and a conceptual neighborhood) that can be used to solve spatial reasoning tasks.

A. RCC-3D Overview

The fundamental axioms of RCC-3D are *Parthood* and *Connectivity*. For two distinct objects A and B, we use $A \cap B$ to represent everything common to both A and B; $A \sim B$ is used to denote everything that is in A, but not in B. The boundary of A is represented as ∂A. The closure of A is denoted by $\bar{A}$; it encompasses everything in A and ∂A. A° denotes the interior of A, which is everything that is in the closure of A, but does not include the boundary of A. A^c is the complement of A; that is, A^c is everything that is not in A. A^e (or $\bar{A}^c$) is the exterior of A; it represents everything that is not in $\bar{A}$. The interior, boundary, and exterior of an object are disjoint, and their union is the universe. $P(A,B)$ is defined as A is part of B if A is a subset of B; specifically, A is part of B if $A \cap B^c = \varnothing$. A is connected to B, denoted by $C(A,B)$, if $\bar{A} \cap \bar{B} \neq \phi$. Additionally, we define a predicate *Proper Part (PP)* as follows:

$$PP(A,B) \equiv P(A,B) \wedge \neg P(B,A) \quad (1)$$

The eight base relations for RCC-3D (the entire set of relations for RCC-8 [17]) can be defined in terms of P, C, and PP as listed below.

$$PO(A,B) \equiv \exists(C \neq \varnothing)[P(C,A^\circ) \wedge P(C,B^\circ)] \wedge \neg P(A,B) \wedge \neg P(B,A) \quad (2)$$

$$EQ(A,B) \equiv A = B \quad (3)$$

$$EC(A,B) \equiv C(A,B) \wedge (A^\circ \cap B^\circ = \varnothing) \quad (4)$$

$$DC(A,B) \equiv \neg C(A,B) \quad (5)$$

$$TPP(A,B) \equiv PP(A,B) \wedge \exists(C \neq \varnothing)[EC(C,A^\circ) \wedge EC(C,B^\circ)] \quad (6)$$

$$NTPP(A,B) \equiv PP(A,B) \wedge \nexists C[EC(C,A) \wedge EC(C,B)] \quad (7)$$

$$TPPc(A,B) \equiv PP(B,A) \wedge \exists(C \neq \varnothing)[EC(C,B^\circ) \wedge EC(C,A^\circ)] \quad (8)$$

$$NTPPc(A,B) \equiv PP(B,A) \wedge \nexists C[EC(C,B) \wedge EC(C,A)] \quad (9)$$

To facilitate automated reasoning, each of the relations must have a converse. For example *TPPc* and *NTPPc* are the converses of *TPP* and *NTPP*, respectively. The converse of a relation is denoted with a *c* at the end of the relation name. However, this convention is not necessary for the symmetric relations *PO(A,B), EQ(A,B), EC(A,B),* and *DC(A,B).*

For some types of spatial reasoning it is necessary to consider the obscuration that can occur when the objects are seen through orthographic projection on any of the principal planes in R^3. When considering that object A obscures object B, it is implied that object A is closer than object B to the perspective reference point. For our discussion, we assume that the direction of the projection (i.e., the line of sight) is orthogonal to the plane of projection, and that the plane of

projection is one of the principal planes (xy, yz, or zx). The obscuration of objects is determined by projecting the objects on a plane P.

Unlike other region connection calculi such as RCC-8, RCC-3D also differentiates between complete and partial obscuration. Partial or complete obscuration is indicated by the inclusion of $POBS_P$ or $COBS_P$, respectively, in the relation name. For example, $EC_POBS_P(A,B)$ is true if objects A and B are externally connected and object A partially obscures object B; $EC_COBS_P(A,B)$ is true if objects A and B externally connect, and object A completely obscures object B. Again, the converse is denoted by including a *C* in the relation name (e.g., $EC_CCOBS_P(A,B)$ is the converse of $EC_COBS_P(A,B)$, and $EC_CPOBS_P(A,B)$ is the converse of $EC_POBS_P(A,B)$). The case of no obscuration is represented by including *NOBS* in the relation name; for example, when A and B externally connect, but there is no obscuration, $EC\text{-}NOBS_P(A,B)$ is true. The complete set of RCC-3D relations (considering the extent of obscuration) are given in equations (10)-(28); A_P and B_P represent the respective projections of A and B on the plane P, where P is xy, yz, or zx. Note that *PO* (previously defined in equation (2)) now has been augmented with PO_POBS_P, PO_COBS_P, and PO_CCOBS_P; *EQ*, *EC*, and *DC* have been augmented similarly. Not every $xOBS_P$ descriptor applies to every relation; for example, for *TPP*, $NOBS_P$ and $POBS_P$ do not apply.

$$DC_NOBS_P(A,B) = DC(A,B) \wedge NOBS(A_P,B_P) \quad (10)$$
$$DC_POBS_P(A,B) = DC(A,B) \wedge POBS(A_P,B_P) \quad (11)$$
$$DC_CPOBS_P(A,B) = DC(A,B) \wedge POBS(B_P,A_P) \quad (12)$$
$$DC_COBS_P(A,B) = DC(A,B) \wedge COBS(A_P,B_P) \quad (13)$$
$$DC_CCOBS_P(A,B) = DC(A,B) \wedge CCOBS(A_P,B_P) \quad (14)$$
$$EC_NOBS_P(A,B) = EC(A,B) \wedge NOBS(A_P,B_P) \quad (15)$$
$$EC_POBS_P(A,B) = EC(A,B) \wedge POBS(A_P,B_P) \quad (16)$$
$$EC_CPOBS_P(A,B) = EC(A,B) \wedge POBS(B_P,A_P) \quad (17)$$
$$EC_COBS_P(A,B) = EC(A,B) \wedge COBS(A_P,B_P) \quad (18)$$
$$EC_CCOBS_P(A,B) = EC(A,B) \wedge CCOBS(A_P,B_P) \quad (19)$$
$$PO_POBS_P(A,B) = PO(A,B) \wedge POBS(A_P,B_P) \quad (20)$$
$$PO_CPOBS_P(A,B) = PO(A,B) \wedge POBS(B_P,A_P) \quad (21)$$
$$PO_COBS_P(A,B) = PO(A,B) \wedge COBS(A_P,B_P) \quad (22)$$
$$PO_CCOBS_P(A,B) = PO(A,B) \wedge CCOBS(A_P,B_P) \quad (23)$$
$$TPP_CCOBS_P(A,B) = TPP(A,B) \wedge CCOBS(A_P,B_P) \quad (24)$$
$$TPPc_COBS_P(A,B) = TPPc(A,B) \wedge COBS(A_P,B_P) \quad (25)$$
$$EQ_CCOBS_P(A,B) = A = B \quad (26)$$
$$NTPP_CCOBS_P(A,B) = NTPP(A,B) \wedge CCOBS(A_P,B_P) \quad (27)$$
$$NTPPc_COBS_P(A,B) = NTPPc(A,B) \wedge COBS(A_P,B_P) \quad (28)$$

Obscuration can be described using a 4-Intersection that considers the intersections of object A_P's interior and exterior with object B_P's interior and exterior; it should be noted that the boundaries are not relevant in this context. A 2×2 matrix, *M*, is used to represent these intersections as shown in Table I.

TABLE I. MATRIX M

$A_P^{\circ} \cap B_P^{\circ}$	$A_P^{\circ} \cap B_P^{e}$
$A_P^{e} \cap B_P^{\circ}$	$A_P^{e} \cap B_P^{e}$

There are 16 combinations for M using the values of empty and not-empty. Of those 16 combinations, only four are possible considering reality constraints; for example, unless one of the objects is the universe (in which case it has no exterior), the intersection of the exteriors cannot be empty. So only 3-Intersections actually are needed for bounded objects. The four relations of the 3-Intersection are as follows, each specified as the triple of intersections ($A_P^{\circ} \cap B_P^{\circ}$, $A_P^{\circ} \cap B_P^{e}$, $A_P^{e} \cap B_P^{\circ}$): $NOBS_P$=(∅, ¬∅, ¬∅), $POBS_P$=(¬∅, ¬∅, ¬∅), $COBS_P$=(¬∅, ∅, ¬∅), and $CCOBS_P$=(¬∅, ¬∅, ∅).

B. The Compositon Table

A composition table is used to answer questions such as, "given three objects, A, B, and C, and knowledge of relations R_1(A,B), and R_2(B,C), what can be said about the relation between A and C" [4]. The composition operator is an important function for any relation algebra [17]; by creating a composition table, calculating the composition of two relations becomes a simple table look-up.

In [4], the author shows a method to generate all the relations from the composition of two existing topological relations. Key to this approach is the 9-Intersection model [3], which considers each of the nine intersections of object A's interior, boundary, and exterior with object B's interior, boundary, and exterior. A similar approach was used to generate the composition table for the RCC-3D relations (for brevity, that table is not included in this paper).

C. The Conceptual Neighborhood Graph

The conceptual neighborhood graph was first introduced in [9] to find those transitions that can occur when we gradually change the geometry of one object in a pair. The authors of [5] introduced the idea of a topological distance, building upon the 9-Intersection calculus. The topological distance is the number of intersections that have changed from empty to not empty (or vice versa), from one relation to another. Consider the 9-Intersection matrices for *DC* and *TPP*, as shown in Tables II and III, respectively. In this example, the topological distance between *DC* and *TPP* is five because the values for five intersections have changed; namely, the values have changed for ($A^{o} \cap B^{o}$), ($A^{\circ} \cap B^{e}$), ($\partial A \cap B^{\circ}$), ($\partial A \cap \partial B$), and ($\partial A \cap B^{e}$).

TABLE II. 9-INTERSECTION FOR DC

	$\mathbf{B^o}$	$\mathbf{\partial B}$	$\mathbf{B^e}$
$\mathbf{A^o}$	∅	∅	¬∅
$\mathbf{\partial A}$	∅	∅	¬∅
$\mathbf{A^e}$	¬∅	¬∅	¬∅

TABLE III. 9-INTERSECTION FOR TPP

	$\mathbf{B^o}$	$\mathbf{\partial B}$	$\mathbf{B^e}$
$\mathbf{A^o}$	¬∅	∅	∅
$\mathbf{\partial A}$	¬∅	¬∅	∅
$\mathbf{A^e}$	¬∅	¬∅	¬∅

As discussed previously, here also, only 8 out of 9-Intersections actually are needed for bounded objects. The conceptual neighborhood graph is created by considering the closest topological distance (as calculated for each pair of relations), as well as possible deformations of scaling, translation, and rotation. The neighborhood of an object is an equivalence class of all objects that are topologically transformable into each other; that is, it is the equivalence class of all objects that are topologically transformable into it with the least number of transformations.

Expanding on this approach for RCC-3D, we combined the 8-Intersection for the RCC-8 base relations with a 3-Intersection for the projections onto the specified plane. The 3-Intersections are applied to 2D projections of 3D objects to represent obscurations of the objects. The results of applying these intersections to the RCC-3D relations (in terms of the 8- and 3-Intersections, where applicable) are shown in Table IV.

The 8-Intersection can be used to determine *inter-relation* topological distance, and the 3-Intersection can be used for the determination of the *intra-relation* topological distance. Here intra-relation means *within* the same (base) relation (e.g., DC_POBS_P to DC_COBS_P, or from EC_COBS_P to EC_POBS_P), whereas inter-relation means *between* (base) relations (e.g., from EC_POBS_P to DC_POBS_P, or PO_COBS_P to $TPPc_NOBS_P$). It is important to distinguish between these two types of distances. A total topological distance of (8-Intersection + 3-Intersection = Total) 0+2=2 is different from a distance of 2+0=2. This also is significant when we consider that a change in the 8-Intersection often implies a change in the 3-Intersection.

The computed topological distances for the RCC-3D relations are shown in Table V, where the first number is the topological distance in the 8-Intersection and, where applicable, the second number (y in the expression $x+y$) is the topological distance in the 3-Intersection.

The entries in Table V are important for building the RCC-3D conceptual neighborhood graph. The inter-relation distance between two base relations is the number of entries that differ between their 8-Intersection representations. For determining obscuration, objects are projected on a plane and the additional 3-Intersection representation on the projections is considered. The intra-relation distance between two relations is then the number of entries that differ between their 3-Intersection representations. In general, the distance between a relation and another relation R is represented by $x_{inter} + y_{intra}$. If $x_{inter} = 0$ and $y_{intra} = 0$, then the neighbor is the relation R itself. If $x_{inter} = 0$ and $y_{intra} \neq 0$, then the neighbor is from the same base relation (e.g., DC_NOBS_P and DC_COBS_P are neighbors, as are PO_COBS_P, and PO_POBS_P). If $x_{inter} \neq 0$, then the neighbors are from different base relations (e.g., PO_POBS_P and EC_COBS_P are neighbors). The complete conceptual neighborhood graph for RCC-3D is depicted in Fig. 2 as a minimum spanning tree.

IV. VRCC-3D

The composition table and the conceptual neighborhood play an important role in the VRCC-3D system's consistency checking and determination of spatial transformations across abstract time periods. This functionality can be illustrated with two possible use cases.

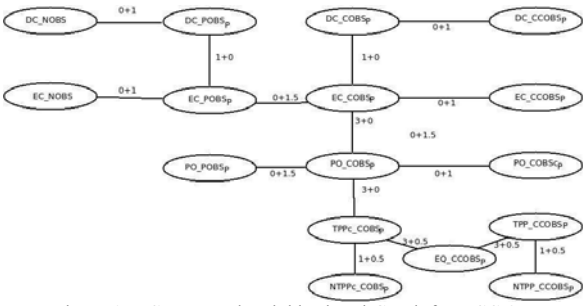


Figure 2. Conceptual Neighborhood Graph for RCC-3D

A. An Example of Inconsistency Checking

When changes are made to the graphical representation of the collection of objects (in the left hand pane of the VRCC-3D interface), the system must update the corresponding textual display of RCC-3D relations in the right hand pane. Because the change was made directly to the graphical representation, we know that the resulting spatial arrangement of the objects (and associated RCC-3D relations) will be consistent. Updates to the matrix of RCC-3D relations will take place in both the entire column and row associated with the object that has been altered. Specifically, the All-Pairs-Relation-Detection Algorithm [1] is applied to recalculate the affected relations; this algorithm uses the composition table to narrow down the possible relations that *could* hold, in the process of determining the single relation that *does* hold between a pair of objects.

A simple example of this use case is illustrated in Fig. 3, wherein the user has altered the scene depicted earlier in Fig. 1 by moving object A. Clearly this requires recalculation of the RCC-3D relations in row A and column A (which are highlighted in the RCC-3D relation matrix that is also displayed in Fig. 3). The relation for (A,B) changes from $PO_POBS_P(A,B)$ to $DC_POBS_P(A,B)$. The relations for *(A,C)* and *(A,D)* do not change in this particular case, but still are recalculated; in future work, we will investigate ways to detect and avoid such unnecessary computational efforts.

Suppose that the user instead directly updates entries in the RCC-3D relation matrix (in the right hand pane of the VRCC-3D interface). All object pairs which have been changed are then added to an "updated list." We use the Dynamic Path Consistency Algorithm [1], which in turn uses the composition table, to check whether those changes have introduced any inconsistencies in the associated spatial relations. If any inconsistencies are detected, the VRCC-3D system displays an error message, and gives the user an opportunity to make corrections. In no inconsistencies are found, the graphical display needs to be updated; an efficient

implementation of this particular functionality is currently under development.

B. An Example of Constraint Determination over Time

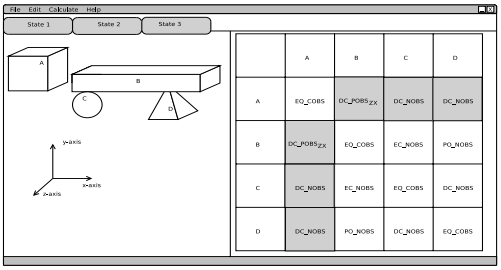


Figure 3. VRCC-3D System After Moving Object A

If the user has created two or more states that represent arrangements of the objects at different abstract time periods, the VRCC-3D system can be asked to determine if state t_i is one deformation removed from state t_j where $i > j$. This is accomplished by finding all object pairs that are one deformation from their current relation. We use the conceptual neighborhood graph to determine if there is a path of length one in the neighborhood graph from the relation at t_i to the relation at t_j. After finding all object pairs with such a path, we determine if that change creates a spatial arrangement that is equivalent to the arrangement at time t_j. This can easily be done by making that change, and using the All-Pairs-Relation-Detection Algorithm to recalculate the relations that could have changed. The Dynamic-Path-Consistency Algorithm can then be used to find inconsistencies. After the Dynamic-Path-Consistency Algorithm is completed, we compare the resulting arrangement to state t_j; if they match, state t_j is a deformation of state t_i. For example, suppose that the configuration of objects shown in Fig. 1 has been designated as state 1, and the arrangement shown in Fig. 3 represents state 2. Since state 2 is determined to be a valid deformation of state 1, a notification message to that effect is displayed.

If the two arrangements (resulting from application of the Dynamic-Path-Consistency Algorithm) are not the same, the next pair is chosen and the same process is applied. If no pair of relations is found that results in a match with t_j, the system will ask the user to create another state t_x, where $t_i < t_x < t_j$, or the system can be asked to generate a new possible intermediate state t_x. The user may pick from the system-generated possible states that are created by interpolating the 8-Intersection and 3-Intersection of the configurations at times t_i and t_j. This amounts to discovering new states (and hence knowledge) interactively.

V. Future Work

As previously stated, our future work will focus on efficiently updating the graphical representation of the collection of objects when a change has been made to the textual display of spatial relations. Another area of improvement will be in detecting and avoiding recalculation of relations that ultimately will not change. We hope to accomplish the latter by keeping track of the old and new locations of objects, and using quantitative methods to efficiently determine if the move necessitates a recalculation of the spatial relations; similarly, the effects of resizing of objects will be considered.

VI. Summary

In this paper we introduced VRCC-3D, a system which utilizes a region connection calculus to check the consistency of spatial relations between 3D objects. By providing a visual interface to this system, the user is able to make changes to both the graphical representation of a collection of objects as well as the textual representation of the associated spatial relations. Because the spatial changes that hold from one time period to the next may be of particular interest, we briefly discussed the roles of a composition table and a conceptual neighborhood graph to check the consistency between "states", as well as to suggest what changes might have occurred in intermediate states.

With the explosion in he availability of 3D image data, practical applications for this work lie in many areas, including the studies of organism development and evolution. To examine (or even hypothesize) changes that occur in a species from one development stage to another, or from the anatomy of one species to the anatomy of another potentially related species over time, VRCC-3D can be used to determine and display the topological relations that hold in associated collections of images. The user can then ask the system to verify the consistency of the physical transformations between the images; if inconsistencies are found, the system can suggest possible transformation scenarios. This could aid in the identification of previously undetermined development patterns within a single species, or could lead to the description of other species that had to have existed as an "intermediate step" in the evolution of the species under consideration. Given that such images are comprised of hundreds of objects, with thousands of data points, automation of this analysis, combined with a visual display that associates the formal topological descriptions with the images, hopefully will facilitate such research advances.

Acknowledgments

This work was supported by NSF under award DBI-0640053.

References

[1] J. Albath, J. Leopold, C. Sabharwal, K. Perry, Efficient 3D Qualitative Spatial Reasoning with RCC-3D, *Proceedings of KSEM 2010*, pp. 470-481, 2010.

[2] R. Bartak, Constraint Programming: In Pursuit of the Holy Grail, *Proceedings of WDS99 (invited lecture),* 1999.

[3] M. Egenhofer. Reasoning about Binary Topological Relations. In *Advances in Spatial Databases*, pages 141–160. Springer, 1991.

[4] M.J. Egenhofer. Deriving the Composition of Binary Topological Relations, *Journal of Visual Languages and Computing*, 5(2): 133-149, 1994.

[5] M. Egenhofer and K. Al-Taha. Reasoning about Gradual Changes of Topological Relationships. *Theories and methods of spatio-temporal reasoning in geographic space*, pp. 196–219, 1992.

[6] M.J. Egenhofer and R.G. Golledge, Spatial and Temporal Reasoning in Georgraphic Information Systems, Oxford University Press, USA, 1998.

[7] M. Erwig, R.H. Guting, M. Schnieder and M. Vazirgiannis, Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases, *GeoInformatica*, Springer, 3(3):269-296, 1999.

[8] M.T. Escrig and F. Toledo, Qualitative Spatial Reasoning: Theory and Practice: Applications to Robot Navigation, ISO Press Amsterdam, The Netherlands, 1998.

[9] C. Freksa. Temporal Reasoning Based on Semi-Intervals. *Artificial intelligence*, 54(1-2):199–227, 1992.

[10] J. Jaffar and M.J. Maher. Constraint Logic Programming: A survey. *The Journal of Logic Programming*, 19:503-581, 1994.

[11] D. Ladret and M Rueher. VLP: A Visual Logic Programming Language. *Journal of Visual Languages & Computing*, 2(2):163-188, 1991.

[12] J. Puigsegur, J. Agusti and D. Robertson. A Visual Logic Programming Language. *IEEE Symposium on Visual Languages*, pp. 214-221, 1996.

[13] A. Raffaeta and T. Fruhwirth, Spatio-Temporal Annotated Constraint Logic Programming, *Practical Aspects of Declarative Languages,* Springer, pp. 259-273, 2001.

[14] D.A. Randell, Z. Cui, and A.G. Cohn. A Spatial Logic Based on Regions and Connection. *KR*, 92:165–176, 1992.

[15] C. Schulte, OZ Explorer: A Visual Constraint Programming Tool, *Programming Languages: Implementations, Logics, and Programs*, Spring, pp. 477-478, 1996.

[16] L. Spratt and A. Ambler, A Visual Logic Programming Language Based on Sets, *Proceedings of the 9th IEEE symposium on Visual Languages,* 1993.

[17] A. Tarski, On the Calculus of Relations. *The Journal of Symbolic Logic*, 6:73-89, 1941.

TABLE IV. THE 8+ 3-INTERSECTIONS FOR RCC-3D

	$A^o \cap B^o$	$A^o \cap \partial B$	$A^o \cap B^e$	$\partial A \cap B^o$	$\partial A \cap \partial B$	$\partial A \cap B^e$	$A^e \cap B^o$	$A^e \cap \partial B$	$A^o_P \cap B^o_P$	$A^o_P \cap B^e_P$	$A^e_P \cap B^o_P$
DC_NOBS	∅	∅	¬∅	∅	∅	¬∅	¬∅	¬∅	∅	¬∅	¬∅
DC_POBS$_P$	∅	∅	¬∅	∅	∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅
DC_COBS$_P$	∅	∅	¬∅	∅	∅	¬∅	¬∅	¬∅	¬∅	¬∅	∅
DC_CCOBS$_P$	∅	∅	¬∅	∅	∅	¬∅	¬∅	¬∅	¬∅	∅	¬∅
EC_NOBS	∅	∅	¬∅	∅	¬∅	¬∅	¬∅	¬∅	∅	¬∅	¬∅
EC_POBS$_P$	∅	∅	¬∅	∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅
EC_COBS$_P$	∅	∅	¬∅	∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	∅
EC_CCOBS$_P$	∅	∅	¬∅	∅	¬∅	¬∅	¬∅	¬∅	¬∅	∅	¬∅
PO_POBS$_P$	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅
PO_COBS$_P$	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	∅
PO_CCOBS$_P$	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	¬∅	∅	¬∅
TPP_CCOBS$_P$	¬∅	∅	∅	¬∅	¬∅	∅	¬∅	¬∅	¬∅	∅	¬∅
TPPc_COBS$_P$	¬∅	¬∅	¬∅	∅	¬∅	¬∅	∅	∅	¬∅	¬∅	∅
EQ_COBS$_P$	¬∅	∅	∅	∅	¬∅	∅	∅	∅	¬∅	∅	∅
NTPP_CCOBS$_P$	¬∅	∅	∅	¬∅	∅	∅	¬∅	¬∅	¬∅	∅	¬∅
NTPPc_COBS$_P$	¬∅	¬∅	¬∅	∅	∅	¬∅	∅	∅	¬∅	¬∅	∅

TABLE V. THE TOPOLOGICAL DISTANCES FOR THE RCC-3D CALCULUS

	DC_ NOBS	**DC_ POBS$_P$**	**DC_ COBS$_P$**	**DC_ CCOBS$_P$**	**EC_ NOBS**	**EC_ POBS$_P$**	**EC_ COBS$_P$**	**EC_ CCOBS$_P$**	**PO_ POBS$_P$**	**PO_ COBS$_P$**	**PO_ CCOBS$_P$**	**TPP_ CCOBS$_P$**	**TPPc_ COBS$_P$**	**EQ_ COBS$_P$**	**NTPP_ CCOBS$_P$**	**NTPPc_ COBS$_P$**
DC_NOBS	0	0+1	0+2.5	0+2.5	1+0	1+1	1+2.5	1+2.5	4+1	4+2.5	4+2.5	5+2.5	5+2.5	6+3	4+2	4+2
DC_POBS$_p$	0+1	0	0+1.5	0+1.5	1+1	1+0	1+1.5	1+1.5	4+0	4+1.5	4+1.5	5+1.5	5+1.5	6+2	4+1	4+1
DC_COBS$_p$	0+2.5	0+1.5	0	0+1	1+2.5	1+2.5	1+0	1+1	4+1.5	4+0	4+1	5+1	5+0	6+0.5	4+1.5	4+0.5
DC_CCOBS$_p$	0+2.5	0+1.5	0+1	0	1+2.5	1+1.5	1+1	1+0	4+1.5	4+1	4+0	5+0	5+1	6+0.5	4+0.5	4+1.5
EC_NOBS	1+0	1+1	1+2.5	1+2.5	0	0+1	0+2.5	0+2.5	3+1	3+2.5	3+2.5	4+2.5	4+2.5	5+3	5+2	5+2
EC_POBS$_P$	1+1	1+0	1+2.5	1+1.5	0+1	0	0+1.5	0+1.5	3+0	3+1.5	3+1.5	4+1.5	4+1.5	5+2	5+1	5+1
EC_COBS$_P$	1+2.5	1+1.5	1+0	1+1	0+2.5	0+1.5	0	0+1	3+1.5	3+0	3+1	4+1	4+0	5+0.5	5+1.5	5+0.5
EC_CCOBS$_P$	1+2.5	1+1.5	1+1	1+0	0+2.5	0+1.5	0+1	0	3+1.5	3+1	3+0	4+0	4+1	5+0.5	5+0.5	5+1.5
PO_POBS$_P$	4+1	4+0	4+1.5	4+1.5	3+1	3+0	3+1.5	3+1.5	0	0+1.5	0+1.5	3+1.5	3+1.5	6+2	4+1	4+1
PO_COBS$_P$	4+2.5	4+1.5	4+0	4+1	3+2.5	3+1.5	3+0	3+1	0+1.5	0	0+1	3+1	3+0	6+0.5	4+1.5	4+0.5
PO_CCOBS$_P$	4+2.5	4+1.5	4+1	4+0	3+2.5	3+1.5	3+1	3+0	0+1.5	0+1	0	3+0	3+1	6+0.5	4+0.5	4+1.5
TPP_CCOBS$_P$	5+2.5	5+1.5	5+1	5+0	4+2.5	4+1.5	4+1	4+0	3+1.5	3+1	3+0	0	6+1	3+0.5	1+0.5	7+1.5
TPPc_COBS$_P$	5+2.5	5+1.5	5+0	5+1	4+2.5	4+1.5	4+0	4+1	3+1.5	3+0	3+1	6+1	0	3+0.5	7+1.5	1+0.5
EQ_COBS$_P$	6+3	6+2	6+0.5	6+0.5	5+3	5+2	5+0.5	5+0.5	6+2	6+0.5	6+0.5	3+0.5	3+0.5	0	4+1	4+1
NTPP_CCOBS$_P$	4+2	4+1	4+1.5	4+0.5	5+2	5+1	5+1.5	5+0.5	4+1	4+1.5	4+0.5	1+0.5	7+1.5	4+1	0	6+2
NTPPc_COBS$_P$	4+2	4+1	4+0.5	4+1.5	5+2	5+1	5+0.5	5+1.5	4+1	4+0.5	4+1.5	7+1.5	1+0.5	4+1	6+2	0

On the Relative Expressiveness of Second-Order Spider Diagrams and Regular Expressions

Peter Chapman and Gem Stapleton
Visual Modelling Group, University of Brighton, UK
{p.b.chapman, g.e.stapleton}@brighton.ac.uk

Abstract

This paper is about spider diagrams, an extension of Euler diagrams that includes syntax to make assertions about set cardinalities. Like many diagrammatic logics, spider diagrams are a monadic and first-order, so they are inexpressive. The limitation to first-order precludes the formalisation of many fundamental concepts such as the cardinality of a set being even. To this end, second-order spider diagrams have recently been proposed. In this paper, we establish their expressiveness relative to that of regular expressions, advancing novel research connecting diagrammatic logics with formal languages. In particular, we establish that every regular language is definable by a second-order spider diagram, demonstrating a substantial increase in expressiveness over spider diagrams.

1. Introduction

Diagrams are widely used in application areas such as software engineering and information visualization, since they can be more readily accessible to users than their symbolic counterparts. Examples of diagrammatic logics include Euler diagrams. Hammer developed a simple sound and complete Euler diagram logic which had just three inference rules [7]. An example can be seen in figure 1 where, from the premise diagram d_1 we deduce d_2 by erasing a curve. The diagram d_1 tells us that B is disjoint from both A and C; erasing B loses this information.

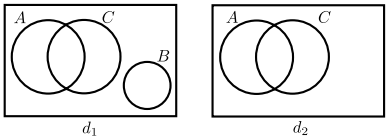


Figure 1. Euler diagram reasoning.

A logic based on Venn diagrams was formalised by Shin [11], building on work by Peirce [9]. Shin's work is regarded as seminal, since this was the first serious attempt at producing a sound and complete diagrammatic logic. Substantial advances have now been made in terms of our understanding of how to formalise the syntax and semantics diagrammatic logics, what constitutes an effective diagram, and how to develop diagrammatic inference rules. Recent examples include Swoboda and Allwein's Euler/Venn logic, which augments Euler diagrams with named individuals to represent particular elements [12]. For example, in figure 2, the Euler/Venn diagram d_3 expresses that $John$ is a $Lecturer$ and a $Person$. In fact, the diagram expresses that all lecturers are people.

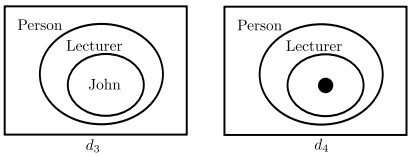


Figure 2. Extending Euler diagrams.

Spider diagrams [6] use trees, called spiders, to represent the existence of elements. To illustrate, the spider diagram d_4 in figure 2 expresses that there is a lecturer and that all lecturers are people. Recently, spider diagrams have been extended to spider diagrams of order allowing properties of an ordered universe to be specified [2]. The motivation for this extension derived from the realisation that spider diagrams could define some regular languages, but could not specify ordering constraints on the letters in those languages; we give examples later in the paper demonstrating how diagrams define languages. Spider diagrams of order can define precisely the star-free regular languages [4] and spider diagrams can define precisely the commutative star-free regular languages [3]; a language is star-free if it can be defined without the use of the Kleene star, such as $a|b$ but not $(aa)^*$, and is commutative if it is closed under permutation. The primary result of this paper is that second-order spider diagrams [1], can define all regular languages.

Section 2 presents the syntax and semantics of second-order spider diagrams and section 3 provides a formal definition of the language of such a diagram. Moreover, we informally describe a translation process that converts a reg-

ular expression, r, into a second-order spider diagram, d, such that r and d define the same language. This conversion shows that every regular language is definable by a second-order spider diagram. Consequently, we have established a substantial increase in expressiveness over spider diagrams of order, which can only define the star-free regular languages. We conclude in section 4.

2. Second-Order Spider Diagrams

Spider diagrams use closed curves to represent sets and assert relationships between those sets. For instance, the enclosure of one curve by another curve corresponds to a subset/superset relationship between the represented sets. In addition, as shown in d_4 of figure 2, they use existential spiders to assert the existence of elements; the existential spiders place lower bounds on set cardinalities. Shading places upper bounds on set cardinality: in a shaded region, all elements are represented by spiders. The diagrams described so far are called *unitary* spider diagrams. Spider diagrams can be negated and joined with binary connectives: $\wedge$ to denote 'and', $\vee$ to denote 'or', $\Rightarrow$ to denote 'implies' and, in the case of spider diagrams of order, joined with $\lhd$ (read as 'product') to specify ordering information. The spider diagram of order $d_1 \lhd d_2$ in figure 3 expresses that there is an element in A ordered before an element in B.

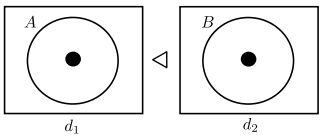


Figure 3. Compound diagrams.

In [1], spider diagrams of order were extended by incorporating additional syntax to make second-order statements. Perhaps the simplest form of second-order quantification is over subsets of the universal set. Extending spider diagrams to incorporate such quantification results in a monadic second-order logic with an order relation. Such an extension is achieved by using unlabelled curves to assert the existence of a set. Negating diagrams, therefore, brings in universal quantification over subsets. Here, we note that the free-rides [10] that occur in the underlying Euler diagram notation also occur when we incorporate unlabelled curves. Since we are concerned with ordering information, second-order spider diagrams also extend the syntax to allow the use of arrows to talk about successors.

In figure 4, d_1 expresses that there are two disjoint, non-empty subsets of A that have sets of successors that are disjoint from A. These sets of successors are also disjoint and, in fact, disjoint sets always have disjoint sets of successors, since successor functions are injective. The diagram d_2 expresses that the universe can be partitioned into two disjoint sets where the successors of one set are all in the other set. That is, the diagram expresses that the universe has even cardinality. Negating this diagram would, therefore, express that the universe has odd cardinality; negation is depicted using a line above the diagram, shown as d_3.

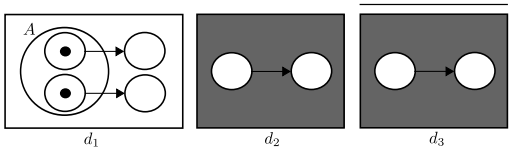


Figure 4. Second-order spider diagrams.

We now provide an overview of the syntax and semantics. We have a finite set of *fixed contours*, $\mathcal{FC}$, and a countably infinite set of *existential contours*, $\mathcal{EC}$. Each contour in $\mathcal{FC}$ corresponds to a labelled curve in a diagram, whereas the contours in $\mathcal{EC}$ correspond to unlabelled curves. A *zone* is a pair of finite, disjoint sets of contours. Given a diagram, d, with a finite set of contours $C \subset \mathcal{FC} \cup \mathcal{EC}$, one defines the *possible zones* of d as pairs of sets of contours (in, out), such that $in \cup out = C$; a diagram's zone set, Z, is a subset of its possible zones. Intuitively, the zone thus defined is inside every contour of *in*, and outside every contour of *out*. The set of all zones is $\mathcal{Z}$. Zones may be *shaded*: the set of shaded zones of a diagram is denoted *Sh*, where $Sh \subseteq Z$. An *existential spider* is an element the set $\mathcal{ES}$. Spiders are placed in a set of zones in a unitary diagram. The set of such zones is called the spider's *habitat*. It is assumed that the sets $\mathcal{FC}, \mathcal{EC}$ and $\mathcal{ES}$ are pairwise disjoint. Properties of the successor function will be denoted by arrows, formally defined as a set of pairs, $(source, target)$ with the set of arrows in a unitary diagram denoted A. The source and target for each arrow in A are drawn from $C \cup ES$. In addition, $\bot$ is a unitary diagram which denotes 'false'. As stated above, we incorporate negation, depicted as a line above the diagram, along with $\wedge$, $\vee$, $\Rightarrow$ and $\lhd$.

At the semantic level, an *interpretation*, I, is a universal set, U, together with an assignment, Ψ, of a subset of U to each fixed contours, and a strict, total order, $<$, on U. We write $I = (U, \Psi, <)$. An *extended interpretation* is an extension of Ψ to interpret existential contours, zones, regions and existential spiders. A zone, (in, out), represents the set

$$\Psi(in, out) = \bigcap_{l \in in} \Psi(l) \cap \bigcap_{l \in in} (U - \Psi(l))$$

where $\Psi(l)$ is the set assigned to contour l. A region represents the set which is the union of the sets represented by the region's constituent zones. Briefly, we say that an interpretation, $I = (U, \Psi, <)$, is a *model* for unitary diagram d if $d \neq \bot$ and there exists an extension of I that ensures

1. the union of the sets represented by d's zones is U,

2. each spider in d maps to a singleton set that is a subset of the set represented by the spider's habitat,
3. no two spiders map to the same set,
4. in a shaded region, r, all of the elements in the represented set, $\Psi(r)$, are in sets mapped to by spiders of d, and
5. for each arrow in d, there exists a bijection, s, between the sets denoted by the arrow's source and target such that s is the successor function induced by $<$.

The semantics extend to compound diagrams in the obvious way with, perhaps, $\lhd$ being the exception. In this case, an interpretation, I, is a model for $d_1 \lhd d_2$ if there exist two interpretations, I_1 that models d_1 and I_2 that models d_2, whose ordered sum is I [5].

3. Connections with Regular Languages

Recall a (not necessarily regular) language over an alphabet Σ is a subset of Σ^* (all the words formed from letters in Σ). A language is *regular* if it can be defined by a regular expression [8]. *Regular expressions* are formed using letters in Σ, λ to denote the empty word, disjunction, denoted $|$, concatenation, denoted $\cdot$, and the Kleene star, denoted *. In addition, we allow the use of the $+$ operator, where $r^+ = r \cdot r^*$.

To define regular languages using logic, one assigns letters to predicate symbols and then uses logical statements, in our case diagrams, to make assertions about the letters [13]. In second-order spider diagrams, the fixed contour labels act as predicate symbols. In d_1 of figure 5, there is one spider in the contour A, and one spider outside the contour A. Suppose we assign the letter a to the contour A, and the letter b to the rest of the diagram. Because there is no shading in the diagram, the number of letters in a word is not restricted. However, the spiders and the successor arrow means any word in the language of d_1 must contain the subword ab. The shading in d_2 means that there is exactly one element in the contour A, so every word contains exactly one a. In terms of regular expressions, d_2 defines $b^*(ab)b^*$ whereas d_1 defines $(a|b)^*(ab)(a|b)^*$.

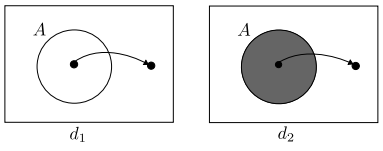


Figure 5. Defining languages using diagrams.

To formalise the language of a diagram, we assume that the sets $\mathcal{FC}$ and Σ are fixed and, moreover, that a fixed function $l\colon \mathcal{FC} \to \mathbb{P}\Sigma$ is given; we call l a *letter assignment*. We extend l to interpret zones, such that for each zone (in, out)

$$l(in, out) = \bigcap_{c \in in \cap \mathcal{FC}} l(c) \bigcap_{c \in out \cap \mathcal{FC}} (\Sigma - l(c)).$$

To illustrate, consider the alphabet $\Sigma = \{a, b\}$, fixed contour set $\mathcal{FC} = \{A\}$ and letter assignment $l(A) = \{a\}$. In figure 4, d_1 contains four zones that are inside existential contours. The letters assigned to these zones are derived from the 'underlying' zone in which the existential contours are placed. The three zones inside the contour A all contain the letter a under l whereas the three zones outside of A all contain the letter b under l. Using l, we are able to associate letters with zones. First, we define for each letter $a \in \Sigma$, the set of fixed contours (fc) that 'contain' that letter:

$$fc(a) = \{c \in \mathcal{FC} : a \in l(c)\}.$$

We can think of the zone $(fc(a), \mathcal{FC} - fc(a))$ as containing a. For instance, if we take the mapping $l(A) = \{a, b\}$ and $l(B) = \{b, c\}$ over the alphabet $\Sigma = \{a, b, c\}$ and with $\mathcal{FC} = \{A, B\}$ then the letter a has $fc(a) = \{A\}$ and is contained by the zone $(\{A\}, \{B\})$. Thus, a unitary diagram containing the zone $(\{A\}, \{B\})$ with a spider placed in that zone would assert the existence of a letter a in each word of the language it defines. A unitary diagram containing the zone $(\{A\}, \emptyset)$ with a spider placed in that zone would assert the existence of a letter a or a letter b in each word of the language it defines; see figure 6, where the diagram d_1 defines $(a|b|c)^*(a|b)(a|b|c)^*$ with the sub-expression $(a|b)$ arising from the spider. By contrast, a diagram such as d_2 containing both the fixed contours A and B can distinguish the letters a and b. Here, d_2 tells us that there is at least one a, using the spider, and no bs, using shading.

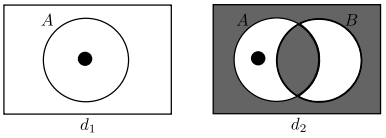


Figure 6. Distinguishing the letters a and b.

If two letters, a and b say, have the same image under fc then no diagram can define the language a^* for example. We require that no two letters are placed in any zone that arises from the function fc. Given $\mathcal{FC}$, Σ and a letter assignment l, we say that l is *valid* if the induced function fc is injective. From this point forward, we are assuming that a valid letter assignment, l, has been specified.

The language of a diagram, d, is defined by appealing to its semantics. In particular, we define when an interpretation models a word. A word, w, is in the language of d if all of the models of w are models for d. For example, in figure 6, d_1 has a model where $U = \{1, 2\}$, $\Psi(A) = \{1\}$, $\Psi(B) = \{1, 2\}$ and $<$ is the natural order over U. This interpretation models bc since there exists a bijection, f, from its letters to U ($f(b) = 1$ and $f(c) = 2$) such that

1. the order of the letters in bc agrees with the order of the elements under f and $<$, and
2. for each letter, w_i, in bc, the element to which it maps is in the set denoted by the zone which contains it.

We now define word models, extending and adapting [13].

Let $I = (U, \Psi, <)$ be an interpretation and let $w = w_1...w_n$ be a word drawn from Σ^*. Then I is a *model* for w iff there exists a bijection f, from the multi-set $\{w_1, ..., w_n\}$ to U such that for each w_i:

1. $f(w_i)$ is the ith element of the total ordering induced by $<$
2. $f(w_i) \in \Psi(fc(w_i), \mathcal{FC} - fc(w_i))$.

Furthermore, I is a *model* for language L iff there exists a word, w, in L such that I models w.

Let d be a second-order spider diagram and let L be a language. We say that L is the *language of* d, denoted $\mathcal{L}(d)$, iff the models of d are the same as the models of L. If L is the language of d then we say that d *defines* L [1].

For example, under the just defined function l with alphabet $\Sigma = \{a, b, c\}$, the diagram $d_1 \lhd d_2$ in figure 7 defines the same language as the regular expression $r_1 \cdot r_2$ where

$$r_1 = \big((a|b|c)^*a(a|b|c)^*b(a|b|c)^*\big)|\big((a|b|c)^*b(a|b|c)^*a(a|b|c)^*\big)$$

and $r_2 = cc^*$. For r_1, each disjunct arises from an ordering of the letters arising from the spiders.

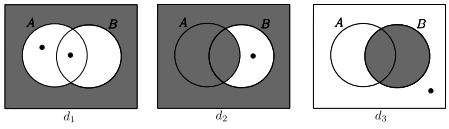


Figure 7. Defining regular languages.

Under the same l, the diagram d_3 in figure 7 does not define any language. This is because the zone $(\emptyset, \{A, B\})$ is (a) not shaded, and (b) is not assigned any letters under l. More formally, d_3 has models that are not models for any words and, therefore, not models for any language. Such a model for d_3 is given by $I = (U, \Psi, <)$ where $U = \{1, 2\}$, $\Psi(A) = \{1\}$, $\Psi(B) = \emptyset$ and $<$ is the natural order over U; here, we have $\Psi(\emptyset, \{A, B\}) = \{2\}$. Any word, $w = w_1w_2$ that is modelled by I would need to have exactly two letters, in order that a bijection, f, exists from $\{w_1, w_2\}$ to U. Suppose that I does indeed model w. For the ordering of letters in w to respect $<$, we therefore have $f(w_1) = 1$ and $f(w_2) = 2$. Since I models w,

$$f(w_2) \in \Psi(fc(w_2), \mathcal{FC} - fc(w_2)). \quad (1)$$

Noting that, for any letter, w_j, in Σ we have $fc(w_j) \neq \emptyset$,

$$(fc(w_2), \mathcal{FC} - fc(w_2)) \neq (\emptyset, \{A, B\}). \quad (2)$$

It is relatively easy to show that distinct zones, (in_1, out_1) and (in_2, out_2) where $in_1 \cup out_1 = in_2 \cup out_2$ represent disjoint sets in any interpretations. Thus, from (1) and (2) it cannot be the case that

$$2 = f(w_2) \in \Psi(\emptyset, \{A, B\}) = \{2\}.$$

We reach a contradiction and, hence, the supposition that I models w is false. Therefore, d has a model that does not model any word. Thus, under l, no language has the same model set as d, so d does not define any language.

3.1 Properties of Regular Expressions

To establish that all regular languages, r, are definable by second-order spider diagrams our approach is to provide a construction of a diagram, d given r and show that d defines r. However, there are some technical difficulties for which we need to account. In particular, these difficulties center around four factors: the presence of λ (which denotes the empty word), the presence of $\emptyset$, the presence of *, and disjunctive expressions involving single letters that are nested within the $+$ operator (e.g. in $(a|bc)^+$, the single letter a is nested within $+$). To overcome the first three issues, we eliminate the use of these symbols where ever possible.

In particular, it can be shown that for every regular expression, r_1, there exists a regular expression, r_2, such that r_2 does not contain the symbols $\emptyset$, λ, or *, but may include $\cdot$, $|$ and $+$, and either

1. r_2 defines the same language as r_1,
2. $\lambda|r_2$ defines the same language as r_1,
3. λ defines the same language as r_1, or
4. $\emptyset$ defines the same language as r_1.

Thus, we only need to consider expressions like r_2, $\emptyset$ and λ. We give an example in the next section concerning the issues around expressions like a^+. In the expression a^+, we have a trivial disjunction involving just a; more complex regular expressions where problems arise include those like $(a|ab|c)^+$.

3.2 Defining Regular Languages

Our construction of a diagram, given a regular expression, is inductive mirroring the inductive construction of regular expressions and taking account of the observations in section 3.1. First, we describe how to construct diagrams for the expressions $\emptyset$ and λ. Then, we can focus on expressions that do not involve $\emptyset$, λ or * (i.e as described in section 3.1, namely expressions like r_2). Expressions like r_2 follow an inductive construction, starting with letters and joining them using $|$ and $\cdot$, and using $+$. Throughout most of this section, we take an example alphabet to be $\Sigma = \{a, b\}$ and fixed contour set to be $\mathcal{FC} = \{A\}$. A valid letter assignment is given by $l(A) = \{a\}$.

When $r = \emptyset$, the diagram we construct is $\perp$, the unsatisfiable unitary diagram. Clearly, r defines the empty language in this case and, since $\perp$ has no models it too defines the empty language. For $r = \lambda$, we take a unitary diagram, d, that contains all fixed contours and is entirely shaded. Here, r defines the language containing precisely the empty word. This word is satisfied by the interpretation, $I_\emptyset$, with empty universal set. Clearly, $I_\emptyset$ is the only model for d, since it is entirely shaded and contains no spiders. Thus, d defines the language containing precisely the empty word. We refer to d as $d + \lambda$, which can be seen in figure 8.

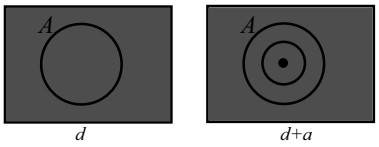


Figure 8. Constructing $d + \lambda$ and $d + a$.

For expressions like r_2, as a base case we have $r = w_i$ for some letter in Σ. We start with unitary diagram d that contains all fixed contours and is entirely shaded (i.e. $d+\lambda$). We add a single existential contour, placed inside the zone $(fc(w_i), \mathcal{FC} - fc(w_i))$, and a single existential spider whose habitat is the single zone inside the new existential contour. The diagram remains entirely shaded. We refer to this diagram as $d + w_i$ and it defines the same language as w_i. An example is shown in figure 8, where we have constructed $d + a$. At this stage in the construction, the existential contour is redundant. However, we are interested in providing an inductive construction and this will, depending on the regular expression, make use of the existential contour later in the process.

Our next example can be seen in figure 9, where the letter a has $fc(a) = \{A\}$ and b has $fc(b) = \emptyset$. Here, d_2 is the diagram $d + a$. To construct $d + ab$ we take $d_2 = d + a$, add a further existential contour, for the letter b, and place an arrow sourced on the existential contour for a and targeting the existential contour for b; this arrow asserts b is the successor of a. Thus, $d_3 = d + ab$. We can think of d_3 as being $d + a + b$, together with an arrow arising from the use of $\cdot$ (recall ab is really $a \cdot b$). Thus, this example gives an indication of how use use the inductive construction to build a diagram for a regular expression of the form $r_1 \cdot r_2$.

To create $d + (ab)^+$, we need to add some further syntax to d_3, accounting from the presence of $+$. In this case, the syntax is a pair of fresh existential contours with an arrow between them, as shown in d_4. Together, these contours and the arrow tell us that the only successors of bs are as. The original arrow, together with the alteration to the shading, now tells us that the only successors of the as are bs. It should be apparent that the language defined by $d_4 = d + (ab)^+$ is the same as that defined by the regular expression $(ab)^+$. Again, this example illustrates the general idea behind our construction of a diagram for regular expressions of the form r^+.

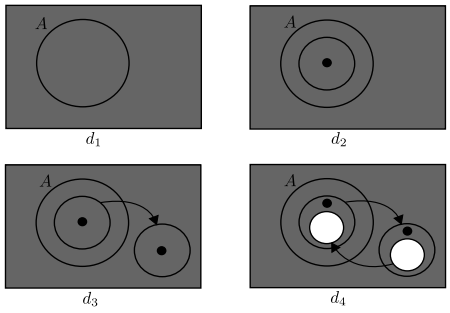


Figure 9. Constructing $r = (ab)^+$.

Regarding $|$, consider $r = (a|b)$. Here, we construct $d + a + b$, as shown in figure 10, clearly following the inductive construction (a and b are base cases). Next, we join the two spiders, creating a single spider that denotes the existence of an a or a b.

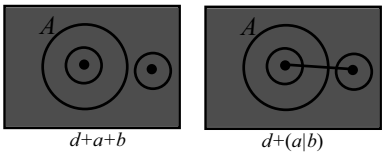


Figure 10. Constructing $r = a|b$ and $r = ab|a$.

Some technical difficulties arise from the presence of expressions like a^+. When we use a successor arrow between two existential contours, we are forcing every element in the source set to have a successor. Intuitively, this type of construction places a constraint on the relative numbers of elements in the source and elements in the target: the cardinalities of the source set and target set must be the same. In the case of $(ab)^+$ above, this causes no problem, since we require the same number of as and bs. However, in the case of a^+, the language defined using this type of approach is not $\{a, aa, aaa, aaaa, ...\}$.

A diagram like d in figure 11 does not define a^+ but instead defines $(aa)^+$. The righthand diagram $d + a^+$, with existential contours EC_1, EC_2 and EC_3, defines a^+. To justify this, consider the letter that arises from the presence of the spider. This letter must be an a due to the location of the spider. Clearly, because of the shading, all words in the language defined by $d + a^+$ contain only as. Moreover, because of the spider, the language defined by $d + a^+$ does not include the empty word. In summary, the language defined must be a subset of $\{a, aa, aaa, ...\} = a^+$. Finally, we must justify that given any word containing only as, except the empty word, is in the language of $d + a^+$. To do so, it

is sufficient to argue that $d + a^+$ has models with any non-zero cardinality (where the cardinality of a model is taken to be the cardinality of its universal set). Clearly, $d + a^+$ has a model with cardinality 1. Suppose $d + a^+$ has a model $I = (U, \Psi, <)$ with cardinality n; without loss of generality, we take $U = \{1, 2, ..., n\}$. Create a new interpretation $I' = (U \cup \{n+1\}, \Psi', <')$ where $n+1$ is in the set $\Psi'(A)$ and $<'$ agrees with $<$ but with $n+1$ ordered last. It can then be shown that, taking $\Psi(s) = \{1\}$, $\Psi(EC_1) = \{n+1\}$, $\Psi(EC_2) = U - \{n+1\}$ and $\Psi(EC_3) = \{1, n+1\}$, I' is a model for $d + a^+$. By induction, $d + a^+$ has models of every cardinality and, thus, defines the same language as a^+, as required.

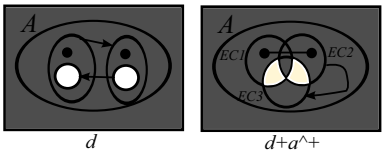


Figure 11. Difficulties constructing a^+

Finally, as a more complex example, we take $\Sigma = \{a, b, c, d\}$, $\mathcal{FC} = \{A, B\}$, $l(A) = \{a, c\}$ and $l(B) = \{b, c\}$. Then the lefthand diagram in figure 12 defines $(ab)^+cd$, whereas the righthand diagrams defines $((ab)^+cd)^+$.

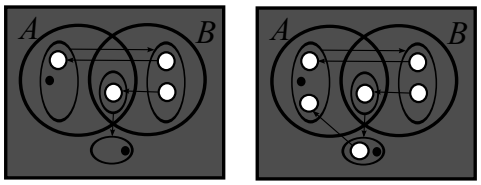


Figure 12. Constructing $(ab)^+cd$ and $((ab)^+cd)^+$

For space reasons, we are unable to elucidate on all of the details and omit discussion on various technical issues that arise. However, we give our primary result:

Theorem 1. *Given any regular expression, r, there exists a diagram that defines the same language as r.*

4. Conclusion

In this paper, we have presented a second-order diagrammatic logic for expressing constraints over an ordered universe. We have established that second-order spider diagrams can define all regular languages. Since spider diagrams of order can only define the star-free regular languages (being a monadic first-order logic with an order relation) this constitutes a substantial increase in expressive power. Indeed, this also demonstrates that second-order spider diagrams are substantially more expressive than all of the other monadic first-order diagrammatic logics that have been developed to date (such as Shin's Venn-II system and Swoboda and Allwein's Euler/Venn system). This increase in expressiveness has been achieved with relative small additions to the syntax (existential contours and arrows).

There are various avenues for future work. First, we aim to characterise the expressiveness of the new logic with respect to symbolic logics. It is known that monadic second-order (symbolic) logic with an ordered universal set can define the regular languages. However, the exact relationship between second-order spider diagrams and monadic second-order symbolic logic is unclear. This paper also lays the foundations for devising more expressive, yet intuitive, visual logics. The restriction to a monadic second-order logic with an order relation still precludes the formalisation of many higher-order concepts, such as constructing the transitive closure of a relation.

Acknowledgement This research is supported by EPSRC grants EP/H012311/1 and EP/H048480/1.

References

[1] P. Chapman and G. Stapleton. Introducing second-order spider diagrams for defining regular languages. *Visual Languages and Human-Centric Computing*. IEEE, 2010.

[2] A. Delaney and G. Stapleton. Spider diagrams of order. *Visual Languages and Logic*, *CEUR vol 274*, 27–39, 2007.

[3] A. Delaney, G. Stapleton, J. Taylor, and S. Thompson. A diagrammatic characterisation of commutative star-free regular languages. In preparation.

[4] A. Delaney, J. Taylor, and S. Thompson. Spider diagrams of order and a hierarchy of star-free regular languages. *Diagrams*, 172–187. Springer, 2008.

[5] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Springer, 1991.

[6] J. Gil, J. Howse, and S. Kent. Formalising spider diagrams. *Visual Languages*, 130–137. IEEE, 1999.

[7] E. Hammer. *Logic and Visual Information*. CSLI, 1995.

[8] A. Mateescu and A. Salomaa. Formal languages: an introduction and a synopsis. *Handbook of formal languages*, 1–39. Springer, 1997.

[9] C. Peirce. *Collected Papers*, vol. 4. Harvard University Press, 1933.

[10] A. Shimojima. Inferential and expressive capacities of graphical representations: Survey and some generalizations. *Diagrams*, Springer, 2004.

[11] S.-J. Shin. *The Logical Status of Diagrams*. CUP, 1994.

[12] N. Swoboda and G. Allwein. Using DAG transformations to verify Euler/Venn homogeneous and Euler/Venn FOL heterogeneous rules of inference. *Journal on Software and System Modeling*, 3(2):136–149, 2004.

[13] W. Thomas. Classifying regular events in symbolic logic. *Jnl of Computer and System Sciences*, 25:360–376, 1982.

Image Annotation System Using Visual and Textual Features

Swapnil Barai and Alfonso F. Cardenas
Computer Science Department
University of California, Los Angeles, USA

Abstract

We present an automated Image Annotation system called I-Tag which uses both visual and textual information of the images and recommends relevant tags for them. The automatic generation of metadata would allow image searches and content-based image retrieval (CBIR) to be more effective. We use state of the art tools on text based retrieval and image content based retrieval to retrieve similar images. We use an open source image retrieval engine called FIRE [8], to perform content based image search. We use the visual features as well as the title of the image to recommend tags for the image with high relevance. Experiments are conducted on approximately 5000 images downloaded from FLICKR to compare the accuracy of the recommended tags. Our experiments show that the mean average precision of the recommended tags is 69%.

Index terms: image annotation, tagging, image retrieval, content-based image retrieval.

1. Introduction

Image annotation also known as image tagging is the process of assigning metadata in the form of captions and keywords to digital images. Developing effective methods for automated annotation of images and the consequent retrieval of images aided by such annotations is a relatively new and highly active area due to its wide importance in the internet, and still an open research area [1]–[35]. Major search engines, such as Google, MSN, and Yahoo! that provide image search depend heavily on the name of the images, their textual description and the text surrounding the images, found on the Web pages containing these images. Hence computer assisted annotation of images would guarantee their visibility as search engine result.

Manual image annotation has been in existence for quite a while. Image sharing websites such as Picasa[34], Flickr[13], etc. give their users the ability to annotate the images while uploading them. These annotations are generally one or two words describing the image. However the major drawback of this technique is that the tags are user specific. Various attempts have been made towards automating this manual process [1]. Some analyze the information generally present in the tags that people give to their images, and henceforth recommend tags for their images [10]. Some World Wide Web image search engines attempt to automate annotation by using text from the image filename and text near the image on a webpage. However, search results using this method usually contain many irrelevant images. Most of the real world systems just rely on the textual features that are present with the image such as the title, the text present around an image on the webpage from where it is uploaded [11]. Also Google has made attempts towards this by involving the user in understanding the semantics of the images by developing an Image-Tag game Image Labeler [11]. However they have not come up with an Image tag recommendation system yet.

Efforts have been devoted also for more computer-aided automatic annotation based on image features. To annotate an image, different visual features, for example, color, texture, and shape are extracted. Then, using some classification algorithm, an automated system will be able to annotate the image. An image annotation system is very closely related to image retrieval system. An image retrieval system can retrieve images using the text in the annotated images which is more naturally represented in comparison to non-annotated images. There are three basic image annotation approaches: free text annotation, keywords and ontologies [1]. Keywords annotation is the use of arbitrarily chosen keywords or keywords from controlled vocabularies and is the most widely used. Free text annotation does not have any restriction on how to use text to annotate image. Ontologies are the use of a large classification system that classifies different aspects of life into hierarchical categories [23]. ImageCLEF provides the visual concept detection task [25] and medical image annotation task [24] for this area. Both task aims to annotate images using ontologies. There have also been many efforts to reduce the annotation retrieval gap [1],[2],[3],[4],[7] which refers to the difference in metadata generated by the human experts and those by systems for the same image. Automatic annotation of images using Statistical Learning Approach [5] and hidden Markov models [6] has also been studied. However reliable automatic image annotation still remains an open problem.

We present in this paper an automated Image Annotation system called I-Tag which uses both visual and textual information in the title of the images and recommends relevant tags for them. We believe that

by accounting for both visual as well as such textual features of images it is possible to reduce the annotation retrieval/semantic gap and provide highly relevant tags for them. [35] tries to utilize both visual and textual features, but therein the textual features refer to the text surrounding the image on the webpage from where it is taken, and this would lead to the semantic gap problem.

2. Dataset

The dataset consists of 5000 images downloaded from Flickr [13]. Flickr is a website that allows users to upload images to it and provides online image management. Through Flickr, users can share their images with other Flickr users as well as other people in the world.

There are several reasons for choosing to work with Flickr. Firstly, they have a large number of images available for download since they have a lot of users uploading images to their website every day. This makes it easier to build our data set. Also along with the images the metadata associated with them is also available. Flickr provides an API for developers to easily access its information, such as the images, user information, and image information. Flickrj is an API kit developed for Java to access the Flickr API and is a Java API which wraps the REST-based Flickr API. It covers most of the functionality that Flickr API provides. Images in the interesting photo section, along with their titles, tags, and number of favorite, are downloaded directly using Flickrj.

The downloaded images are chosen such that at least one of the following words, “Clouds”, “Flower”, “Tree”, “Beach” , “Sunset”, “Winter”, “Night”, “Grass”, “Dog”, and “Water”, in their title or tags are chosen to build up the Flickr data set to be used in the experiment. We also downloaded images having two of these keywords occurring together in their title or tags such as, “Beach” and “Water” or “Winter” and “Night” etc. Similarly we also had in our dataset images with triplets of these keywords and images with quadruples of these keywords occurring together in their title or tags. We conducted separate experiments on these sets of images and observed the variation of results. We do not use more than quadruples of keywords as there was no significant improvement in our results after that. Using interesting photos and restricting the images to have certain keywords ensured the quality of the photos and reduced the noise of the dataset. The reason for using these keywords is because they appear more frequently in the title or the tag of an image and therefore, the data set is easier to build. However, our system should work well with other less popular terms and image data sets – depending of course in the quality of tags assigned by the public to those images.

3. Image annotation system and implementation

In this section, we describe the image annotation system I-TAG that we have built, the components of the system and their implementation, and content-based image retrieval system highlights.

3.1 Overview

The objective of the textual and content based image annotation system is to allow the users to give images to the I-TAG tool and get back relevant tags for them as the result. Fig. 1 shows a high level block diagram illustrating our system. There are three main tasks involved in building an annotation system for images. The first task is to build a dataset for the annotation system, comprising of images from Flickr. This is done by the administrator of the system. The images for the dataset are downloaded from Flickr along with their respective title and tags. Then the user of the system gives an image to the system as input, which is to be annotated automatically by the system, referred to as the Query Image X. The next task for the administrator is to find images in the downloaded dataset, that are visually similar to the query image X. To perform this task state of the art content-based image retrieval engine is used that can extract visual features of a given image and return images that are visually similar to it. Also along with each of these similar images it generates a similarity score ranging from 0 to 1. Higher similarity score implies that the image is more similar to the original query image. The administrator orders these images based on this similarity score. Then the administrator extracts textual features (tags and/or titles) of all these similar images and the query image itself and forms a list of all the keywords. Finally the administrator finds the most frequently occurring keywords in this list and recommends it as the tags for the query image X.

3.2 Image Database

First we download images along with their tags from Flickr and build up our dataset as mentioned earlier. Obtaining this information is relatively easy since Flickr provides an API for developers. Flickrj [14] is an API kit developed for Java to access the Flickr API. It is a Java API which wraps the REST-based Flickr API. We store the image id, image title and the image tags in a MySQL database [9]. Next we divide the dataset into testdata (for which to generate its tags) and traindata i.e. we store them as two different MySQL tables. The testdata table has the following columns: image id, image title and tags that are to generated. The traindata table has the following columns: image id, image title, original tags (tags that are downloaded along with the image from flickr) and recommended tags (this column would be empty initially). The images

in the traindata form the dataset to submit to the image retrieval engine FIRE. The images in the testdata form the set of query images X, i.e. those images for which our tool has to recommend tags. The tags that have been downloaded along with these test-data images from Flickr are the Original or the 'Ideal' tags and serve as the comparison base in the evaluation of the tags recommend by I-TAG for these images.

3.3 Content Based Image Retrieval

The Flexible Image Retrieve Engine (FIRE)[8] is used as a "black box" to perform content-based image search in our I-TAG system. FIRE has participated in the ImageCLEF since and has shown to be one of the best among other content based image retrieval system (CBIR) participants in both photo and medical retrieval tasks. It is available for download on the FIRE website. The initial screen of FIRE shows several randomly chosen from the database to present to the user. The user can choose one of the images to start a query, ask the system to present more random images or upload one's own image as a query image. Unique to FIRE, but not to most of the several available CBIR systems, is the flexibility for a user to upload an arbitrary set or database of images and then obtain from FIRE the images in it most similar to an arbitrary query image or images provided by the user.

After getting back the result from FIRE, the user can start a new query by choosing other random images or uploading his own image. The user can refine the retrieval result by giving feedback to FIRE using three buttons below the image, the user can tell FIRE if the image is relevant or not: the green plus sign means the image is relevant, the red minus sign means the image is not relevant and the grey sign is chosen by default and means that the user does not have any opinions on that image. The user can look at the information of a particular image in the retrieval result, the full size image and the information of the visual features extracted.

Feature Extraction and Image Retrieval

For each image, FIRE extracts five visual features: color histogram, Tamura features, global texture descriptor, invariant feature histograms, and sparse patch histograms. Invariant feature refers to a feature that does not change when transformations are applied to an image. The transformations considered are translation, rotation, and scaling [30]. The invariant feature histogram is used in FIRE.

To decide which image in the database should be retrieved for an image query X, the distance score between the query image and each image I in the database is calculated. Each feature is assigned a weight and the distance score is the weighted sum of all the features.

We have also manipulated the code of FIRE such that it returns the name of the similar images as well in the final result. This is required specifically for our tool, as once we get the name of these similar images then only can we retrieve their textual information (title and tags) from our database.

3.4 Tag Recommendation

Now we can recommend tags for the query images X s. Using FIRE we got the most similar images (from the train dataset) to the query image. We use the image titles of these similar images to form a list of all the tags that have been associated with these similar images. We then find the most popular tags among this list and add them to the recommended tags column in the test dataset for that particular query image.

To improve our results we have also added the title of the similar images as well as the title of the query image to the list that we formed of all the associated tags. This helps as follows: for example the query image has title 'beach volleyball', but due to some reason the tag 'volleyball' is not a part of the most popular tags list and hence is not recommended. Thus by adding the title of the query image we are increasing the chances of the tag 'volleyball' to be returned as one of the recommended tags. The same reason holds good for why we are adding the titles of the similar images to our list. To summarize the tool should use all the data that is available to it for the task of recommending tags.

4. Experiments and Results

This section explains the four experiments performed over our dataset. The images in the test dataset act as the query images for our system. Similar images for these query images are searched from the images in the train dataset. The experiments compare the tags that are recommended for the query images by our system, with the "Ideal Tags" of these images or the tags that were associated with the Image on Flickr. Our goal is to show that the tags recommended by our system are accurate and relevant.

We have four different sets of images in our dataset. The first one comprises the images that have at least one of the following words, "Clouds", "Flower", "Tree", "Beach" , "Sunset", "Winter", "Night", "Grass", "Dog", and "Water", in their title or tags. The second has images with two of these keywords occurring together in their title or tags such as, "Beach" and "Water", "Winter" and "Night" etc. Similarly third and fourth datasets have images with triplets of these keywords and images with quadruples of these keywords occurring together in their title or tags. We refer to these sets of images as 'One-Word queried images', 'Two-Word queried images', 'Three–Word queried images' and 'Four-Word queried images' respectively. We have

conducted separate experiments on these sets of images and observed the variation of results.

For each of these sets of images we divide them into test-dataset and train-dataset. Test-dataset has the images that are uploaded to our system one by one for annotation i.e., they contain all the Query images. Each query image is then uploaded to our system. Next the administrator of the system finds images in the train dataset that are visually similar to the query image using FIRE. FIRE retrieves the ten most similar images to the query image, based upon a similarity score. After retrieving the ten most similar images to the query image, the administrator passes the titles of all these similar images to our database which has the textual information of all these images. Next using the titles of these similar images the administrator finds the tags that are associated with these images in the traindata table and forms a list that contains the titles and tags of the similar images. To the list we also add the title of the query image. This is done to improve the accuracy of recommended tags. Now from this list thus formed, we find out the most frequently occurring keywords and recommend them as tags for the query image. Finally we store these recommended tags in the "recommended tags" column of the testdata table for the respective query image. We repeat these steps for each of the query images in the four different sets of images we have. The following example illustrates all this visually:

Suppose the user gives the following image as input to the system.
Title: Beach Volleyball
Ideal/Original Tags: Beach, Volleyball, Sport, Sand
Recommended Tags: To be recommended by the annotation system



This image becomes our Query image X and has been selected from the test dataset. Table 1 shows a snapshot of the test dataset (before the annotation of the image). The tags that have been downloaded along with these test-data images from Flickr are the Original or the 'Ideal' tags. It is clearly visible that this image is a part of the One-Word queried images set, as it contains only one keyword 'Beach', out of our list of ten keywords, in either its title or tags.

Then our system extracts the visual features of the query image X and using the CBIR FIRE system finds visually similar images to it, in the images in the train dataset of our system. FIRE returns the following images are the four most similar images in the train dataset to the query image X.

Title: Volleyball 123X104





Game



Volleyball Match



Next the titles of these similar images are sent to the train dataset table, to get the Tags associated with the respective images. Table 2 shows a snapshot of the Train Dataset table.

Image Id	Image Title	Ideal Tags	Recom. Tags
1	Beach Volleyball	Beach,Volleyball, Sport,Sand	

Table 1: Snapshot of Test Dataset Table before the recommendation process

Image Id	Image Title	Tags
1	Volleyball	Players,Shot,Ball,Sport
2	123X104	Beach,Net,Volleyball
3	Game	Sand,Volleyball,Jump,Sport
4	Volleyball Match	Sport, Net, Beach, Shot

Table 2: Snapshot of Train Dataset Table

Now the administrator forms a list of Tags and titles of the similar images and also including the title of the query image. Hence the list is now:
Beach Volleyball, Volleyball, 123X104, Game, Volleyball Match, Players, Shot, Ball, Sport, Beach, Net,

Volleyball, Sand, Volleyball, Jump, Sport, Sport, Net, Beach, Shot.

Finally from the list thus formed, the administrator finds out the most frequently keywords and recommends them as tags for the query image to the user, as shown below.

Query/Input Image
Title: Beach Volleyball



Output, recommended tags:
1) Volleyball, 2) Beach, 3)Shot, 4) Sport

The test dataset after this is:

Image ID	Image Title	Ideal Tags	Recom. Tags
1	Beach Volleyball	Beach, Volleyball, Sport, Sand	Volleyball, Beach, Shot, Sport

Table 3: Snapshot of Test Dataset Table after the recommendation process

Result evaluation

The results are evaluated using precision, average precision and mean average precision (MAP)

Precision = (Number of recommended tags) ∩ (Number of original/'Ideal' Tags) / (Number of recommended tags)

Average precision refers to the average of the precision for the five query images of the same word. Here is the definition of average precision:

Average Precision = Sum of Precision / Number of Query images

MAP is the average of all the average precisions. It is calculated for each sets of images, i.e. One-Word Images, Two-word Images, Three-word images and Four-word Images.

MAP = Sum of Average Precision of all query images / Number of Query images

Usually, recall and F-score are also used to evaluate the result of retrieval systems. However, they are not used here as they do not have much applicability.

Recall = (Number of recommended tags) ∩ (Number of original/'Ideal' Tags) / (Number of original/'Ideal' Tags)

$F\beta = (1 + \beta 2)$ (Precision x Recall) / ($\beta 2$ Precision + Recall)

The number of original/Ideal tags that any test-image has, is totally dependent on the user who has uploaded that image (thus is quite variable). As recall is significantly dependent on the number of original/ideal tags, its value does not us tell much about the efficiency of our system. Also F-score is dependent on recall, hence it is also not a useful measure of the efficiency of our system.

Precision results

The following table shows the results for Three-word query images, i.e. images that have any three of the following keywords "Clouds", "Flower", "Tree", "Beach" , "Sunset", "Winter", "Night", "Grass", "Dog", and "Water" occurring together in their title or tags. We have approximately 1200 Three-Word query images (divided into testdata and traindata) in our dataset.

Query Image having keywords:	Average Precision (%)
Winter,Night,Tree	67
Beach,Sunset,Water	68
Flower,Tree,Grass	68

Table 4: Results of Three-Word Query Images

Thus the Mean Average Precision for the Three-Word query images is Sum of all Precisions/3= **67.6%.**

The comparison of results for One-word, Two-word, Three-word and Four-Word query images are shown below in Fig. 2. We observe good precision values in recommending tags for the images. We can easily see that the precision value for Four-Word query images is as high as 69%. Our remarks are qualitative opinions because there is no comparable data set in the literature for comparison.

Thus it is clearly visible that as the number of keywords increase in the query images, the precision values for the annotations also improve as our results show in Fig. 2. This can be accounted for the fact, that as the number of keywords increase in the query image, the image becomes more specific and users ought to tag it with similar keywords and hence our system gets better results

During the course of the experiments another important observation was made. Until now we were assuming the tags for the query images given by the users on Flickr as 'ideal tags', and we were trying to compare the tags recommended by our system for that image with its 'ideal tags' and calculate the precision. However it might be that these ideal tags originally given by the user(s) are not very accurate. Hence we manually tried to observe the relevancy of tags (as a neutral user) recommended by our system for a sample set of 100 images. We found out that for every 1 in 10 images the precision of tags was actually better than the precision calculated based upon the 'ideal tags'. Hence the precision goes up by 1.1% incrementally, and to a total level of 70.1% (for Four-word queried images).

Comparison to other annotation approaches is very difficult because their image test data sets and experiments details are not published.

.5.Conclusions

In this project we have developed an automatic image annotation tool I-TAG which uses both visual and textual features and recommends relevant tags for them with high relevance. We showed that by accounting for both visual and textual features of the image that we are able to annotate the images with high precision. We chose the Flexible Image Retrieval Engine (FIRE) to perform similar image search based on visual features. FIRE is one of the best content based image retrieval systems in the ImageCLEF photographic retrieval task.
The quality or accuracy of the image similarity by FIRE, or any other system that could be used instead, has of course a direct influence on the quality and accuracy of the tags. We have used MYSQL as our database, to store the textual information associated with images and extract this information when required.

To annotate an image, first the user gives an image to the system which is called the query image. Then our system finds images in our database that are visually most similar to the given image using FIRE. Then the administrator of the system extracts the textual information i.e. titles and tags of these similar images and forms a list. To this list, the title of the query image is also added. Finally the most frequently occurring keywords in this list are recommended as tags for the query image.

We have downloaded a dataset of 5000 images from FLICKR between the dates of August 2009 and October 2009. The title or tags of these images has at least one of the following keywords "Clouds", "Flower", "Tree", "Beach" , "Sunset", "Winter", "Night", "Grass", "Dog", and "Water" for restricting the dataset.

Experiments were conducted to evaluate the accuracy of our annotation system. The tags recommended by our system are compared with the original/ideal tags of these images provided by users to calculate precision. Testing with independent user judges should be done in the future. We have tested four kinds of image sets: One-word query images, two-word query images, three-word query images and four-word query images. Experiments are performed separately on these four sets of images. Our results based on average precision and mean average precision are shown in Fig. 2, indicating maximum accuracy for four-word query images.

6.References

[1] A. Hanbury, "A Survey of Methods for Image Annotation", Journal of Visual Languages and Computing, vol. 19, Issue 5, pp. 617-627, October 2008.

[2]Datta, R., Ge, W., Li, J., and Wang, J. Z. 2006. "Toward bridging the annotation-retrieval gap in image search by a generative modeling approach", In Proceedings of the 14th Annual ACM international Conference on Multimedia (Santa Barbara, CA, USA, October 23 - 27, 2006), MULTIMEDIA '06. ACM, New York, NY, 977-986.

[3] Irfanullah, Nida Aslam, Kok-Keong Loo,Roohullah, "Semantic Annotation Gap: Where to put Responsibility?", JDCTA: International Journal of Digital Content Technology and its Applications, Vol. 3, No. 1, pp. 94 -97, 2009

[4] L. Hollink, G. Schreiber, J. Wielemaker, and B. Wielinga, "Semantic annotation of image collections", In Workshop on Knowledge Markup and Semantic Annotation, KCAP'03, 2003.

[5] Baigang Huang, Junshan Li, Shuangyan Hu, "Automatic Annotation of Images by a Statistical Learning Approach," *International Conference on Natural Computation*, vol. 4, pp. 477-480, 2008 Fourth International Conference on Natural Computation, 2008.

[6] Ghoshal, A., Ircing, P., and Khudanpur, S. 2005. Hidden Markov models for automatic annotation and content-based retrieval of images and video. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM, New York, NY, 544-551

[7] Stephan Bloehdorn, Kosmas Petridis, Carsten Saathoff, Nikos Simou, Vassilis Tzouvaras, Yannis S. Avrithis, Siegfried Handschuh, Ioannis Kompatsiaris, Steffen Staab, Michael G. Strintzis – "Semantic Annotation of Images and Videos for Multimedia Analysis". In ESWC pp. 592-607, 2005.

[8] Deselaers, T., "Features for Image Retrieval", *Computer Science Department*, Aachen, Germany, RWTH Aachen University, 12/2003.

[9] MySQL http://www.mysql.com/

[10] Sigurbjörnsson, B. and van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international Conference on World Wide Web* (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 327-336

[11] Behold http://www.behold.cc/
[12] Google Image Labeler http://images.google.com/imagelabeler/
[13] Flickr. http://www.flickr.com/
[14] Flickrj. http://flickrj.sourceforge.net/
[15] M.Grubinger, P. Clough, A. Hanbury, and H. Muller, "Overview of the ImageCLEFphoto 2007 Photographic Retrieval", Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007.
[16] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, W. Equitz, "Efficient and effective querying by image content", J. Intell. Inf. Syst. 3 (3–4), 1994.
[17] A. Pentland, R.W. Picard, S. Scaroff, "Photobook: content-based manipulation for image databases", Int. J. Comput. Vision 18 (3), 1996.
[18] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu, "The Virage image search engine: An open framework for image management", In I. K. Sethi and R. C. Jain, (editors), IS&T/SPIE Proceedings of Storage & Retrieval for Image and Video Databases IV, vol. 2670, pp. 76-87, San Jose, 1996.
[19] J.R. Smith, S.F. Chang, "VisualSeek: a Fully Automatic Content-Based Query System", Proceedings of the Fourth ACM International Conference on Multimedia, pp. 87–98, 1996.
[20] Elisa Drelie Gelasca, Joriz De Guzman, Steffen Gauglitz, Pratim Ghosh, JieJun Xu, Emily Moxley, Amir M. Rahimi, Zhiqiang Bi and B. S. Manjunath, "CORTINA: Searching a 10 Million + Images Database", Technical Report, VRL, ECE, University of California, Santa Barbara, Sep. 2007.
[21] J.Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 9, pp. 947–963, 2001.
[22] Gass, T., T. Weyand, T. Deselaers, and H. Ney, "FIRE in ImageCLEF 2007: Support Vector Machines and Logistic Regression to Fuse Image Descriptors in for Photo Retrieval", Advances in Multilingual and Multimodal Information Retrieval 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, vol. 5152, Budapest, Hungary, Springer, September 2007.
[23] E. Hyvonen, A. Styrman, S. Saarela, "Ontology-based image retrieval", Proceedings of XML Finland Conference, pp. 27-51, 2002.
[24] T. Deselaers, T. Deserno, N. Muller, "Automatic medical image annotation in ImageCLEF 2007 Overview, results, and discussion", Pattern Recognition Letters, vol. 29, issue 15, pp. 1988-1995 November 2008.
[25] T. Deselaers, and A. Hanbury, "The Visual Concept Detection Task in ImageCLEF 2008", CLEF Workshop 2008 / Evaluating Systems for Multilingual and Multimodal Information Access, Aarhus, Denmark, Springer, 2009
[26] Li, J. and Wang, J. Z. 2008. Real-Time Computerized Annotation of Pictures. IEEE Trans.
[27] Pon, R. K., Cárdenas, A. F., Buttler, D., Critchlow, T., "iScore: Measuring the Interestingness of Articles in a Limited User Environment", IEEE Symposium on Computational Intelligence and Data Mining 2007, Honolulu, HI, April 1-5, 2007.
[28] 3. F. Tushabe, M. H. F. Wilkinson (2008): "Content-based Image Retrieval Using Combined 2D Attribute Pattern Spectra", in book: Advances in Multilingual and Multimodal Information Retrieval by Lecture Notes in Computer Science, Springer, Vol. 5152, ISBN 978-3-540-85759-4.
[29] www.wikipedia.com
[30] T. Deselaers, "Features for image retrieval", diploma thesis, RWTH Aachen University, Aachen, Germany, 2003.
[31] H. Tamura, S. Mori, T. Yamawaki, "Textural Features Corresponding to Visual Perception", IEEE Transaction on Systems, Man, and Cybernetcs, Vol. SMC-8, No. 6, pp. 460–472, June 1978.
[32] J. Puzicha, Y. Rubner, C. Tomasi, J. Buhmann, "Empirical Evaluation of Dissimilarity Measures for Color and Texture", International Conference on Computer Vision, Corfu, Greece, 1999.
[33] http://images.search.yahoo.com/
[34] http://picasa.google.com/
[35] X. Li, L. Chen, L. Zhang, F. Lin, and W. Ma, "Image annotation by large-scale content-based image retrieval," in Proceedings of the 14th Annual ACM international Conference on Multimedia, 2006.

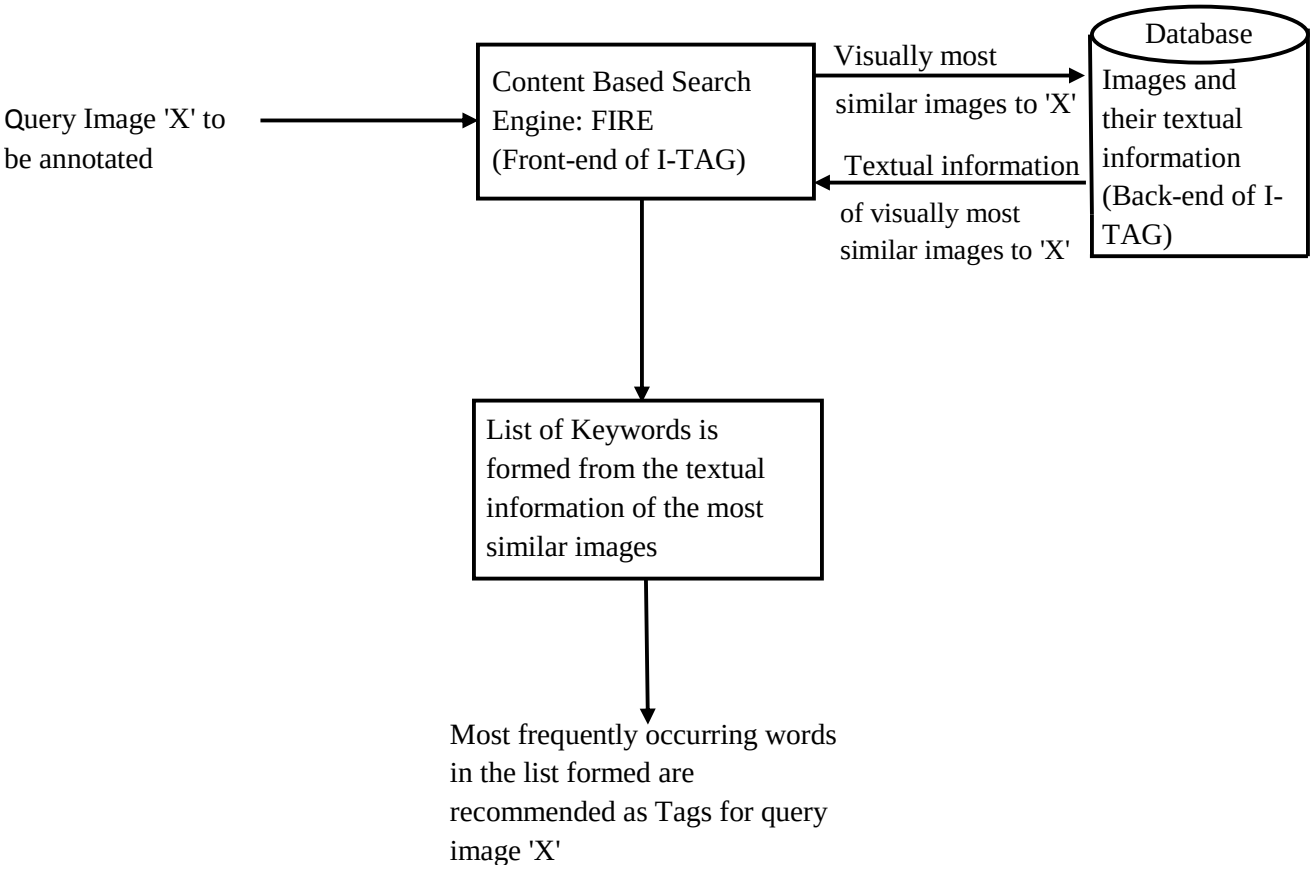


Fig 1: I-Tag system overview

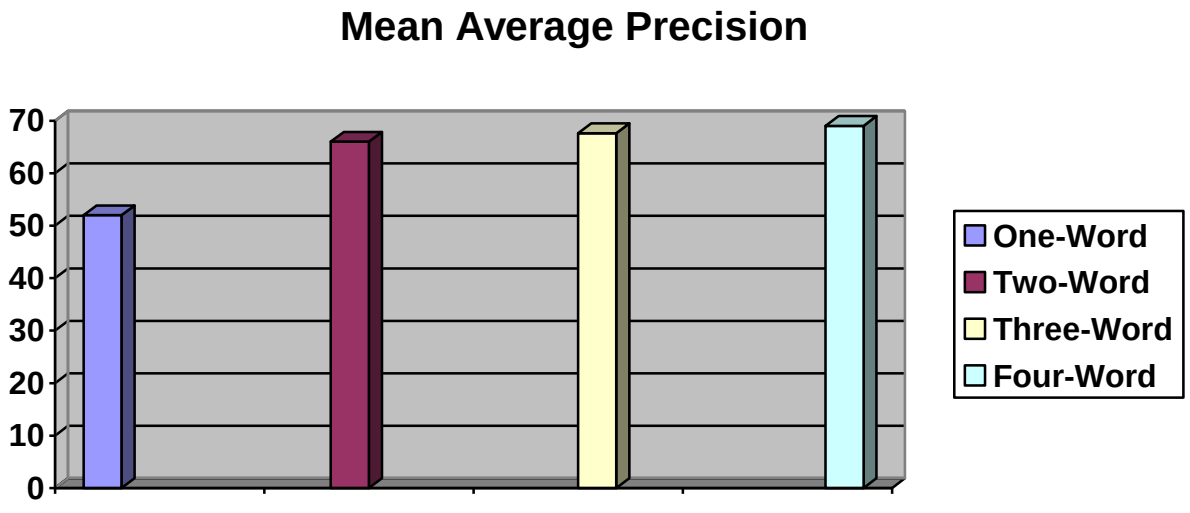


Fig 2: Mean Average Precision comparison

Reverse Engineering Visual Design Models from Code: What are the Issues?

Kendra M.L. Cooper
Department of Computer Science
The University of Texas at Dallas
800 West Campbell Road
Richardson, Texas, 75080, USA
kcooper@utdallas.edu

Jeffrey Koch
Department of Computer Science
The University of Texas at Dallas
800 West Campbell Road
Richardson, Texas, 75080, USA
jeffrey.koch@student.utdallas.edu

Abstract

Model-driven development addresses forward and reverse software engineering issues. Forward engineering issues have received a great deal of attention (model representations, transformations); however reverse engineering has received little attention. This short paper presents a preliminary list of issues in reverse engineering code into design models that researchers may use to help characterize and refine their problem statements. It highlights the extensive scope of the problem and the need for inter-disciplinary research to address the issues.

1 INTRODUCTION

Model-driven development addresses forward and reverse software engineering issues [1]. Forward engineering issues have received a great deal of attention. Modeling notations, specification, verification, and transformation approaches to systematically refine requirements to code have been proposed. In contrast, reverse engineering has received little attention. The ability to create a more abstract, visual design model on demand that accurately reflects (is consistent with) the code would be of great value, as it addresses two key issues. In plan-driven environments, the code and the design models may drift and become inconsistent over time, which reduces the value of the models. In agile environments, the design models are often not available and developers need to understand a (potentially large) code base to add a feature or correct a defect.

This short paper outlines key issues in reverse engineering code to design models. This paper is expected to a) motivate research, which would significantly impact the ability to abstract, understand, verify, and modify code; b) provide a framework for comparing reverse engineering approaches.

2 ISSUES

The reverse engineering issues are organized around the fundamentals of the activity: input, transformation, output (Figure 1). The goal of the activity is to create "good" visual design models. In other words, the models created are correct, complete, understandable, and so on. The event triggering the activity is a request to create the visual design model. The input is the code base; the output is the presentation of one or more design models. Issues related to processing the input to create the output, including abstraction constraints, are listed under the transformation.

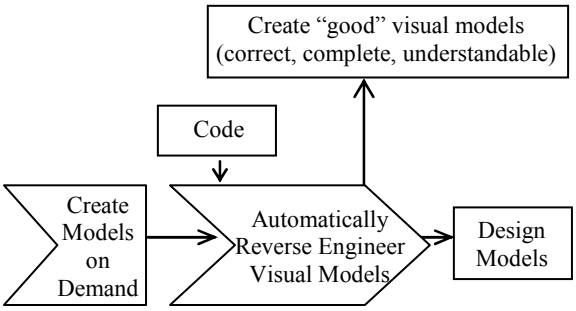


Figure 1. Reverse Engineering Code to Visual Design Models

2.1 Input

For the code to design reverse engineering activity, the input is the code base. The code base may be available as source code, byte code, or an executable. The code base may be a version of a large number of programming languages (currently popular or legacy). The software engineering research problem can be narrowed here by considering which language(s) to support and their form. From a long term perspective, a general solution based in programming language theory may be considered.

2.2 Transformation: Code to Design.

Abstraction Constraints. In addition to established object-oriented paradigm, numerous alternatives have recently emerged: component based, aspect-oriented, product-line, service-oriented, and agent-oriented [2]. As some (not all) of these paradigms have proposed new languages and there is no common re-use meta-model uniting the alternatives, software engineering researchers need to consider which paradigm(s) to support in the

shorter term; a re-use meta-model is needed to provide a theoretical foundation that spans paradigms.

Architectural styles and design patterns also provide abstraction constraints [3][4]. These collections continue to evolve; on-going research is needed to recognize styles and patterns in the code (unmodified and tailored), and how to support up to date collections. Code analysis and data mining techniques may be considered here.

Choice of Design Models. The design activity has evolved to include a collection of models. Multiple levels of design abstraction (analysis, architecture, detailed design) and multiple views of the models (structural, behavioral, process allocation, …) are established [6]. Capabilities to automatically reverse engineer code to the detailed design static model (e.g., class diagram) for the object-oriented paradigm model are available in commercial tool support; semi-automated capabilities are available to reverse engineering to the detailed design behavioral model (e.g., sequence diagram) [7]. A framework has been recently been presented to support the visualization of object-oriented designs at multiple levels of abstraction [8].

Consistency of the Design Models. The consistency of the models at multiple levels of abstraction and across multiple views is a key issue, as is the ability to support traceability relationships across models. The models and their relationships need to be rigorously defined; a survey of model transformation approaches in [9] presents a classification with five categories: direct-manipulation approaches, structure driven approaches, relational approaches, graph-transformation-based approaches, and hybrid approaches. A graph based framework has been recently been presented to support the consistent visualization of object-oriented designs [8]; little work is available to support other paradigms such as the aspect-oriented, agent-oriented, or service-oriented or product-lines.

Behavioral Design Models. The behavioral models reflect the system when it is executing. In forward engineering, the requirements for the system (e.g., use cases) can be used to select the execution(s) included in the behavioral design models (e.g., activity diagrams, sequence diagrams). The design models can be considered complete when all the requirements have been realized. When reverse engineering, key issues are how to create and select the executions to include in the design models. Code analysis and test case generation techniques from the software testing literature in combination with multi-criteria decision making techniques from the artificial intelligence community may yield useful, automated solutions to this problem.

Quality of the Design Models. When reverse engineering code to a design model, there may be more than one solution that is consistent. If so, then the alternatives need to be quantitatively, automatically evaluated using metrics. Metrics to support this evaluation are established for the object-oriented paradigm [5], but little is available for the more recent paradigms.

2.2 Output

The notations used to present the design models (forward or reverse engineering) may be informal, semi-formal, or formal. Informal box and arrow diagrams are useful for communicating design concepts at a high level. UML is a semi-formal, standardized, visual modeling language that has been widely adopted and extended to support alternative paradigms. Formal alternatives specifically for architecture level designs include Architecture Description Languages (ADLs) [10]. Given the extensive number of possibilities (current and future) a general all encompassing solution is a long term research endeavor.

3 CONCLUSIONS

The preliminary list of issues presented here indicates the large scope of the reverse engineering problem. This list can be refined and used as a framework to structure much needed research in the area. A survey of existing approaches in the literature, organized around the list of issues, is underway. We are applying this list to characterize a reverse engineering visualization problem for the aspect-oriented paradigm under investigation (Java/AspectJ code reverse engineered to multiple views of the architecture in a UML-based notation).

The broad range of issues highlights the need for fundamental, interdisciplinary research; they require expertise from communities such as programming languages, graph theory, artificial intelligence, data mining, and software engineering.

4 REFERENCES

[1] OMG Model Driven Architecture (MDA) Guide Version 1.0.1, March 2003. Available at: www.omg.org.

[2] K. Cooper, L. Dai, R. Steiner and R. Mili, "Survey of Software Architecture Approaches", book chapter in Designing Software-Intensive Systems: Methods and Principles, Idea Group Inc., May, 2008.

[3] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1995.

[4] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, Pattern-Oriented Software

Architecture Volume 1: A System of Patterns, John Wiley & Sons, 2001.

[5] M. Lanza and R. Marinescu, Object-Oriented Metrics in Practice: Using Software Metrics to Characterize, Evaluate, and Improve the Design of Object-Oriented Systems, Springer, 2006.

[6] K. Cooper, "Object-oriented Design and Analysis", book chapter in Encyclopedia of Computer Science and Engineering, John-Wiley & Sons Inc., vol. 4, 2013-2024, 2009.

[7] IBM Software Architect. Available at: http://www-01.ibm.com/software/awdtools/swarchitect.

[8] H. M. Kienle and H. A. Müller, .Rigi: An environment for software reverse engineering, exploration, visualization, and redocumentation,. Science of Computer Programming, vol. 75, no. 4, pp. 247.263, April 2010.

[9] K. Czarnecki, S. Helsen, Classification of model transformation approaches, in: Workshop on Generative Techniques in the context of MDA, OOPSLA, 2003. Available at: http://www.swen.uwaterloo.ca/_kczarnec/ECE750T7/czarnecki_helsen.pdf.

[10] R. Pandy, Architectural description languages (ADLs) vs UML: a review, ACM SIGSOFT Software Engineering Notes, Volume 35 , Issue 3 , May 2010, pp. 1-5.

CAVEMANDER: Creating 3-D Command-and-Control Scenarios for the CAVE Automatic Virtual Environment

Muhanna[1] Muhanna, Sermsak Buntha[2], Sohei Okamoto[1], Michael J.[1] McMahon, Jr., Sergiu Dascalu[1], Frederick C.[1] Harris, Jr.

[1] Department of Computer Science and Engineering
University of Nevada, Reno, USA
{muhanna, okamoto, mcmahon, dascalus, fredh}@cse.unr.edu

[2] Department of Weapons and Tactics
Royal Thai Navy Academy, Thailand
serm34@gmail.com

***Abstract* — Command-and-control training scenarios represent interactive exercises in efficient and timely decision making. Most often, maximizing the realism and impact of these scenarios requires the commanders be presented with as much relevant information as possible. While 2-D media are quite capable of presenting scenario information, 3-D media are able to present *and* relate that information in greater quantity and more meaningfully. Towards this end, we have created CAVEMANDER – a framework, API, and set of tools for developing immersive 3-D command-and-control scenarios in the Cave Automatic Virtual Environment (CAVE). This paper presents the main features of CAVEMANDER, focusing on the application of a software engineering process to develop artifacts (i.e., simulation scenarios) under our design-build-execute approach. This development sequence is illustrated using software engineering diagrams, a graphical user interface wizard developed as a part of CAVEMANDER to simplify use, and actual generated artifact file contents.**

Keywords — CAVE; CAVEMANDER; Command-and-Control; GUI Wizard; Software Engineering; Visual Design; VR.

I. INTRODUCTION

Command-and-control – or C&C – is an interactive exercise in which commanders must interpret all information for a given scenario and, based upon that information, reach reasonable conclusions in an efficient, effective, and timely manner [1]. Computers are increasingly taking on a central role in these scenarios, replacing large tactical glass boards or vertical paper charts to more effectively represent and convey the situation to the commander [2]. 2-D media (e.g. wall projectors and/or monitors), however, remain the predominant conveyances of scenario information [3] [4]. These computer-assisted representations of information still suffer from some inherent limitations, including: difficulties understanding, assessing, and relating geometrical parameters, delays in interpreting various data visualizations, and reduced awareness of the C&C situations [5] [3]. To improve C&C scenarios and address these issues, the Cave Automatic Virtual Environment (CAVE) was selected to represent the scenario information [1] – the advantages of using a third dimension in data visualization [6], combined with the rewards introduced by the immersive element of virtual environments [7], make the CAVE an excellent choice for representing C&C scenario information.

Several software tools and libraries are currently available to create virtual reality applications. CAVELib [8], for example, is a cross-platform application programming interface (API), used to develop virtual reality applications for various systems, including the CAVE. Originally developed at the University of Illinois in Chicago, CAVELib is now commercially owned by Mechdyne Corporation.

FreeVR [9] is an API that is used, primarily, to develop CAVE-based applications. This system provides an interface to utilize multiple devices available in virtual reality environments. VR Juggler [10] is a virtual reality application API that supports the execution of a virtual reality application on any display device using various input devices without the need to change or recompile the source code. CoVE [11] is a cross-platform toolkit built on Open Scene Graph that supports multi-user collaboration and multi-tasking efforts.

Quest3D [12] is a real-time 3-D engine and development platform that is intended to be used for developing software, 3-D web pages, and virtual reality simulators through visual interfaces. Equalizer [13] is a cross-platform toolkit designed to optimize parallel rendering while remaining sufficiently scalable to execute on various system configurations. Equalizer allows an application to dynamically adjust configuration at runtime, adapting to multiple processors or graphics cards, or utilizing computing cluster resources without any modification to the core code. Avango [14] is a scene graph based framework for developing virtual reality applications, where the node functionality is extended to provide additional dataflow and network communication.

CaveUT [15] is a modification of the Unreal Tournament 2004 game engine, adapted to run on CAVE-like systems. Unreal Tournament 2004 is a network-based game engine wherein the server maintains the master copy of the virtual world and each user is considered a client connected to the server, each with a different view. Each user's view is based on the position of their head, allowing realistic view shifting. Utilization of the game engine as the basis of the virtual world allows the production and rendering of realistic models of avatars and objects in the scene, as well as real-time interaction amongst users in the virtual world.

The above systems, as well as others [16], naturally have both common and distinguishing features; each focuses on developing virtual reality applications, yet each adds its own unique set of features and functionality. CAVEMANDER is similar to these systems in that it provides tools and functionality to create virtual environment applications. However, compared with other approaches (mostly focused on offering comprehensive API support), CAVEMANDER also applies a software engineering methodology and provides a set of tools to create a structured, easily extensible development system. Furthermore, besides a specialized API built on top of FreeVR [9], CAVEMANDER's set of available resources include a graphical user interface wizard, which simplifies the creation of C&C scenarios that enable immersion in the virtual environment of the CAVE.

The remainder of the paper is organized as follows: Section 2 presents the motivations behind CAVEMANDER, including a brief overview of its main components. Section 3 describes, in detail, the CAVEMANDER software development methodology. Section 4 demonstrates the system using the included graphical wizard interface to quickly create a C&C scenario. Section 5 indicates several potential directions for future work, concluding with a summary of developments in the CAVEMANDER system.

II. CAVEMANDER

CAVEMANDER has been developed to achieve several goals. Firstly, CAVEMANDER is targeted toward advancing the state-of-art in the development of C&C simulations by utilizing the CAVE – one of the most modern environments for experiencing and integrating new software engineering and human-computer interaction solutions. CAVE and associated technologies have been, and continue to be, in a state of rapid research and development [17]. Second, CAVEMANDER is intended to fill the gap between software development methods and software designed specifically to build CAVE-based applications. Third, CAVEMANDER seeks to contribute to the limited pool of CAVE-supporting resources, adding tools and reusable code to advance CAVE software development efforts worldwide. Lastly, CAVEMANDER provides a structured methodology and toolset-based foundation for future research and development of C&C simulation scenarios using the CAVE.

The CAVEMANDER approach consists primarily of a systematic software engineering method for a specialized software platform that includes a set of software resources [1]. The following sections focus on CAVEMANDER's software architecture, scenario development methodology, software artifacts, and GUI-based wizard.

III. METHOD ACTIVITIES AND ARTIFACTS

An overview of the activities involved in using CAVEMANDER is presented in Figure 1 in the form of a UML activity diagram [18], indicating both activities and the artifacts generated by those activities. The activity flow illustrates the steps required to generate a scenario using CAVEMANDER; the artifacts generated and passed amongst each step show the flow of those objects [19]. The four primary activities are: 1) *scene definition*, 2) *simulation build up*, 3) *scenario creation*, and 4) *scenario execution*.

A *scene* is a collection of mobile and/or fixed unit types (e.g. tanks, supply bases, vehicles) and environmental factors (e.g. visibility and weather) related to a C&C scenario. A scene is defined by specifying the units and/or environmental factors involved, as well as any parameters or commands related to them. A *simulation* adds the software implementation of items involved in a *scene*. Specifically, it is a set of classes associated with the unit types included in a *scene*, such as the ClassTank or the ClassSupplyBase. Finally, a *scenario* is a customized instance of a simulation. Here, the concrete types provided by the *simulation* are combined with specific values that are applied to the environment and units. In other words, instances of the unit types, along with specific parameters and environment variables, are defined to create the scenario. For example, a specific scenario might include three tanks and two supply units, each having specific parameters, all working in an environment in which the visibility is no more than five miles.

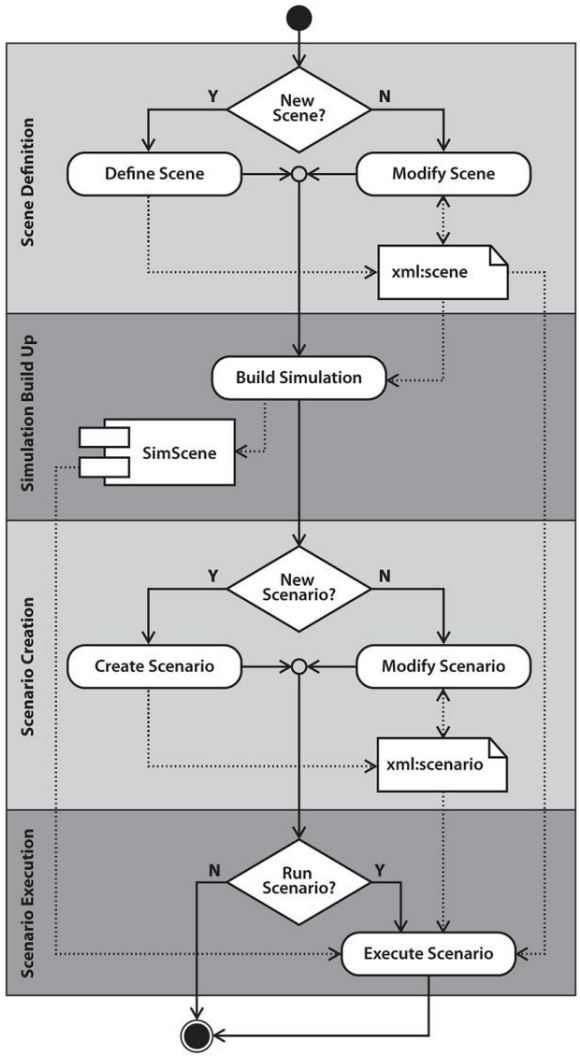


Figure 1. Overview of CAVEMANDER main activities and artifacts

In terms of the software development process, **Scene Definition** is a high-level design activity; **Simulation Build-Up**, however, is an implementation / coding, activity. In this step, existing/reusable code is integrated with new, customized code, to produce an executable software package: *SimScene*. **Scenario Creation** and **Execution**, however, are considered both design artifacts (in terms of describing the C&C situation) and implementation artifacts (as it executes in the CAVE).

A. Scene Defintion

Scene Definition is the first step in creating an executable scenario application. Users must either utilize an existing scene (optionally modifying it) or create a new scene. The scene definition activity is composed of three sub-activities: **Scene Concept Definition**, **Unit Specification**, and **Environment Specification**. **Scene Concept Definition** involves providing a text description of the scene (e.g. "A land C&C simulation" or "Naval operations").

The remaining two activities can be carried out in-parallel. **Unit Specification** entails defining the available models, properties, and commands for each unit type. Similarly, **Environment Specification** embodies the declaration of environmental factors and their properties. The result of this activity is the creation of an XML scene file, which contains the description of all elements that make up the scene (i.e. unit types and environmental factors).

B. Simulation Build-Up

During **Simulation Build-Up**, executable code is generated for the scenario - this code is contained in the package *SimScene*. As with the **Scene Definition** step, users must either use an existing simulation code package available from the CAVEMANDER system (optionally modifying it) or create a new one. Both cases involve code generation, requiring knowledge of a programming language.

Simulation Build-Up is divided into six sub-activities. First, an XML scene file – generated in the previous step – must be selected. Second, the classes needed to implement the unit types in the scene are either written or retrieved from the exiting CAVEMANDER library. Third, the classes needed to implement the environmental factors of the scene are either written or retrieved from the existing CAVEMANDER library. Fourth, the programmer includes all these implementation classes in the main simulation package: *SimScene*. Fifth, the programmer provides an update function for each time-step of the simulation. Finally, the code created in this activity is compiled and linked to the CAVEMANDER library, resulting in an executable file consisting of all elements that make up the scene.

C. Scenario Creation

Scenario Creation is the specification of the initial states and property values of the units and environmental properties of a specific scenario. As with the prior steps, users must either use an existing scenario (optionally modifying it) or create a new one. First, the concept of the scenario is specified as a text string (e.g. "Transporting supplies to front line units"). Second, the individual units in the scenario (not simply their types), their values (IDs), and their initial locations are specified. Third, the initial environment property values are specified. Fourth, the scenario is loaded on the client (CAVE) for later execution. The result of this step is an XML file containing the necessary property settings for all elements that make up the scenario.

D. Scenario Execution

In this activity, the resources required or generated for the simulation are loaded on the server and the client (CAVE) for execution. The user has the option to slow down, speed up, pause, resume, and stop the simulation. Support for a playback feature that allows reversing the progression of a scenario is included in the design of CAVEMANDER, however, the implementation of this feature is part of our future work plans.

IV. USER INTERFACE WIZARD

ServGUI – the CAVEMANDER Wizard – is intended for novice users [17]. Using this wizard, ServGUI enables graphical scene definition, scenario creation, and scenario execution. In other words, ServGUI provides a visual tool that gives the user the ability to create software artifacts without the need to write code.

The user can interact with the ServGUI wizard by creating a scene or a scenario and running a simulation scene, playback, or communication hub. Figure 2 shows the *type* tab of the ServGUI wizard during the creation of a military scene. Here, the user is a adding a Hummer vehicle to the scene. They will then be able to define several properties and commands for the Hummer using the *prop* and *cmd* tabs, respectively. Figure 3 shows the user adding an attack command to the hummer using the *cmd* tab. Using the *env* tab, the user is able to provide several environmental factors for the scene.

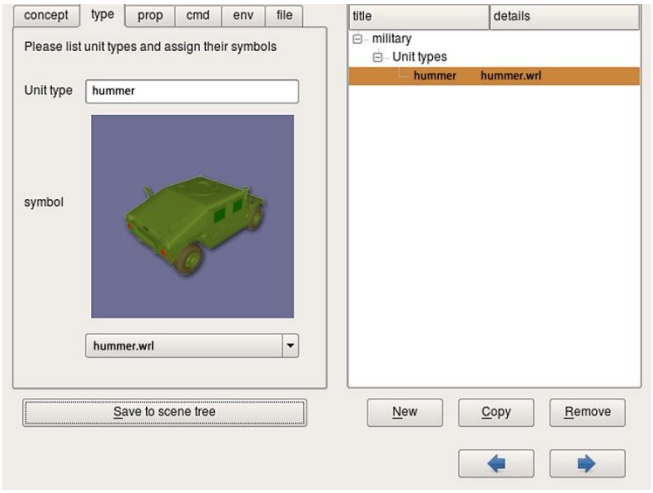


Figure 2. The *type* tab of the ServGUI wizard during the **Scene Defnition** activity

Once the scene is defined, it is saved as an XML file for use in simulation building, as shown in Figure 4.

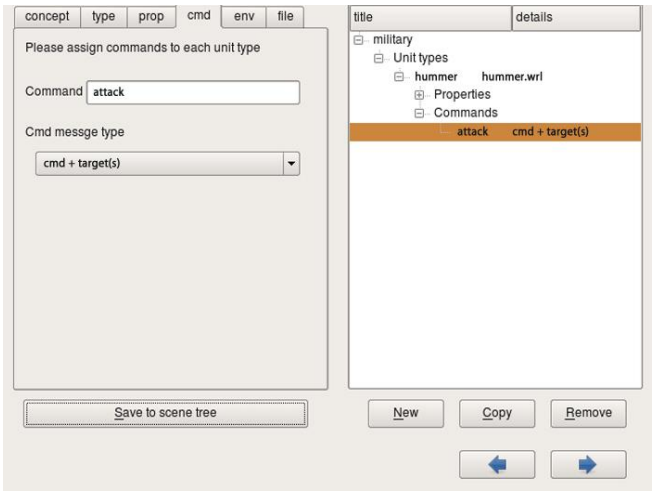


Figure 3. The *cmd* tab of the ServGUI wizard during the **Scene Definition** activity



Figure 4. Example of a scene artifact (XML file)from the **Scene Definition** activity

In the simulation build up software activity, the user assigns specific properties to each element of the scene. Figure 5 depicts the *unit* tab of the ServGUI wizard while the user is applying a different value for each of the Hummer properties. The specified values will be added to the simulation tree, which can be easily visualized by the user in the right-hand side of the user interface tab. Using this activity and the supported GUI, the user is also able to provide different values for environmental factors (e.g. visibility, temperature, wind speed, etc.).

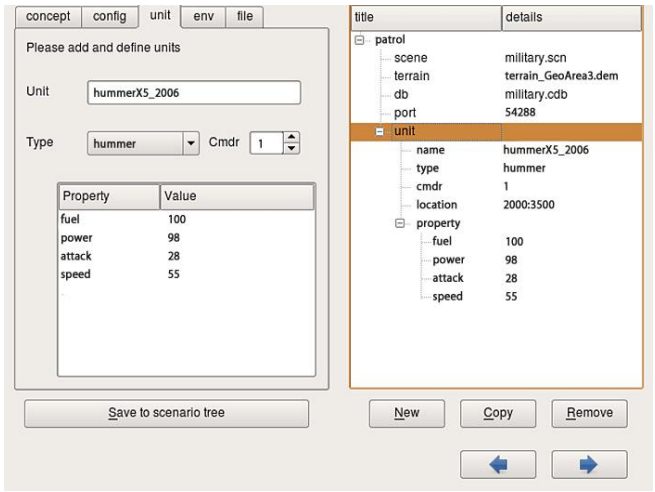


Figure 5. The *unit* tab of the ServGUI wizard during the **Simulation Build-Up** activity

In addition, the ServGUI wizard provides a tool that allows the user to visualize and interact with a simulation scenario during execution. This can be seen in Figure 6 where the user loads a scenario, executes it, and observes the values of the simulation elements and commands.

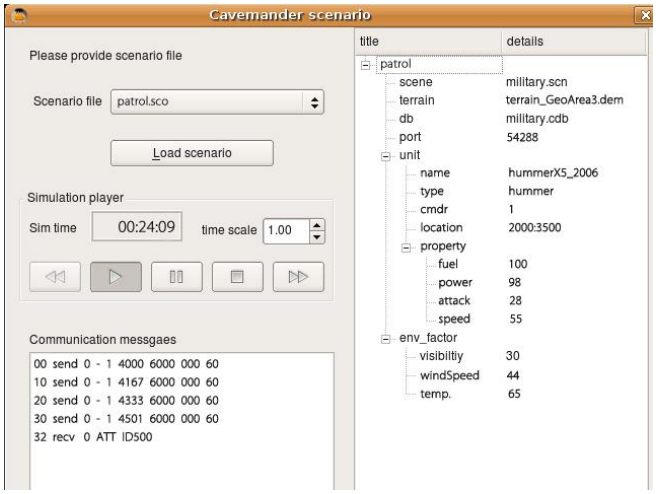


Figure 6. CAVEMANDER GUI for running scenarios

Figure 7 and Figure 8 show an example of a running simulation scenario built using CAVEMANDER. Figure 7 shows the scenario running in the FreeVR CAVE simulator [9], whereas Figure 8 shows the running scenario inside a 4-wall CAVE at the Desert Research Institute in Reno, Nevada.

V. FUTURE WORK AND CONCLUSION

CAVEMANDER has shown promising results during development, with significant potential for use by researchers, application developers, and other CAVE users [1]. It introduces a state-of-the-art software engineering approach to build C&C applications in the CAVE through

the use of a visual GUI wizard that can effectively produce several software artifacts in the construction phase of such applications.

We have identified several potential directions for future work that could enhance the current CAVEMANDER system and related tools. First, new and more complex military and non-military simulations / application scenarios need to be developed. Second, several additional, enhanced, and well-tested sets of software resources should be provided. Third, we plan to conduct usability studies with focus on creating immersive C&C training scenarios. Finally, we intend to further investigate other potential process and architectural improvements. More details about the CAVEMANDER approach, related tools and applications, including a complete C&C scenario, can be found in [1].

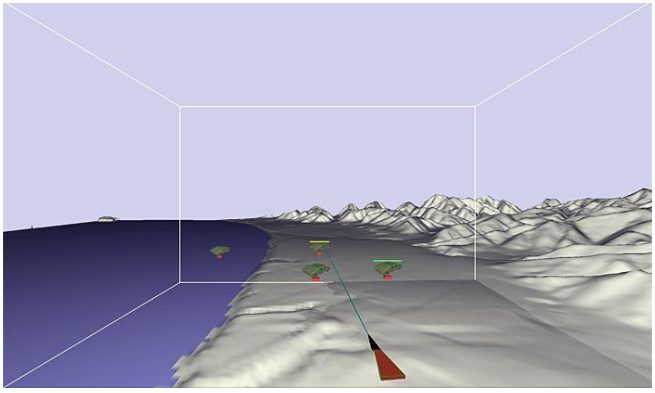


Figure 7. Sample Scenario Execution State in FreeVR CAVE simulator

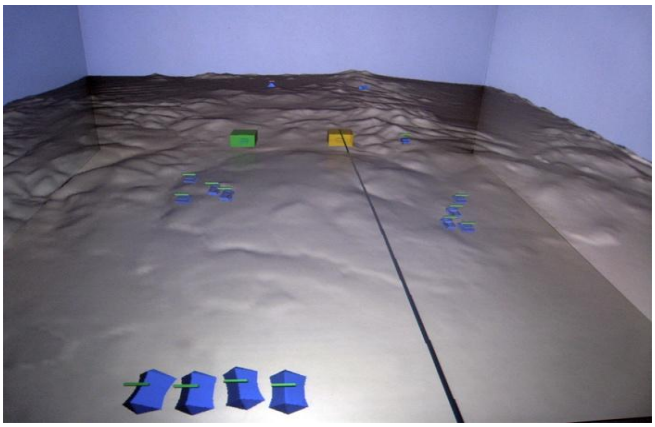


Figure 8. Sample Scenario Execution State in a 4-Wall CAVE

The focus of this paper has been on presenting the process and tools provided by CAVEMANDER to design and develop applications for the CAVE. These advances are significant, showing (amongst other points) that a novice user can visually design and build software artifacts that take an essential role in the construction phase of a CAVE-based system. In this way, CAVEMANDER takes a notable step toward providing non-expert computer programmers the ability to build simulation scenarios that run in the CAVE, thus allowing them to conduct VR immersive training sessions and research studies with reduced effort.

ACKNOWLEDGMENT

This work was made possible through the support provided by NASA grant #NNX07AT65A via a sub-award and with cost share provided by the Nevada System of Higher Education: NSHE-08-51 and NSHE-08-52.

REFERENCES

[1] Buntha, S. CAVEMANDER: An Approach and Software Platform for Building Command-and-control Applications in CAVE, PhD Dissertation, University of Nevada, Reno, 2009.

[2] Nacenta, M. A., Sakurai, S., Yamaguchi, T., Miki, Y., Itoh, Y., Kitamura, Y., Subramanian, S., and Gutwin, C. E-conic: A Perspective-Aware Interface for Multi-Display Environments. In Procs of the 20th Annual ACM Symp. on User Interface Software and Technology, Newport, RI, USA, 2007, ACM, pp. 279-288.

[3] Funke, G. J., and Galster, S. M., The effects of Spatial Processing Load and Collaboration Rechnology on Team Performance in a Simulated C2 Environment. In ECCE '07: Proceedings of the 14th ACM European Conference on Cognitive Ergonomics New York, NY, USA, 2007, ACM, pp. 37-43.

[4] Dudfield, H., Macklin, C., Fearnley, R., Simpson, A., and Hall, P., Big is Better? Human Factors Issues of Large Screen Displays with Military Command Teams. In Proceedings of The Second IEEE International Conference on People in Control Human Interfaces in Control Rooms, Cockpits and Command Centres, 2001, pp. 304-309.

[5] Mancero, G., Wong, W., and Amaldi, P., Looking but Not Seeing: Implications for HCI. In ECCE '07: Procs. of the 14th European Conf. on Cognitive Ergonomics, NY, USA, 2007, ACM, pp. 167-174.

[6] Sowndararajan, A., Wang, R., and Bowman, D. A., Quantifying the Benefits of Immersion for Procedural Training. In IPT/EDT '08: Procs. of the 2008 Workshop on Immersive Projection Technologies/ Emerging Display Technologiges, NY, USA, 2008, ACM, pp. 1-4.

[7] Schuchardt, P., and Bowman, D. A. The benefits of Immersion for Spatial Understanding of Complex Underground CAVE Systems. In VRST '07: Procs. of the 2007 ACM Symposium on Virtual Reality Software & Technology, New York, USA, 2007, ACM, pp. 121-124.

[8] Mechdyne Corporation. CAVE Software, April 2010. Accessed at http://www.mechdyne.com/integratedSolutions/software/products/CAVELib/CAVELib.htm.

[9] Sherman, W. R. FreeVR Homepage, March 2010. Accessed at http://www.freevr.org/.

[10] VR Juggler. The VR Juggler Suite, April 2010. Accessed at www.vrjuggler.org.

[11] Open Tech, Inc. CoVE - An Open Source Virtual Reality Toolkit, February 2010. Accessed at http://cove.opentechinc.com.

[12] Act-3D B.V. Quest 3D: Visual 3D Development Software, April 2010. Accessed at http://www.quest3d.com.

[13] Eilemann, S., Makhinya, M., and Stalder, C. Equalizer: Parallel Rendering, March 2010. Accessed at www.equalizergraphics.com.

[14] Avango. Avango, April 2010. Accessed at http://www.avango.org/.

[15] Jacobson, J., Renard, M. L., Lugrin, J.-L., and Cavazza, M., The CaveUT System: Immersive Entertainment Based on a Game Engine. In ACE '05: Procs. of the 2005 ACM SIGCHI Intl. Conf. on Advances in Comp. Entertainment Technology, 2005, pp. 184-187.

[16] Kreylos, O., and Billen, M. I. 3D-Visualizer: Interactive Gridded Data Volume Visualization Software, March 2010. Accessed at http://keckcaves.org/software/VISUALIZERCG/index.html.

[17] Buntha S., Muhanna, M., Dascalu, S., Harris, F. C., and Okamoto, S., A GUI Wizard for Developing Command-and-Control Applications in CAVE, Procs. of the 4th IASTED Intl. Conf. on Human-Computer Interaction, St. Thomas, US Virgin Islands, USA, 2009, pp. 301-308.

[18] Arlow, J., and Neustadt, I., UML and the Unified Process: Practical Object-Oriented Analysis and Design, 2nd Edition, Addison-Wesley, Boston, MA,USA, 2005.

[19] Sommerville, I., Software Engineering, 8th Edition, Addison-Wesley, Boston, MA, USA, 2006.

Recognizing Sketches of Euler Diagrams Drawn with Ellipses

Aidan Delaney[1] Beryl Plimmer[2] Gem Stapleton[1] Peter Rodgers[3]
[1]Visual Modelling Group, University of Brighton, UK.
{a.j.delaney,g.e.stapleton}@brighton.ac.uk
[2]University of Auckland, New Zealand. beryl@cs.auckland.ac.nz
[3]University of Kent, UK. p.j.rodgers@kent.ac.uk

Abstract

Euler diagrams form the basis of a number of visual languages. However, the existing tool support for creating Euler diagrams is limited to generic diagram editing software that uses mouse and keyboard interfaces. A more natural and convenient mode of entry is via a sketching interface. In addition, it is known that sketching, as opposed to using an editing package, facilitates greater cognitive focus on the task of diagram creation. This paper presents the first sketch tool for Euler diagrams. In particular, we describe the recognition mechanism and conversion that takes a sketched Euler diagram and converts it into a formal diagram drawn with ellipses and circles. The semantics of the sketch is computed.

1. Introduction

Euler diagrams are a popular and widely used tool for information visualization. In addition, they form a component of many visual languages, such as spider diagrams [8], Euler/Venn diagrams [21], Venn-II diagrams [19], constraint diagrams [12], and ontology diagrams [15]. Given their wide-ranging practical use, there is a need to provide convenient ways of creating these diagrams in electronic form. Moreover, the software in which they are created will, ideally, have some understanding of the diagram semantics, so it can further support the user in exploring information conveyed by the diagram. Currently, however, the manual creation of Euler diagrams in electronic form relies on unintuitive mouse and keyboard interfaces in software that has no semantic understanding of the diagram.

A natural creation method for Euler diagrams is using a pen, but no intelligent tool support exists for this mode of entry. The lack of sketching support for Euler diagrams means that, in the vast majority of cases, when using a computer they must be drawn in off-the-shelf diagram editing tools. The current editing support does not provide a natural and convenient interface because the point-by-point specification of the diagrams' curves is slow (compared to sketching) and the act of sketching is simple in comparison to using a diagram editor, such as those found in Microsoft's Word or Visio packages. Sketching the diagram has the further advantage that it allows the user to focus on the actual diagram creation rather than the interface of the editing tools.

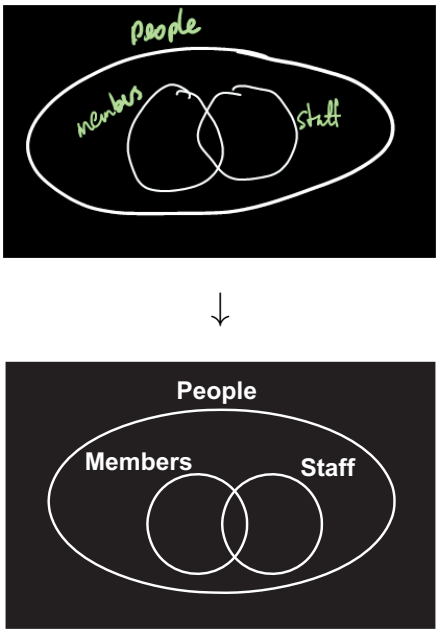


Figure 1. A sketched Euler diagram and its formal visualization.

The sketch recognition software developed to date has focused on user interface design and graph oriented diagrams [11]. With respect to user interface design tools the sketched items are largely independent of each other. In graph oriented diagrams the spatial positioning of nodes and edges are not of semantic significance. By contrast, in Euler

diagrams the spatial relationships between sketched items is fundamental to their semantics. To our knowledge, the work described in this paper is the first to consider these types of complex spatial relationship for sketch recognition.

This paper describes a mechanism for recognizing sketches of Euler diagrams drawn with ellipses. This work forms a basis for developing more sophisticated sketch recognition tools for Euler diagrams and, hence, for the many notations that extend them. To illustrate the process, a user-sketched Euler diagram can be seen in the top of figure 1 whereas a diagram drawn in an editing tool is placed underneath; we call this lower diagram a *formal* diagram. The sketch tool that we have developed can automatically produce a formal diagram from the user sketch. Our mechanism relies on, firstly, dividing the syntax into two classes: handwritten labels and hand-drawn curves. The curves are further classified as ellipses or circles. Syntax identified as a label is subsequently passed to a standard handwriting recognition package. Syntax identified as a curve is converted into a formal representation by finding a suitable approximating circle or ellipse, as appropriate.

Section 2 provides background material on Euler diagrams and sketch recognition, thus providing motivation for the research and contextual information. Section 3 describes our Euler diagram sketch recognition techniques, including our method for deriving the diagram semantics. Finally, we conclude in section 4, where we also discuss further directions for this research.

2. Context and Motivation

Euler diagrams [4] are collections of closed curves used to visualize relationships between sets and generalize Venn diagrams [23]. Euler diagrams are a common component of visual languages because they intuitively represent exclusion, containment and intersection of sets. An example can be seen in figure 2, which shows a categorization of countries in the British Isles. The varied application areas of Euler diagrams include, but are not limited to: crime control [5], computer file organization [3], classification systems [22], education [10], genetics [13], and medicine [20]. Thus, there is widespread interest in these types of diagrams, but under-developed tools to support their creation. Providing sketch tools for Euler diagrams has the potential to be beneficial to a large community of users.

Sketching a diagram on a piece of paper is a useful problem solving and communications technique [7]. In a sketching context, users can produce, evaluate, modify, refine and replace diagram components rapidly. The backtalk from the external representation of the diagram is considered an essential part of the cognitive support for design. If a computer 'understands' the sketch it can be more easily edited, animated and translated from one format to another. Pen-input computers mean that the act of drawing can now be performed directly on the computer screen or tablet. Automatic and reliable recognition of hand-drawn diagrams is now becoming possible.

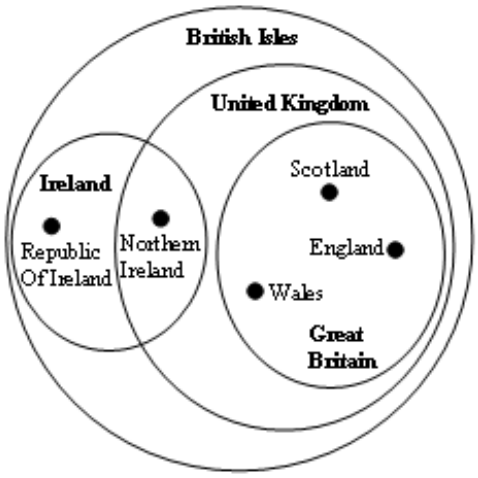


Figure 2. The British Isles, by Sam Hughes [9].

Hand-drawing diagrams has been shown to be more effective for external representation of a problem than using formal computer diagramming tools [6]. However to retain these benefits the computer-based sketch tools interaction must be carefully designed with particular attention to retaining the hand-drawn appearance of the diagram [18] in the initial stages of the diagram's creation. Underlying such a sketching interface there must be a sophisticated and accurate recognition engine. While much of the early work was on recognition engines for specific diagrams, for example [14], there are now a number of configurable or trainable recognition engines [16, 17].

However, there are no recognition engines specifically designed for diagrams where the spatial relationship between the syntactic components is of primary significance. For instance, closed curves (typically circles) represent nodes in graph-like diagrams and their spatial relationship is often taken to have no semantic relevance whereas closed curves in Euler diagrams overlap or enclose one another precisely to convey semantically relevant information: overlapping and containment between curves is unique and critical to Euler diagrams.

3. Recognizing Euler Diagrams

There are various stages to our recognition process which we have implemented in a prototype software tool.

This tool is an extension of the existing SketchNode codebase [18]. Given a user created sketch, the first stage is to divide the sketched components into two categories: labels and curves. We derive the semantics of the sketch by computing its abstract syntax. Independently, we convert the sketched curves into formal curves; in this paper we restrict the curves to being to ellipses, with circles as a special case of an ellipse. The curve labels are processed by a handwriting recognition engine.

3.1 Dividing the Syntax

An example of a sketched Euler diagram, drawn in our software tool, can be seen in figure 3. The first phase of the recognition process is to classify each syntactic component as a curve label or as a curve; curves are further categorized as circles or ellipses. We use AI techniques to derive

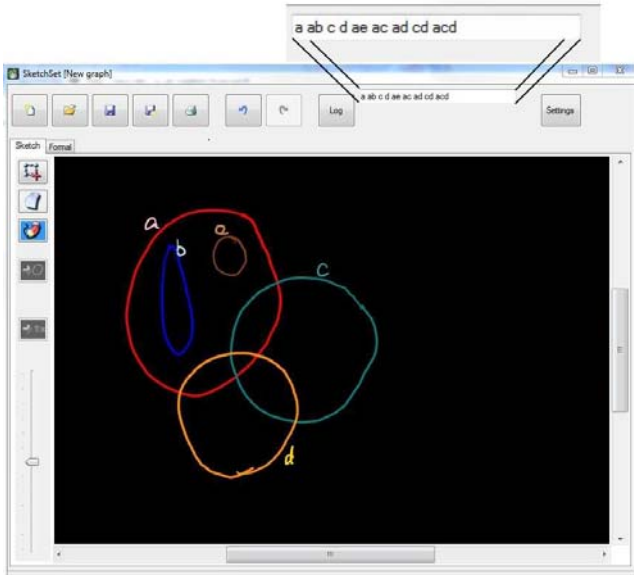


Figure 3. A sketched Euler diagram.

a mechanism for the classification. The recognizer component, Rata.SSR [2], was trained from a number of sample diagrams; we chose 10 Euler diagrams, each with up to 4 curves. The curve labels comprised the letters of the alphabet. We asked 10 participants to sketch each of the 10 Euler diagrams and label their curves appropriately. This provided us with a training set of 100 sketches. The sample diagrams were collected in DataCollector, part of DataManager [1].

To illustrate, figure 4 shows a screenshot of a participant's sketch. The righthand panel shows the description of the task. Here, two descriptions of Euler diagrams are given and the participant is asked to draw them. The top sketch was drawn from the Euler diagram description f, fg, fh which identifies the regions to be present in the diagram; f is a region inside just a curve labelled f, fg is inside just curves labelled f and g and so forth. The participants who were not familiar with this manner of describing an Euler diagram were provided with training.

After collecting the 100 sketches, we labelled each digital ink stroke as: text, curve-circle or curve-ellipse. From the sketches, DataManager generates feature vectors from each ink stroke and uses these to train the Rata.SSR generator. The Rata.SSR is then consumed within our software tool as a component. It separates labels from curves and recognizes curves as circles or ellipses. Labels are passed to the operating system text recognizer while the curves are used to generate the equivalent formal visualization; we discuss this latter aspect in section 3.3.

3.2 Abstract Syntax of the Sketch

A sketch tool needs semantic understanding of the diagram in order to support intelligent interaction such as editing and to generate other representations. Euler diagrams represent sets using curves and the diagram's *zones* completely determine the relationships between the sets. That is, the zones present correspond to the semantics. A zone is a region that can be described as being inside some curves but outside the rest of the curves. For example, the sketch in figure 3 has zones described by a, ab, c, d, ae, ac, ad, cd, acd, as shown in the callout. This list of zone descriptors is called the *abstract description*.

In order to compute the abstract description, we first assign each curve label its closest curve, unless that curve is already assigned a label. The process of label assignment is done on-line, whilst the user is creating the sketch, and adjustments are made as the user performs edits to the diagram. For instance, if a curve is deleted but its associate label remains then that label may be reassigned to the next closest curve, or assigned to no curve if every sketched curve already has a label. Each curve is automatically assigned a colour that is used in both the sketch and formal views. There is an internal list of 8 colours for curves; the colours are automatically assigned in sequence as curves are recognised. Curve labels are assigned a lighter shaded of the colour associated to their curve. This allows the user to readily check the association made by the software and, thus, have an opportunity to change the sketch if necessary.

There are three major steps when building the abstract description of a sketched Euler diagram: first, all the separate zones in the diagram must be identified; second, we establish the containment relationships between zones; lastly, we build the abstract description by iterating through all zones to establish their label.

The zones are created by first identifying intersections between each pair of (real) curves, c_1 and c_2. When an in-

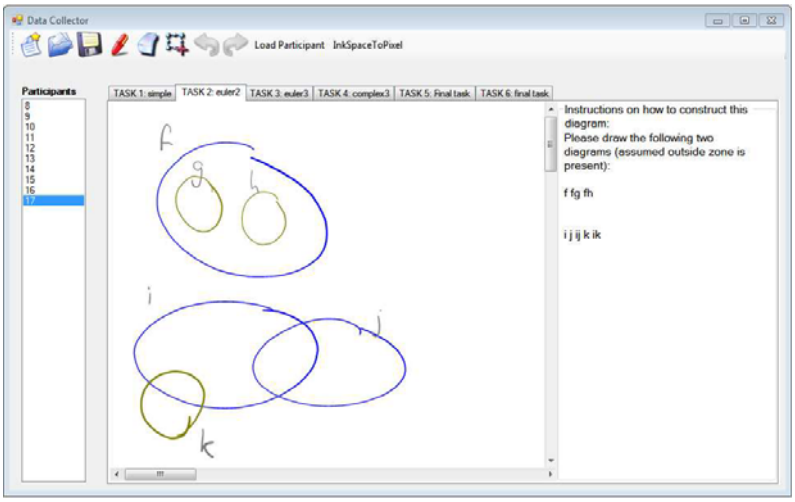


Figure 4. A sketch from the training set.

tersection is found between c_1 and c_2, a virtual curve, c_3, is created with a sequence of points made from the segments of the real curves that form the boundary of the intersection space. Information on the parent curves, c_1 and c_2, is held in the virtual curve, c_3, for later use. In figure 3, this results in virtual curves ac, ad and cd (three regions formed from the pairwise intersecting curves). The algorithm then checks the set of virtual curves looking for intersections. On the next pass, the zone acd is discovered, for example. If more than one new virtual curve is created the algorithm calls itself with the new virtual curves looking for further intersections.

We also need to identify when one curve, virtual or real, contains another. To establish containment, the smallest (if any) container for each curve is found. Each curve (real and virtual) is checked against all other non-parent curves (a virtual curve will always fall within its parents' bounding boxes). When the bounding box of a curve lies within the bounding box of another curve then the other curve is a candidate container. If more than one candidate container is found then the smallest is the container the outer containers can be found by check each containers container.

The abstract description can then generated by building the label for each curve. A label for a real curve consists of the label of its container (which may itself have a container) and its own label. The label for a virtual curve is the set of the labels of the component curves To allow the abstract description to be constructed when a curve is unlabelled the curve's colour is used as shown in the callout in figure 5.

3.3 Converting to the Formal Diagram

We now describe the process by which we convert a sketch drawn curve into a formal curve. First, we describe

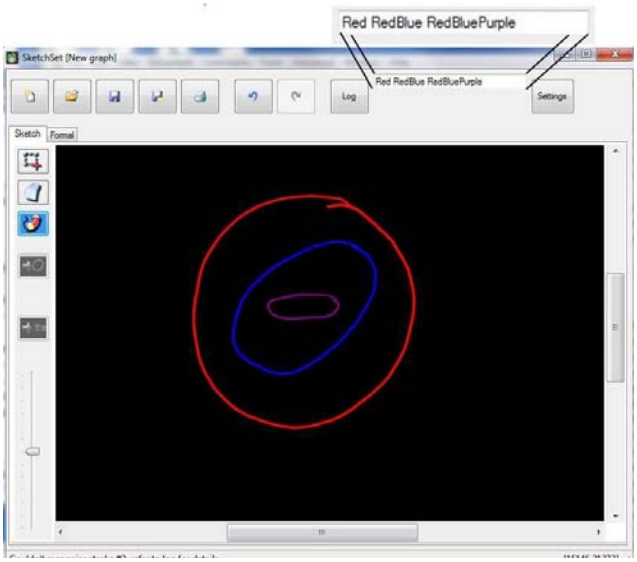


Figure 5. A sketch with no curve labels.

how to convert a sketched curve that has been classified as an ellipse, but not a circle, into a formal ellipse. A sketched curve, see the left of figure 6, is represented internally by a sequence of points, as shown on the right. We find the longest chord across the ellipse by taking the two points that are furthest apart in the internal representation and we then draw a line segment between them. We can calculate the rotation of the sketched ellipse from the angle, α, formed by this chord and the x-axis, as shown in figure 7.

Next, we rotate the sketched ellipse by α so that this chord is parallel to the x-axis. A standard algorithm is then used to find a bounding box of the sketched ellipse, as shown in figure 8. This rectangle is used to create the

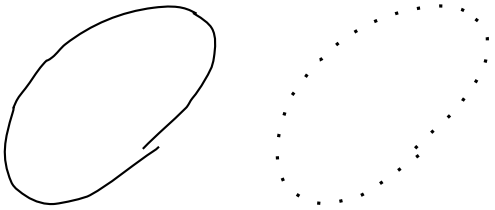


Figure 6. A sketched ellipse and its internal representation.

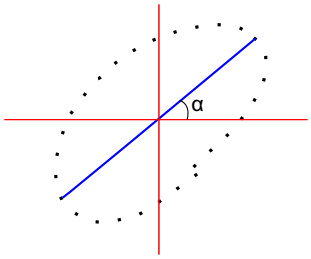


Figure 7. Computing the angle of rotation.

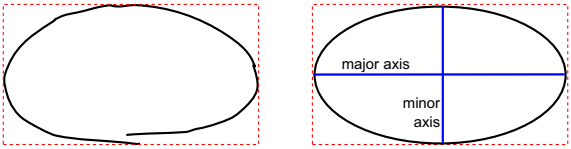


Figure 8. Finding a bounding box to compute the minor axis.

formal ellipse: the centre of the ellipse is the centre of the rectangle, the major axis takes the width of the rectangle, and the minor axis is the height of the rectangle. Finally, the formal ellipse has rotation α, the same angle of rotation as the sketched version, shown in figure 7.

However, we attempt to preserve user intent in this last step, with regard to approximate alignment with the x and y-axis. If α is within 10 degrees of 0, 90, 180, or 270 degrees then we round α accordingly to achieve alignment.

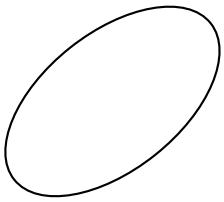


Figure 9. The final formal ellipse.

In the case of a curve classified as a circle by Rata.SSR, we follow a similar procedure, except that angle of rotation is irrelevant. Once we have computed the bounding box, we take the average length of the two sides as the circle's diameter. Of course, it is not necessary to differentiate between circles and ellipses, but circles are more aesthetically pleasing; if a user intended to draw a circle it is preferable to preserve that intention where possible.

Figure 10 shows the formal Euler diagram obtained from the sketch in figure 3. The sketched curves c, d and e have been classified by Rata.SSR as circles and, therefore, rendered as circles in the formal diagram, whereas a and b have been identified as ellipses. Notice that b, in the sketch, is approximately parallel to the y-axis and has, therefore, been given a rotation of 90 degrees in the formal diagram. The computed angle of rotation of a in the sketch has, however, has been maintained in the formal diagram. We plan to extend the software so that it computes the abstract description of the formal diagram, which can be done using the algorithm described for the sketch. This will allow us to automatically compare the abstract description of the sketch and the formal diagram to highlight any differences that may have arisen during the formalization stage.

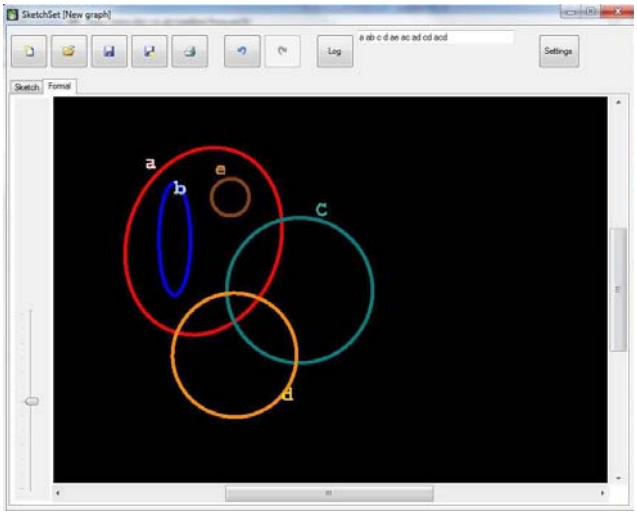


Figure 10. A formal Euler diagram generated from figure 3.

4. Conclusion

In this paper, we have presented the first techniques for recognizing sketches of Euler diagrams and developed a prototype implemtation. We have restricted to the case when all curves are circles or ellipses, which is often seen to be the case in user created Euler diagrams. However, there are collections of sets that cannot be accurately visualized using an Euler diagram drawn with ellipses and more arbitrary shaped curves must be used. We plan to extend this work to allow the recognition of arbitrary curves.

In order to produce quality diagrams from sketches drawn with arbitrary curves there are many challenges to be overcome. For instance, when a sketched curve has some symmetry, we should aim to preserve that symmetry in the formal diagram. In addition, we need to take into account approximate alignment and equality of spacing between sketched curves and preserve the perceived user intention in the formal diagram. The software should also incorporate more sophisticated editing functionality which the user can access in order to make modifications to both the sketch and formal diagram as required. We plan to write an Euler diagram sketch recognition plug-in for InkScape, so we can take advantage of the features already implemented in this freely available diagram editing tool.

The results presented here provide the necessary basis for developing sketch recognition software for notations that extend Euler diagrams with additional syntax. For example, constraint diagrams augment Euler diagrams with graphs, arrows, and shading. A simple example can be seen in figure 11; in this example, the curves of the underlying Euler diagrams do not intersect with each other but more complex diagrams can display arbitrary intersections between the curves. In constraint diagrams, the different style of node (asterisks versus filled circles) in the embedded graphs is of semantic importance and a sketch recognition engine for constraint diagrams would need to be sensitive to this difference.

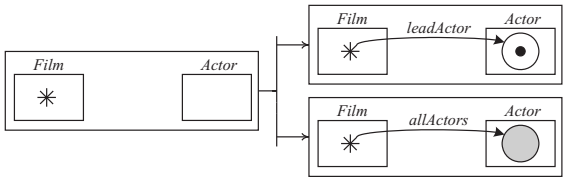


Figure 11. A constraint diagram.

Acknowledgement This research is supported by EPSRC grants EP/E010393/1, EP/H012311/1 and EP/H048480/1 and a Royal Society of New Zealand Marsden Grant.

References

[1] R. Blagojevic, P. Schmieder, and B. Plimmer. Towards a toolkit for the development and evaluation of sketch recognition techniques. In *Sketch Recognition Workshop*, 2009.

[2] S. Chang, B. Plimmer, and R. Blagojevic. Rata.ssr: Data mining for pertinent stroke recognizers. In *Sketch Based Interface Modeling*. ACM, 2010.

[3] R. DeChiara, U. Erra, and V. Scarano. VennFS: A Venn diagram file manager. In *Proceedings of Information Visualisation*, pages 120–126. IEEE Computer Society, 2003.

[4] L. Euler. Lettres à une Princesse d'Allemagne sur divers sujets de physique et de philosophie. *Letters*, 2:102–108, 1775.

[5] G. Farrell and W. Sousa. Repeat victimization and hot spots: The overlap and its implication for crime control and problem-oriented policing. *Crime Prevention Studies*, 12:221–240, 2001.

[6] V. Goel. *Sketches of thought*. MIT Press, 1995.

[7] G. Goldschmidt. *Visual and Spatial Reasoning in Design*, chapter The Backtalk of Self-Generated Sketches, pages 163–184. University of Sydney, 1999.

[8] J. Howse, F. Molina, J. Taylor, S. Kent, and J. Gil. Spider diagrams: A diagrammatic reasoning system. *Journal of Visual Languages and Computing*, 12(3):299–324, June 2001.

[9] S. Hughes. The Great British Venn diagram. http://qntm.org/uk, accessed July 30, 2010.

[10] E. Ip. Visualizing multiple regression. *Journal of Statistics Education*, 9(1), 2001.

[11] G. Johnson, M. Gross, and J. Hong. *Computational Support for Sketching in Design*. Now Publisher Inc., 2009.

[12] S. Kent. Constraint diagrams: Visualizing invariants in object oriented modelling. In *Proceedings of OOPSLA97*, pages 327–341. ACM Press, October 1997.

[13] H. Kestler, A. Muller, H. Liu, D. Kane, B. Zeeberg, and J. Weinstein. Euler diagrams for visualizing annotated gene expression data. In *Proceedings of Euler Diagrams 2005*, Paris, September 2005.

[14] J. Landay and B. Myers. Interactive sketching for the early stages of user interface design. In *Chi 1995 Mosaic of Creativity*, pages 43–50, 1995.

[15] I. Oliver, J. Howse, G. Stapleton, E. Nuutila, and S. Törma. A proposed diagrammatic logic for ontology specification and visualization. In *International Semantic Web Conference*, 2009.

[16] B. Paulson and T. Hammond. Paleosketch: Accurate primitive sketch recognition and beautification. In *Intelligent User Interfaces*. ACM Press, 2008.

[17] B. Plimmer and I. Freeman. A toolkit approach to sketched diagram recognition. In *HCI*, pages 205–213. British Computer Society, 2007.

[18] B. Plimmer, H. Purchase, and H. Laycock. Preserving the hand-drawn appearance of graphs. In *Visual Languages and Computing*, pages 347–352, 2009.

[19] S.-J. Shin. *The Logical Status of Diagrams*. Cambridge University Press, 1994.

[20] J. Soriano, K. D. B. Coleman, G. Visick, D. Mannino, and N. Pride. The proportional Venn diagram of obstructive lung disease. *Chest*, 124:474–481, 2003.

[21] N. Swoboda and G. Allwein. Heterogeneous reasoning with Euler/Venn diagrams containing named constants and FOL. In *Proceedings of Euler Diagrams 2004*, volume 134 of *ENTCS*. Elsevier Science, 2005.

[22] J. Thièvre, M. Viaud, and A. Verroust-Blondet. Using Euler diagrams in traditional library environments. In *Euler Diagrams 2004*, volume 134 of *ENTCS*, pages 189–202. ENTCS, 2005.

[23] J. Venn. On the diagrammatic and mechanical representation of propositions and reasonings. *Phil.Mag*, 1880.

Automated Clock Drawing Test through Machine Learning and Geometric Analysis

Anupam Guha, Hyungsin Kim, Ellen Yi-Luen Do
College of Computing and GVU Center
Georgia Institute of Technology
Atlanta, Georgia
{aguha7, hyungsin, ellendo}@gatech.edu

***Abstract*—In this paper, we discuss the challenges of sketch recognition accuracy and automation of the Clock Drawing Test (CDT). Sketch recognition in the context of the CDT is a complex problem due to the lack of knowledge of the preference bias among the sketches drawn by neuro-atypical patients. However, machine learning provides a viable solution to detect measurable patterns among sketches drawn in the CDT. The paper posits that these sketches are constrained, and the majority of the recognition work can be completed by doing handwritten digit recognition. All other properties to be measured in the CDT are primarily geometric in nature. The paper explores various machine learning classification techniques for handwritten digit recognition. Finally, based on analysis, we present the best one in context of CDT.**

Keywords: Clock Drawing Test, Machine Learning, Geometric Analysis, pattern recognition

INTRODUCTION

The Clock Drawing Test (CDT) is an instrument to screen people with cognitive impairment and dementia [1]. Generally, CDT is administered manually in a hospital environment in which the test subject is instructed to draw the face of a clock, and the hands indicating a specific time (e.g., 10 past 11) [2]. A variation of CDT asks the subject to copy a drawn figure of a clock. The difficulties in developing automated CDT software lies in the amount of variances in the predicted results due to the neuro-atypical nature of the patients involved. Not only is it difficult to predict with any degree of accuracy how much the errors on parts of the subject alter the sketches to be recognized from the expected outcomes but also it is hard to distinguish situational drawing errors from genuine symptoms of cognitive debility.

In this paper, we present a system that attempts to address these challenges and automates some if not all of the processes that occur in the conduction of the CDT. It becomes somewhat easier to address the problem if it is broken down into subparts. Each has multiple potential solutions and then we use build and fix to test each one of them. The computational model of the CDT has several exploitable constraints which shortens our task. First we chose a rubric for scoring once the clock face has been drawn. The system has to generate a score out of 13, each point awarded for a factor drawn correctly. These factors are:

1. Only numbers 1-12 are present without additions or omissions
2. Only Arabic numerals are used
3. Numbers are in the correct order.
4. Numbers are in approximately correct positions
5. Numbers are inside the circle
6. Only two hands are present
7. The hour target number is indicated
8. The minute target number is indicated
9. The hands are in correct position
10. There are no superfluous markings
11. The hands are relatively joined
12. A centre of the circle is present or inferred at the joining of the hands

Looking at this scoring rubric one fact emerges. Recognition of handwritten digits is the key. If all the digits are correctly identified it becomes easier to process the other sketches, like that of the hands, where they are pointing to, etc.

HANDWRITTEN DIGIT RECOGNITION

We posit that handwritten digit recognition provides a solution to the automation of CDT. For one it is a fairly robust and tested paradigm [4], which is being currently developed for various other applications [3] using a lot of different techniques. If the digits are determined to be all present and accounted for (on conversely unidentifiable) then more than half of the work of sketch recognition in CDT is over. Also even if the recognition is not completely accurate it does not alter the score much. All that needs to be decided is that most of the sketches are Arabic digits and their locations. The first task of the system is to identify the discrete sketches drawn in the clock-face. Counter intuitively this proves to be a non-trivial task because the patient undergoing the test may not draw every digit neatly away from each other (See Figure 1). Often the patient would join adjoining sketches like clock hands with digits. While this is a small error from the human point of view and presents no problem to the medical examiner it may play havoc with an Machine Learning (ML) system. Conversely the patient may choose to draw the

sketches of the digits in such a manner that the digits themselves are drawn with separations which ought not to occur (Figure 2).

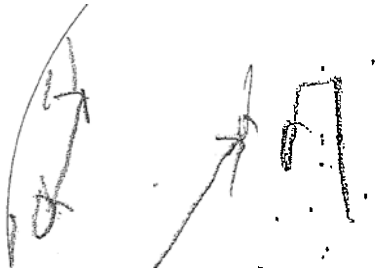


Figure 1. Cases of patients smudging two discrete sketches

The data collected for analysis of such anomalies was taken from a set of 50 CDTs obtained from the tests conducted by the Emory's Alzheimer Research Center on a random sample set patients suffering from mild cognitive impairment due to aging. The results were:

1. Three cases of sketches being smudged together – for example, hands overlapping or connecting the digits as shown in Figure 1.
2. Eight cases of sketches being discrete without cause – for example, digits drawn with space between different strokes in Figure 2



Figure 2. Cases of patients breaking a continuous sketch

Considering that 50 CDTs yield over 700 discrete sketches these are fairly low numbers. Patients undergoing treatment for cognitive impairment somehow are able to maintain digit integrity. Analyzing the CDTs we came to a conclusion that sketch integrity is fairly stable across a spectrum of patients and there is a justifiable case for doing machine learning classification techniques on them after some elementary noise cleaning of the CDT images.

The second test we conducted was to identify which machine learning classification algorithm worked best when the data to be classified was handwritten digits made by neuro-atypical patients. The goal was not to test for algorithms which are considered state of the art for handwritten digit recognition but to use algorithms which can be used easily to determine the suitability of other algorithms based on the way they work and which are relatively fast ('lazy learning' like kNN) as we wanted to see what, if any, differences are there from 'normal' digits drawn by neurotypical patients besides extracting features which were 'interesting'. It was quite possible that the handwritten digit data here would be sufficiently different from that of a standard dataset and it was interesting to find out if so then where. It was not expected that these general machine learning algorithms will perform as well as those specifically designed for handwritten digit recognition but absolute accuracy was not the goal here, rather comparisons and identifying useful features to be extracted was. The results obtained are different from those obtained by standard datasets like MNIST (http://yann.lecun.com/exdb/mnist/). We have a dataset of handwritten digits that have already been identified for the purpose of recognizing the sketches. The dataset was obtained from the MNIST dataset of which the first 5000 instances were taken as training data. The algorithms used were kNN, Multilayer Perceptron and Decision Trees.

Feature Extraction

To test for which algorithm would prove to be the best, the test set was created out of the digits obtained from the 50 datasets. Each was resized into a 16 x 16 pixel binary image. If all the pixels were to be taken as features then each sketch was represented as an array of 256 '0's and '1's. Also a second dataset was created by taking these images, and taking the horizontal, vertical, right diagonal and left diagonal histograms. The histograms will be arrays of 16, 16, 31, and 31 integers, concatenating them would give an array of 94 integers which were used to represent the sketches.

Table 1: Decision Trees Digit-Raw

Pruning	Train%	Test%
---	100%	59.5%
minimum leaf instance=2 stop splitting nodes with less instances than=2	90.3%	58.6%
minimum leaf instance=4 stop splitting nodes with less instances than=4	80.4%	54.8%
post pruning m=2	79.9%	56.1%

Table 1 displays the results for decision trees on the raw data with no feature extraction. The results are not very encouraging and even a little amount of 'pruning' (lossy compression of the trees) results in further degradation. When the same algorithms have been used for 'normal samples' from the MNIST dataset with the same algorithms for the same settings the accuracy has crossed 95% in kNN as well as neural networks for digit-hist and 90% for digit raw and a little above 72% for unpruned decision trees for digit hist and 65% for digit raw. Accuracy increases in Digit-Hist (Table 2) which uses histograms as features compared to Digit-Raw. We believe this is due to the fact aggregation of attributes rather than taking them individually smoothens out the errors and decreases 'overfitting' (which means that the algorithm trusts faulty data too much and takes into account each little deviation in the graph and misclassifies when test data doesn't follow those deviations exactly). We believe some of the damage done by overfitting is slightly improved by post pruning as post pruning smoothens out the small deviations.

This behavior is consistent on both Digit datasets. It is evident that decision trees are not a viable option for such noisy data.

Table 2: Decision TreesDigit-Hist

Pruning	Train%	Test%
---	99.98%	63.1%
minimum leaf instance=2 stop splitting nodes with less instances than=2	89.1%	62.4%
minimum leaf instance=4 stop splitting nodes with less instances than=4	78.6%	61%
post pruning m=2	78.1%	61.4%

Table 3 and 4 display the results for kNN. A drastic improve in accuracy is observed. This has to do with the properties of the datasets and the way kNN works. kNN is extremely suitable for handwritten digit recognition as we are taking into consideration all attributes simultaneously, and errors are more likely to get smoothened out. 'Lazy' machine learning (machine learning simply involving storing and matching like kNN, no 'training' involved) makes a lot of sense for classification where there is a pretty straightforward feature extraction from images. It takes a lot less time than, say, neural networks while giving comparable accuracy. However there is not much scope for improvement, once a maximum has been reached.

Table 3: k-Nearest NeighboursDigit-Raw

Nearest Neighbours	Train%	Test%
1	99.9	85.6
5	99.9	89.4
10	99.9	85.8
20	99.9	85.8
50	99.9	84.9

Table 4: k-Nearest NeighboursDigit-Hist

Nearest Neighbours	Train%	Test%
1	99.9	88.7
5	99.9	89.1
10	99.9	89.6
20	99.9	89.5
50	99.9	88.4

Table 5 displays the results of the perceptrons on the two datasets. We trained the neural networks for different number of epochs. A very surprising result is observable. In the two digit datasets, doubling the epochs lead to a slight decrease in accuracy. This may be due to overfitting. Also unlike the last two algorithms in neural network, Digit-Raw would actually give a better result than Digit-Hist. The reason is simple, while in other classification algorithms, presence of too many attributes becomes a hindrance at times, and prevents smoothening over small deviations, in neural networks; it means a large input network, which is almost always good news. Also the input network in Digit-Raw is binary, whereas that in Digit-Hist is numerical. The same is observed with the 'normal digits' but with accuracy crossing 95%. Evidently there is some degree of 'sameness' inside the MNIST digits which is fundamentally different from the digits obtained from the CDT. Whether this difference is the result of the neuroatypiccal patients drawing the digits or whether this is due to the test not done in conditions the MNIST digits were collected is a discussion beyond this paper. We observe that the digits are different; why so is something we will not speculate on.

Table 5: Multilayer Perceptron

Digit-Raw	
256 input layer, 50 hidden layer, 10 output layer	
500 epochs	
Correctly Classified Instances	87.1 %
1000 epochs	
Correctly Classified Instances	86.8 %
1500 epochs	
Correctly Classified Instances	86.5 %
2000 epochs	
Correctly Classified Instances	86.4 %
Digit-Hist	
94 input layer, 20 hidden layer, 10 output layer	
500 epochs	
Correctly Classified Instances	84.5 %
1000 epochs	
Correctly Classified Instances	84.3 %
1500 epochs	
Correctly Classified Instances	84.1 %
2000 epochs	
Correctly Classified Instances	84 %

As can be observed it proved that the decision tree algorithm was not adequate at all in identifying accurately the sketches made by the patients in the CDT. Though the perceptrons proved most accurate they were incredibly slow. We posit that for the wants of the CDT the kNN algorithm suffices as it is sufficiently accurate as well as quite fast. These results advocate the necessity of large scale collection of digits drawn by neuroatypical patients and comparing them with the 'standard' datasets to observe repeatable patterns of differences. As our work here was to develop a system which does automated CDT, The best algorithm which is reasonably fast and can be used for the CDT system is kNN with 10 neighbors using the histograms as the features.

COMPONENTS OF THE SYSTEM

The first order of business of the system under development is to extract the individual blobs. Each blob is a sketch segment from the drawing of the clock face. The system uses MATLAB® as its platform. It uses the connected component

functionality to discover these sketches. The idea is that each sketch component is a sequential connected group of pixels removed from any other sketch segment. Thus extraction of each 'blob' as a separate image is possible where each is a separate sketch. In this manner the entire drawing is broken up into fragments, each is resized and cleaned. The larger sketches are hypothesized to be the hands and stored separately. 'Larger' is defined as the sketches whose major axis is longer by 100% of the average length.

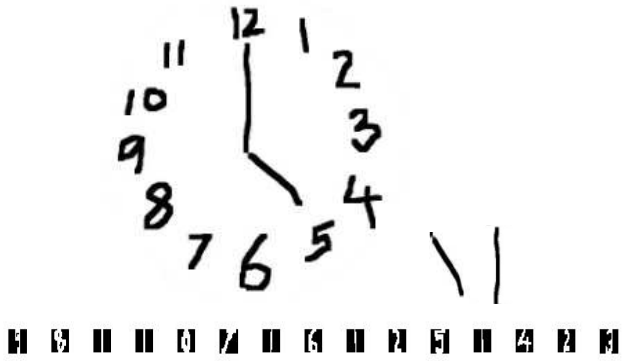


Figure 3. Clock face being broken up into discrete components like hands and digits

After this is done those digits are identified which are '1' using kNN. The '1's are important as there are five '1's. Then each digit is identified. The '1's which are in close proximity with a '0', a '1' and a '2' ('proximity' defined as Euclidean distance between centroids being 50% lesser than average) are designated as 10, 11 and 12. Now the system knows whether there all the 12 numbers are present or not and where their centroids are. This solves about half of the CDT. The general observation was that the patients are most prone to errors like skipping a few digits, or drawing them at faulty distances or confusing the centre of the clock face.

Since the system assumes the larger sketches to be the hands, the lengths of their major axes (measured by functionality provided in MATLAB®) are known which is used to identify the minute and the hour hand (shorter). The system identifies the angles of the hands in relation to the horizontal axis to determine the pointing directions. Also it is easy to check for superfluous markings, extra hands, and position of centre. The 'normal' position of centre is defined as anywhere within 20% distance of the radius from the centre. Any more deviation would not look like the clock centre to the human eye as well. Of course it may be possible that the patient does not draw (or imply) the centre at all. But since the major axes of hands are known it is possible to measure whether they will converge within that 20% from centre distance.

Thus the entire CDT is evaluated. Of course to process these sketches a lot of assumptions are being made. Firstly, we concede that ML doesn't seem robust enough. Secondly, the hands may be so short they get confused for digits. Thirdly the major axes of the hands might not be the direction the patient was intending to point to, fourthly there are chances that 10, 11 and 12 are identified incorrectly (both false positives and negatives). Some smudges may be construed as superfluous markings. And there is no real way to test point 2 that only Arabic numerals are present. If the patient uses roman (or any other numeral system) the system will diagnose those as superfluous markings and not as numerals. Similarly some patients have been known to write the words for the numbers rather than draw them.

ANALYSIS AND CONCLUSION

Even if this system will almost never give a highly accurate score for each patient, it will be useful to detect crests and troughs to measure the progress (or otherwise) of a large number of patients easing the load on the medical officers. The handwritten digit detection part almost reaches 90% accuracy. What this system displays is that in a constrained environment it is easier to do sketch recognition. Also a lot of extra information is being generated which might prove useful for the medical investigators who use the CDT. In the CDT the sketches are either digits or hands which can be handled by machine learning techniques and some geometrical measurements. We state that it is the context of the sketch which plays a vital role in determining what techniques should be used to analyze those sketches because the term 'sketch' may cover a wide spectrum. To do any meaningful sort of analysis of sketches, we need to have a preference bias, something which is a well known rule of thumb in machine learning. Without an a priori preference bias it is difficult to determine what will work before truing it out which becomes intractable. Also when dynamic recognition (pen motion) is not essential, machine learning may be used to identify sketches taking them as final images and converting sketch recognition into a classification problem. We posit that knowing features such as motion points would make recognition even more accurate.

REFERENCES

[1] Steenland NK, Auman CM, Patel PM, et al. Development of a rapid screening instrument for mild cognitive impairment and undiagnosed dementia. Journal of Alzheimer's disease: JAD. 2008;15(3):419-27.

[2] Strauss E, Sherman EM, Spreen O. A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. Oxford University Press, USA; 2006

[3] FAISAL TEHSEEN SHAH, KAMRAN YOUSAF (2007)*Handwritten Digit Recognition Using Image Processing and Neural Networks* Proceedings of the World Congress on Engineering Vol I WCE 2007, July 2 - 4, 2007, London, U.K.

[4] CHENG-LIN LIU, KAZUKI NAKASHIMA, HIROSHI SAKO, HIROMICHI FUJISAWA, *Handwritten Digit Recognition Using State-of-the-Art Techniques* 2002 IEEE

Interpreting Gestures for Text Entry on Touch Screen Devices

Gennaro Costagliola, Vittorio Fuccella
University of Salerno
Via Ponte Don Melillo
84084 Fisciano (SA), Italy
{gencos, vfuccella}@unisa.it

Michele Di Capua
Unlimited Software
Centro Direzionale, Isola F/11
80143 Napoli, Italy
md@unlimitedsoftware.it

Abstract

Text entry on touch screen devices is often performed through Soft keyboards. One of the latest research trends is to abandon the traditional tapping interaction in favor of more natural gesture-based interactions on these keyboards. The interpretation of the gestures is performed through sketch-based techniques.

In this paper we present the sketch-based technology related to the interpretation of the gestures needed to enter text through KeyScretch, a novel text entry method which has been proven to be more efficient than the traditional tapping-based method. The method enables text entry through the input of gestures associated to word chunks. The gestures are guided by a menu which appears around a key of the keyboard as soon as the user presses it.

The proposed procedures improve the interpretation of the gestures and consequently increase the accuracy of the method. This improvement is measured through a simulation in which the sketch-based recognition is compared to the target-based menu item selection through an estimation of the error rate.

1 Introduction

The efficiency of text entry on mobile devices is limited due to the reduced dimension of these devices. In most cases, palmtop and smart phones equipped with touch screens require the users to interact with a *soft keyboard*, a keyboard drawn on the screen. The main method of interacting with these keyboards is *tapping*. Each *tap* corresponds to a single character input. The latest research trend is to abandon such an interaction in favor of a more natural gesture-based interaction performed on these keyboards. As an example of this, we can cite *Shapewriter* [15], in which the user draws over a keyboard, i.e. with a *QWERTY* layout, to link up the letters of a word s/he wishes to write.

Recently, Costagliola et al. [2] introduced a novel text entry method, called *KeyScretch*. The method uses a combination of soft keyboards and menus: a menu is shown around the pressed key enabling the input of a text unit through a gesture touching the menu items associated to the characters. *KeyScretch* is based on previous works: the use of menu-augmented keyboards has been studied by Isokoski [5]. In its basic form, the use of the menu enables the input of digraphs through a single interaction, called a *flick*. The improvement introduced with *KeyScretch* is the support of multiple-selection of menu items with a pointer gesture. The use of these gestures allows the user to enter particularly frequent text chunks instead of simple digraphs. *KeyScretch* outperforms the traditional tapping-based method and significantly improves the previous menu-based one too.

The interpretation of the gestures is performed through sketch-based techniques. In this paper we present the sketch-based technology related through the interpretation of the gestures needed to enter text through *KeyScretch*. These techniques provide a better interpretation of the user gestures, compared to the target-based menu item selection, and consequently relax the error rate. The error reduction has been estimated by running the sketch-based recognizer on the tracks produced by the user gestures during a longitudinal usability study. For both the sketch-based and the target-based methods, the error rate is calculated and reported for each session of the study. In this paper we give details of the procedures employed by the sketch-based recognizer, based on geometrical sketch recognition techniques.

The rest of the paper is organized as follows: the next section contains a brief survey of the main text entry methods related to touch screens; section 3 describes the text entry method and its performances; section 4 discusses the proposed recognition model in detail; lastly, the evaluation of the model is described in section 5.

2 Text Entry Methods: A Brief Survey

Several methods for accelerating the text entry task have been proposed. In this brief survey we only consider text entry methods related to touch screens. For a more comprehensive survey on text entry on mobile devices, the reader should refer to [10]. Some of the main directions followed by researchers are:

- handwriting and shorthand writing recognition;

- proposal of *more efficient* keyboard layouts;
- prediction of text;

Handwriting recognition is not an easy task and raises several challenges, such as segmentation [10]. Furthermore, handwriting is rather slow: its speeds are commonly in the 15 - 25 WPM range. To ease the recognition and avoid segmentation, the methods for handwriting on mobile devices have been limited to the recognition of one character at a time. Examples of these methods are *Unistrokes* [3] and *Graffiti* [1]. As for *shorthand*, it allows users to reach high performances in terms of speed (up to 100 WPM [9]). Nevertheless it appears unlikely that common users could adopt such a difficult to learn method. A method allowing common users to easily learn the symbols is that of using schemes indicating the path the pointer should follow, such as the *pentagrid* used in the *VirHKey* method [13].

The *QWERTY* layout is the most familiar to the users. Unfortunately, it is not the most efficient with *tapping*. An efficient keyboard layout should minimize the distance between characters with a high probability of being consecutive in the words of target languages. To this aim, alternative layouts have been proposed, such as *OPTI* [11], *Metropolis* [16] and others. These layouts enable a faster and more accurate entry (more than 40 WPM with expert users), compared to the traditional *QWERTY* layout.

Predictive input methods reduce the effort required to enter text by predicting what the user is entering. A typical predictive handwriting system presents possible next words as a list and allows the user to select one to skip manual writing. As argued in [8], predictive text entry with English language is not necessarily faster than just simply finishing typing the words, since it requires the cognitive load of selecting the candidates from a list.

A noteworthy attempt to improve the interaction with soft keyboards through gestures, implemented by the *SHARK2* system [6], is to use a single gesture to produce an entire word. Instead of *tapping*, the user draws over a keyboard, i.e. with a *QWERTY* layout, to link up the letters of a word s/he wishes to write. This idea has been also exploited in commercial initiatives [15]. This approach let the user learn the gestures for common words and draw them without looking at the keyboard. The gesture is initially visually guided (for novice users): the pointer must touch the characters located on the keyboard in order to enter a word. Then, when learned, it is recognized through the comparison of the drawn shape (pattern recognition) to a vocabulary of gestures. The limit of this approach is that only a small set of frequent gestures can be remembered by the user.

3 The *KeyScretch* Text Entry Method

The *KeyScretch* method enables text entry through the input of gestures associated to word chunks instead of entire words. The gestures are guided by a menu which appears around a key of the keyboard as soon as the user presses the key and disappears as soon as the user releases the pointer.



Figure 1. The QWERTY keyboard layout augmented with a menu.

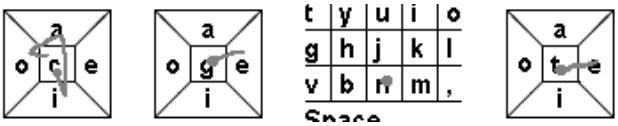


Figure 2. The 4 strokes needed to enter the text 'ciao gente'.

While the menu is shown, it is possible to sweep out a gesture that touches only the desired characters in succession by dragging the pointer, without lifting it (this gesture recalls the one used for the *Cirrin* method [12]). A space character can be inserted by releasing the pointer after returning inside the character key area. If the pointer is lifted in the menu area, instead, the text unit is terminated with the last selected character. Note that in all the other cases the space bar is directly used.

KeyScretch can be instantiated differently for different languages. As a first instance, we have associated each of the vowels 'a', 'e', 'i' and 'o' to each of the sides of the squared character keys. As shown in [2] this choice is particularly suited for the Italian language since these vowels are the most frequent letters in Italian and, furthermore, they are more easily remembered by users. Different menu characters may be needed by different languages. By arranging the menu items as shown in Figure 1, the interaction sequence necessary to enter the Italian text *ciao gente* (*hello folks*, in English) is shown in Figure 2. The string is ten characters long but it can be entered with a sequence of four strokes (*taps* or *gestures*). The strokes correspond to the input of the following sequence of text units $\{ciao\}\{ge\}\{n\}\{te\}$. Three text units out of four are entered through a gesture and only one through a *tap*. With the first stroke we can enter up to five characters.

3.1 Performances

A theoretical analysis on the expected performances in terms of speed obtainable by an ideal user has been performed in [2]. Furthermore, a longitudinal usability study with 6 users has also been recently performed. The methodology of this experiment follows the standard procedures used in the field [11]: the users entered text on a touchscreen with both methods for 20 sessions, for 15 minutes with each method. During the sessions the users entered as fast and as accurate as possible short text phrases, selected in a random

order from an 80 Italian phrases database representative of the language.

The study allowed us to evaluate the learning curve of both the traditional and the *KeyScretch* method. The scatter diagrams of the two metrics, speed and accuracy, are shown Figure 3. As for the former, we can point out that the performance crossover can be reached rather soon (close to the 10th session, after about 2 hours and a half of use) across the sessions and the *KeyScretch* method can outperform the traditional tapping-based method in the subsequent sessions and, all the more so, in the estimation of future sessions. The best average speed with the *KeyScretch* method has been reached at the 19th session, with 42.7 WPM with a moderate error rate of 3.35%. The fastest user has exceeded the the considerable speed of 52 WPM at the last session with an error rate of 3.31%. For each method, we derived standard regression models in the form of the power law of learning. The prediction equations and the squared correlation coefficients are illustrated the figure.

As for the accuracy, we point out that, preserving the status quo, our method has a lower accuracy, on average. In fact, the *KeyScretch* method has an error rate of 3.18%, versus the 1.93% of the traditional method. Nevertheless, the difference is more marked in the earlier sessions and tends to diminish with practice. Since this paper is more focused on the methods for sketch recognition, more details on the usability study are going to be published in an upcoming paper.

4 The Sketch-Based Recognition Model

In the user study described in the previous section, the recognition of user gestures has been performed through a target-based recognizer: the initial character of the text unit is selected by detecting the position where the pointer is pressed; the following characters are selected when the pointer exactly enters a menu item area; the possible final space is selected by detecting the position where the pointer is released.

An exploratory visual analysis of the errors committed by the users during the use of the menu has been carried out by comparing the track of the gestures on the keyboard and the corresponding produced text. From such an analysis, we have realized that:

- Most of the errors are committed in the use of the menu (73.5%) rather than in tapping interactions. Furthermore, the greatest part of them is due to articulatory concerns (43.3% of all errors) instead of cognitive ones. That is, the errors are primarily due to inaccuracy in the physical action necessary to select an item from the menu and not to mistakes of the user in remembering the position of a character in the menu.
- It is possible to recover from most of the articulatory errors by better interpreting the user gestures. In particular, a geometrical recognition of the basic movements intended by the users can have the best performances over other recognition methods.

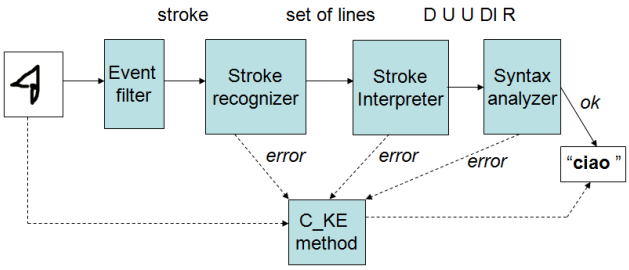


Figure 4. The data flow.

In this section we describe the sketch-based procedures carried out for the interpretation of the gestures. The sketch-based recognizer basically works as follows: the recognizer tries to obtain the text chunk. If it fails (it can fail because the stroke cannot be recognized as a polyline, or the lines obtained do not represent a legal movement inside the menu), the stroke is recognized through an optimized version of the target-based C_KE method, where C stands for *Centering* while KE stands for *Key Enlargement*. In fact, the optimization lies in translating the gesture such that its starting point is centered in the key. Furthermore, the area available to the user to end the stroke inside the squared key (to enter a space) is enlarged of a margin in both its width and height.

To elaborate, the architecture of the recognizer is shown in Figure 4. The data flows through the following modules:

1. The *Event Filter*: captures the input events of the device and constructs the stroke;
2. The *Stroke Recognizer*: transforms the stroke to a polyline;
3. The *Stroke Interpreter*: transforms the polyline to a set of movements among the menu items;
4. The *Syntax Analyzer*: transforms the movements to a text chunk.

4.1 Event Filter

This module captures the input events of the device (pointer *press*, *drag* and *release*) and constructs the stroke. Following a common approach in sketch recognition, the stroke is represented through an ordered sequence of *TimedPoints*: triples in the form (t, x, y), where t is the timestamp, and x and y are the x-coordinate and the y-coordinate of the point on the screen, respectively. In the simplest scenario, the stroke can be easily constructed by adding a *TimedPoint* for all of the *drag* events between a starting *press* and a final *release* events. This layer also includes methods which cope with defects of the device: in particular, the device used during the experiment outlined in section 3.1 (A *Sympodium ID250* Interactive Pen Display), sometimes generated undesired additional events during the use of the pointer. Thus, the events had to be suitably filtered.

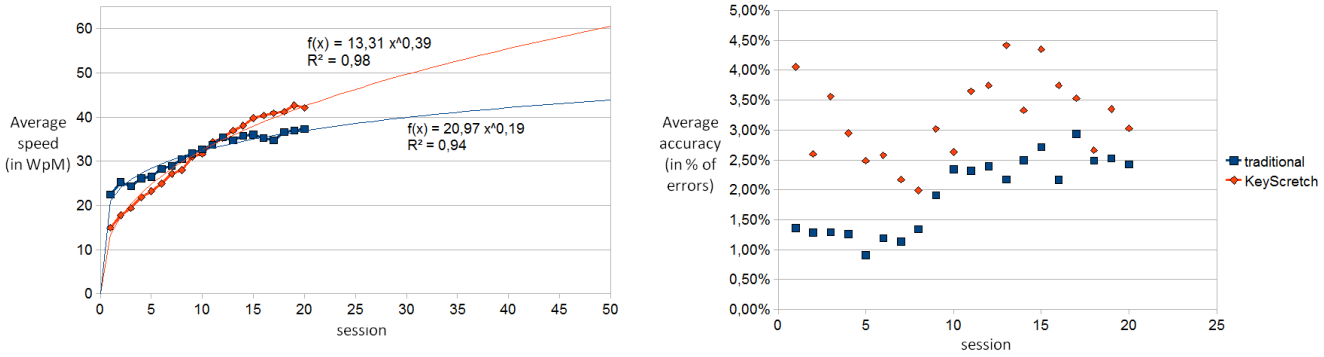


Figure 3. The text entry speed (left) and accuracy (right) measured across the 20 sessions.

4.2 Stroke Recognizer

Firstly, the stroke recognizer takes the strokes from the *Event Filter* and produces a polyline. A polyline is a continuous line composed of one or more line segments. The *Ladder* framework [4] has been used for recognizing the polylines. Ideally, each segment should represent the movement of the pointer to reach the desired menu item. Without firstly cleaning the stroke, this rarely occurs: we then implemented the *Line Fusion* and *Tail Removal* steps to refine the polyline before passing it to the *Stroke Interpreter*. An example of the recognition of a polyline from a stroke is shown in Figure 5(a) and (b) where the menus are not shown. The original stroke is represented in light gray color, the polyline in black. The resulting polyline as recognized by *Ladder* is composed of 6 segments: $b_1..b_6$.

The *Line Fusion* step finds segments which cannot be considered complete movements to reach a menu item, but parts of a unique movement (they are recognized as separate segments due to curvature). The fusion is always performed on the segment of the polyline whose angle with the previous segment is the minimum on the whole polyline. The fusion is performed only if such an angle amplitude is less than a threshold value. The application of the *Line Fusion* procedure on the sample stroke in Figure 5(b) is shown in Figure 5(c): starting from the $b_1..b_6$ polyline, we obtain a simplified polyline with only 4 segments $c_1..c_4$. In two consecutive iterations the procedure fuses in a single segment c_3 segments b_3, b_4 and b_5 while the tails c_1 and c_4 are still to be processed.

The tail removal is a typical procedure in sketch recognition. Paulson and Hammond [14] propose a procedure based on the length and the curvature of the first and last 20% of the stroke. In our procedure we have not considered the percentage length of the segment over the stroke at all, since even a simple tap can result in an unintended snatch on a key, which can be erroneously interpreted as a menu item selection (see Figure 8(c)). We based our procedure on the absolute length of the segment and the difference between the timestamps of the segment endpoints. If the segment length is less than a threshold *l* or the time difference is less than a treshold *t*, the segment is regarded as a tail and removed. This is the case of the segments c_1 and c_4 shown in Figure 5(c). Their removal leads to the resulting clean polyline in Figure 5(d).

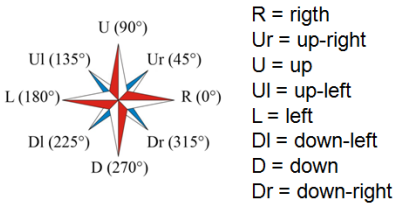


Figure 6. The directions attributed to the polyline segments.

4.3 Stroke Interpreter

The *Interpretation* step transforms the clean polyline representing the calculated user intended character selections in a sequence of movements among menu items. The sequence is represented through a string composed of a concatenation of strings representing possible movements in the 8 main directions shown in Figure 6. The slope of a segment is approximated to the nearest direction. The slopes of the directions are shown in the figure. As for the sample in Figure 5(d), the two segments are attributed the directions 'R' and 'Dl'; thus, the final string representing the interpretation of the stroke is 'RDl'.

If a horizontal segment is approximately (a multiplying factor must be set as a threshold) twice the first segment of the polyline, the movement is doubled (i.e. 'R' becomes 'RR'), in order to interpret the movement from a menu item to the opposite (left-to-right, right-to-left, top-to-bottom or bottom-to-top).

4.4 Syntax Analyzer

In order to formally specify the language describing the set of legal movement strings we define the automaton state diagram represented in Figure 7. Here the states represent

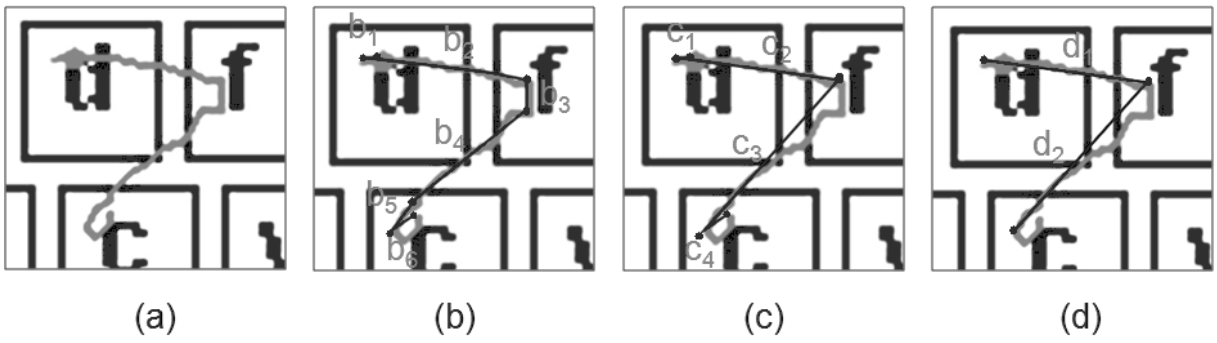


Figure 5. A sample of the Stroke Recognition process.

the 4 menu items (q_1, q_2, q_3, q_4) and the key (q_0), while the transitions represent the legal movements that may be produced from each state. All the states are final and q_0 is also the initial state. In order to formally specify the translation from movement strings into text chunks we build a Moore machine starting from this automaton. The machine associates each of the $c_1..c_4$ states to the output of a character. In particular, in the tested instance it associates q_1 to the output 'a', q_2 to the output 'e', q_3 to the output 'i', and q_4 to the output 'o'. To reflect our semantics we need to split state q_0 in three states q_{01}, q_{02}, and q_{03} where:

1. q_{01} is both the initial and a final state associated to the key character as output, and with only the 4 q_0 outgoing transitions;

2. q_{02} is a non final state associated to no output (empty word) and with both the 4 outgoing and 4 incoming transitions of q_0;

3. q_{03} is a final state associated to the 'space' character as output and with only the 4 q_0 incoming transitions.

As an example, by producing the legal movement string 'DUUDlR' starting from the key 'c', we obtain the sequence of Moore machine states q_{01}, q_3, q_{02}, q_1, q_4, q_{03} producing the output text chunk 'ciao '.

5 Evaluation

The sketch-based recognizer described in the previous section can translate a gesture into a legal sequence of movements among the menu items in 95.09% of the cases (from the data of the longitudinal study outlined in section 3.1). In the remaining 4.91% cases, the recognition must be performed through the *C_KE* target-based method, described in the previous section.

Some cases of better interpretation of the user's will are shown in Figure 8. A tail removal is shown in Figure 8(a) the tail caused a re-enter in the key area, interpreted as a trailing space by the previous target-based recognizer. The sketch-based recognizer has correctly recognized the movement as composed of a successive selection of the bottom and the right menu items ('DUr'). The movements shown

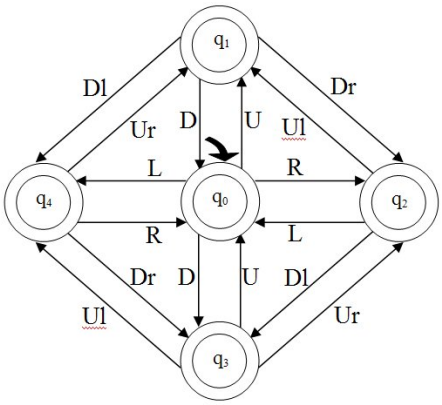


Figure 7. The automaton state diagram used to recognize the movement string.

in Figure 8(b) were interpreted as a sequence of a movement to the right followed by a movement to the left menu item, due to an excessive amplitude of the return movement. The comparable length of the two lines has been correctly interpreted by the sketch-based recognizer as a movement to the right menu item followed by a return to the key area ('RL' instead of 'RLL'). Lastly, the gesture in Figure 8(c) is a simple tap. Nevertheless, the presence of a tail caused an erroneous interpretation of the gesture as a selection of the right menu item. The tail has been removed and the tap has been correctly recognized.

The translation of a gesture in movement among menu items does not necessarily produce a correct text chunk, that is, the text can contain typing errors, due to user incorrect typing or to incorrect recognition. The effectiveness of the sketch-based recognizer has been measured by comparing the error rate obtained by the target-based recognizer in the longitudinal study, to the one that could have been obtained by the sketch-based recognizer. For the latter, the text entry has been simulated through a suitable software module which implements the recognizer: it takes as input the logged tracks and generates the corresponding phrases.

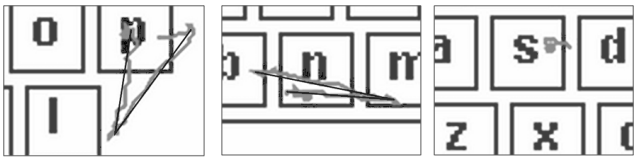


Figure 8. Some cases of correct interpretation of the sketch-based recognizer.

This approach (the use of a simulation run on the data previously gathered in an experiment) is based on previous work: Kristensson and Zhai [7] run a simulation on pen traces recorded from participants in order to show accuracy improvement on *SHARK2* [6] text entry method.

The recognizer has been tuned by setting its parameters on 4 out of the 20 sessions (the 3rd, 8th, 13th and 18th) and tested on the remaining 16 sessions. The improvement obtained on the training data, that is the percentage of error reduction on the 3rd, 8th, 13th and 18th sessions was 32.59%, 8.55%, 35.03%, 15.08%, respectively, with an average of 22.81%. The error reduction on the test set was even better, with an average of 24.75%. Overall, the use of the sketch-based recognizer redured the error rate from 3.18% to 2.20%, thus making the accuracy of *KeyScretch* comparable to the one measured for the traditional tapping-based method from the 9th/10th session on. An analysis of variance revealed a significant difference in error rates between the two recognizers ($F_{1,5} = 68.1$, $p < .0001$).

To be sure that the sketch-based technique is effective also with other users than those with which it has been tuned, we ran another user study. In this experiment, 6 users performed text entry only with the *KeyScretch* method and for just one session of 15 minutes. 3 users were completely novices, while 3 of them had some previous experience with the method occurred about 3 months before this experiment, quantifiable in about 2 hours of practice. The recognizer, applied on the novice users, produced an error reduction of 10.46%, and a reduction of 13.07% on the 3 lightly experienced users. These values are in the range obtained on the tracks of the longitudinal study, even if they are under their average values.

6 Conclusion

In a previous research we have defined the new text entry method *KeyScretch*. Both theoretical and usability studies show that *KeyScretch* has good performances from the point of view of the speed, outperforming the tapping-based method.

In this paper we presented the sketch-based technology related to the interpretation of the gestures needed to enter text through *KeyScretch*. The proposed procedures improve the interpretation of the gestures and consequently increase the accuracy of the method. The evaluation shows that geometric sketch recognition techniques, associated to other calibrations can improve the accuracy of the method and make it comparable to the one measured for the traditional method just after a few hours of use. Future work include: further studies aimed at investigating the possibility of achieving improvements on the recognizer, in order to further tune the accuracy; the comparison of the presented recognizer with other recognizers based on different techniques (i.e. pattern matching); the exploitation of other methods (i.e. dictionary-besed ones) to correct errors. A demo of *KeyScretch* can be run at
`http://keyscretch.altervista.org/KeyScretch.html`

References

[1] C. H. Blickenstorfer. Graffiti: Wow!!!! *Pen Computing Magazine*, pages 30–31, 1995.

[2] G. Costagliola, V. Fuccella, M. Di Capua, and G. Guardi. Performances of multiple-selection enabled menus in soft keyboards. In *Proceedings of DMS'09 (VLC Workshop)*, pages 359–364, 2009.

[3] D. Goldberg and C. Richardson. Touch-typing with a stylus. In *Proceedings of INTERCHI '93*, pages 80–87, Amsterdam, The Netherlands, The Netherlands, 1993. IOS Press.

[4] T. Hammond and R. Davis. Ladder, a sketching language for user interface developers. *Comput. Graph.*, 29(4):518–532, 2005.

[5] P. Isokoski. Performance of menu-augmented soft keyboards. In *Proc. of CHI '04*, pages 423–430, New York, NY, USA, 2004. ACM.

[6] P.-O. Kristensson and S. Zhai. Shark2: a large vocabulary shorthand writing system for pen-based computers. In *Proc. of UIST '04*, pages 43–52, NY, USA, 2004. ACM.

[7] P. O. Kristensson and S. Zhai. Improving word-recognizers using an interactive lexicon with active and passive words. In *Proceedings of IUI '08*, pages 353–356, New York, NY, USA, 2008. ACM.

[8] K. Kurihara, M. Goto, J. Ogata, and T. Igarashi. Speech pen: predictive handwriting based on ambient multimodal recognition. In *Proceedings of CHI '06*, pages 851–860, New York, NY, USA, 2006. ACM.

[9] Y. Ma, G. Leedham, C. Higgins, and S. M. Htwe. Segmentation and recognition of phonetic features in handwritten pitman shorthand. *Pattern Recogn.*, 41(4):1280–1294, 2008.

[10] I. S. MacKenzie and R. W. Soukoreff. Text entry for mobile computing: Models and methods, theory and practice. *Human-Computer Interaction*, 17:147–198., 2002.

[11] I. S. MacKenzie and S. X. Zhang. The design and evaluation of a high-performance soft keyboard. In *Proc. of CHI '99*, pages 25–31, New York, NY, USA, 1999. ACM.

[12] J. Mankoff and G. D. Abowd. Cirrin: a word-level unistroke keyboard for pen input. In *Proc. of the ACM UIST 98*, pages 213 – 214. ACM, 1998.

[13] B. Martin. Virhkey: a virtual hyperbolic keyboard with gesture interaction and visual feedback for mobile devices. In *Proceedings of MobileHCI'05*, pages 99–106, New York, NY, USA, 2005. ACM.

[14] B. Paulson and T. Hammond. Paleosketch: accurate primitive sketch recognition and beautification. In *Proceedings of IUI '08*, pages 1–10, New York, NY, USA, 2008. ACM.

[15] Shapewriter. http://www.shapewriter.com/, 2010.

[16] S. Zhai, M. Hunter, and B. A. Smith. The metropolis keyboard - an exploration of quantitative techniques for virtual keyboard design. In *Proc. of UIST '00*, pages 119–128, New York, NY, USA, 2000. ACM.

Evaluation of Paper-Pen based Sketching Interface

Pankaj Rajan
Sketch Recognition Lab
Dept. of Computer Science
Texas A&M University
pankaj@cse.tamu.edu

Paul Taele
Sketch Recognition Lab
Dept. of Computer Science
Texas A&M University
ptaele@cse.tamu.edu

Tracy Hammond
Sketch Recognition Lab
Dept. of Computer Science
Texas A&M University
hammond@cse.tamu.edu

Abstract

The retrieval and browsing of diagrammatic information extracted from hand-drawn diagrams would open up a rich form of information interaction. However, such sketches currently require hand-annotations in order to be understood by the computer. While improvements in sketch recognition algorithms have enabled automatic recognition for Tablet PC-sketched diagrams, such progress has been constrained to online algorithms. As a result, offline algorithms that are relevant to diagrams sketched on paper remain dominantly domain-dependent, and are also restrictive in the number of diagrams that can be understood. In this paper, we discuss our research aims for providing users with information interaction that exploit the advantages of automatic correction capabilities found in online sketch recognition algorithms, with the low-cost advantages found in paper usage.

1. Introduction

Tablet PC-based interfaces allow users to sketch diagrams directly on a computer, then have them recognized automatically in real-time by a computer. These computer-recognized hand sketches allow the users to interact with a computer as is most natural for them, and also allow them to take advantage of the powerful capabilities provided by a computer (e.g., CAD and CASE tools); this provides further understanding, processing, analysis, and simulation of the design.

The literature [22] tells us that using such interactive tools divert the focus of the designer from the important issues like behavior and structural layout to more trivial aesthetic details like color and alignment; this diverted focus is not necessary at the initial design stage. There is a need for a bridging medium that does not limit the imagination of the user, and at the same time provides the same level of aesthetic layout that any conventional designing software can provide. One direction is the use of tablet-based sketch recognition systems, which have recently gained considerable popularity. Sketch recognition systems [9,11,12] have been developed over period of time that can build the foundations for complex CAD designs. However, although tablets provide a more sophisticated sketching platform capable of automated understanding and intelligent interaction [7], they cannot match the naturalness, flexibility, ease, and costs provided by a piece of paper [8].

The aim of our research is to provide users with the naturalness and flexibility of sketching on a piece of paper, and at the same time provide them with an option to use developing and current state-of-the-art stroke-based sketch recognition systems [2,7,14]. We do so by presenting sketch recognition results from the LADDER sketch system on stroke data automatically generated from our algorithm.

2. Related Work

Tablet-based sketches can exploit the strokes' temporal and spatial data for segmentation. Since paper-based sketches lack this temporal information, identifying the correct pixels to the corresponding stroke is a challenge; this holds especially true for overlapping strokes. Natural cues from users sketching, such as those based on Gestalt's law of continuity [21], should be able to assign pixels to extracted strokes that maintain the stroke trace's continuity. The term *saliency measure* from the stroke extraction literature does just that, selecting pixels that favor long, smooth, and continuous strokes in the image, though others argue that the features from global saliency better represent the stroke's actual perception [6]. The latter have used probabilistic measurements to calculate the contribution of the single unit edge to the stroke's global perception. Other works maximize the saliency measure using the curvature and curvature variation of the network of closely connected components [20] and as a gain of energy from the addition of a new edge to the curve being traced [19].

Other similar works include grouping and clustering points of the perceptual pattern [1], using the figural closure as a saliency measure [17], converting paper scribbles to single vector lines, and employing co-occurrence matrix-based approach to simplify and describe drawings more meaningfully [4]. The method

we employ differs in that those techniques attempt to recognize the primitives as lines, where our method matches the lines to users' original stroke path and time-stamp them afterwards for use in sketch recognition algorithms. Furthermore, our approach is not restricted to a small set of chosen primitives, and can use both the traced strokes' global and local characteristics to closest match users' drawing intentions perceptually.

There also exist similar domain-specific systems for extracting strokes from paper: one on extracting interpretations of paper-sketched UML diagrams [13] and another on extracting shape structures of handwritten signatures [10]. The former incorporates domain-specific rules to recognize the shapes, while the latter converts the sketch into a binary image and thins the strokes to preserve the structure. Our work differs from theirs in that ours is domain-independent, unlike [13]; and can handle local ambiguities, unlike [10].

We have also summarized the details of our algorithm [16] in the next section, but this paper differs from the details in that we focus on the viability of a paper-pen interface for use in tasks instead of the technical aspects.

3. Approach

The technique used for the envisioned paper-pen interface consists of three basic steps [16]. First, as in most image processing techniques, we have to preprocess the image to remove noise. Second, we "thin" the pen strokes to a thickness of only a single pixel. Third, we trace the line using both the local and global characteristics of the previously identified stroke points to determine the future points of the stroke. Further details of the algorithm are as follows:

1. If scanned input is noisy, remove noise by using median filtering and global thresholding [5].

2. Dilate pen strokes to single pixel thickness using Pratt's morphological thinning algorithm [15].

3. For overlapping strokes, associate correct pixel to correct stroke in accordance with Gestalt's law. To do this, identify local and global characteristics of traced stroke by analyzing extracted points. Use local characteristics to identify different paths that a given stroke can take, and global characteristics to identify most optimal perceptually correct path.

4. Because of overlaps, it is possible that a stroke may lose some pixels to another stroke, which may prematurely halt the traversal of a stroke during extraction. To deal with this, define a merging function which looks for possible combinations of all extracted strokes in terms of a confidence value. Then select the combination for merging that has the best confidence value. Compute confidence value of merging using merge function, taking into consideration angle of the merge and the distance between the two end points of two strokes. Obtain missing pixels between two stroke ends from merging process with linear interpolation.

After extracting and merging all the possible strokes, artificial time stamps are appended to each extracted stroke point. We use these generated time stamps because speed information cannot be extracted from static images. However, points are still ordered consecutively, such that the continuity of the stroke is preserved. This data is then converted into XML, which is later used as input to the online sketch recognition system (i.e., LADDER [7] in our case).

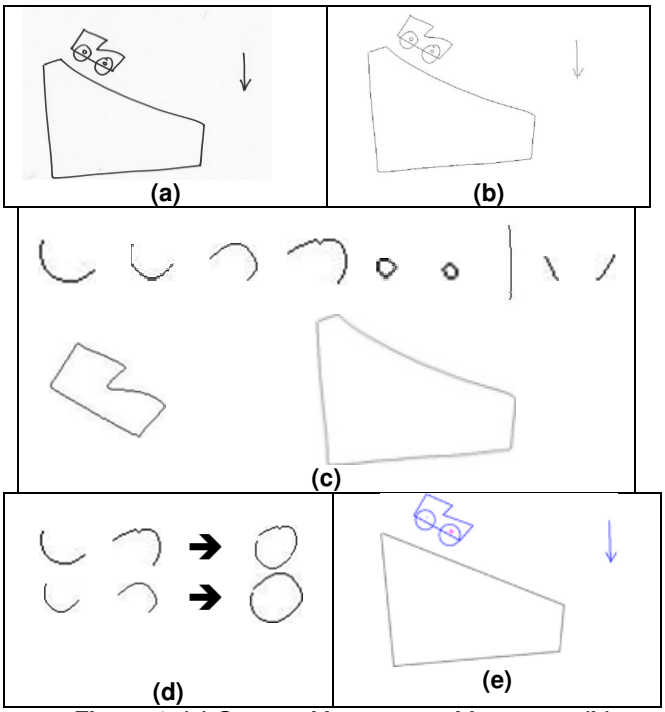


Figure 1. (a) Scanned image on white paper, (b) thinning and binary image created, (c) extracted strokes, (d) extracted strokes that were merged, and (e) final recognized result in LADDER.

LADDER is chosen as the recognition engine because it provides users with an ability to sketch a given shape naturally without enforcing specialized sketching gestures for a given shape. Moreover, any new domain that can be described with the available primitives can be easily defined in LADDER. LADDER uses a context-based grammar to define a sketching domain in which

multiple compound sketch shapes (i.e., shapes formed by combining multiple primitive sketch shapes) can be defined and later recognized. Figure 1 demonstrates a complete working of our algorithm from extraction to merging and from merging to recognition.

4. User Study

During the course of our research, we conducted various user studies to verify our different claims:

1. Certain features of our paper-pen based interface would be more preferred than tablet-based interface. (Interaction user study)

2. Scanned-in images could be used in our online sketch recognition algorithms. (Integration user study)

3. Our algorithm is able to extract out correct strokes from paper as would be extracted by a human. (Verification user study)

4. Our system can be used in a large environment setting. (Verification user study)

4.1. Interaction User Study

In order to verify that our system provides users with a more natural interactive experience, we conducted an interaction user study. Four subjects participated in this user study with each subject asked to sketch a diagram first on paper and then on a tablet with a pen and mouse. At the end of the user study, participants were asked to complete a survey in which they had to report their responses on a scale of 1 (best) to 3 (worst).

Table 1. Bracketed number shows how many subjects gave the score on left.

Which of the two interfaces:	Pen	Mouse
Has Lower Learning Curve	2[4]	3[4]
Is easier to use	2[4]	3[4]
Has better expected Feedback	2[4]	3[3]; 1[1]
Can be easily used in large settings	2[4]	3[3]; 1[1]
Gives freedom of sketching	2(4)	3(4)

The study demonstrated that all of the users agreed with our perception about the advantages of a paper-based interface compared to a tablet-based interface. Users preferred the paper-based input and its more naturalness and freedom over tablet-based interfaces. One of the subjects commented that even though tablet-based interfaces were interesting and sophisticated, they lacked the frictional feedback that paper provided while sketching. Some users with extensive experience using tablet-based interfaces argued that even though paper-based sketching is more natural, tablet-based sketching allowed the ability to store the information digitally which they either can be shared through e-mails or stored in storage devices. Some users also complained that since we are using paper-based sketching for a much longer time compared to tablet-based sketching, it was not a fair comparison. However, they did agree with our claim that paper-based sketching is much more universal and easily adaptable than tablet-based systems.

4.2. Integration User Study

This user study was aimed at verifying that our system can be integrated with an online sketch recognition system such as LADDER. We also verified that a paper-based sketch interface matched the recognition accuracy of a tablet-based system under similar drawing constraints.

In our integration user study, three subjects participated. Each of the subjects were provided with a pen and white paper, and asked to sketch a music sheet diagram and a tic–tac-toe game. Their response papers were scanned and presented to our algorithm to generate the artificially time-stamped extracted stroke points. These stroke points were then used as input to LADDER, and recognition accuracy was computed as:

$$\frac{Number\ of\ sketch\ components\ recognized}{Total\ number\ of\ sketch\ components\ in\ the\ sketch}$$

The sketch component is the constituent of a sketch which represents a complete shape. For example, a music note is a single sketch component.

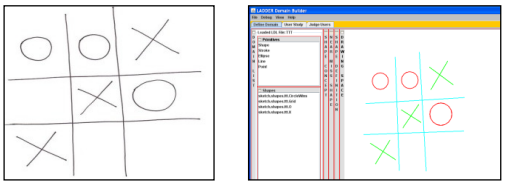


Figure 2. The left side shows the scanned in image and the right side shows the scanned image recognized in LADDER [8].

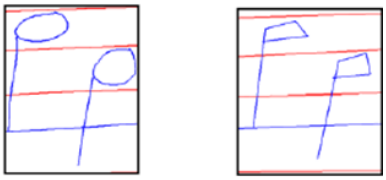


Figure 3. (Left) The original strokes in LADDER after applying the stroke-tracing algorithm. (Right) Strokes after the low-level recognizer was applied.

It was observed that music notes in the sheet music domain were not recognized correctly. This was because the sketch recognition system was unable to effectively fragment the stroke into an oval and a shaft, which comprised a music note. This can be attributed to the observation that there is no effective corner available in the music note that can be used for segmentation of a stroke into two constituents (i.e., oval and a shaft).

In order to compare our results with the online system, we conducted another user study in which the same three users whom had participated in the previous user study participated. They were shown their sketches which they had on paper and told to sketch the same on a tablet. However, they were instructed to sketch music notes as a single stroke. Table 2 shows the comparison of the accuracy results obtained.

Table 2. Tablet vs. Paper

Sketch	Tablet	Pen
Music-1	0.7	0.8
Music-2	0.667	0.778
Music-3	0.778	0.667
TicTacToe-1	1	1
TicTacToe-2	1	1
TicTacToe-3	1	1

It can be seen that we obtained nearly similar results with the online system as we obtained using a paper-based interface.

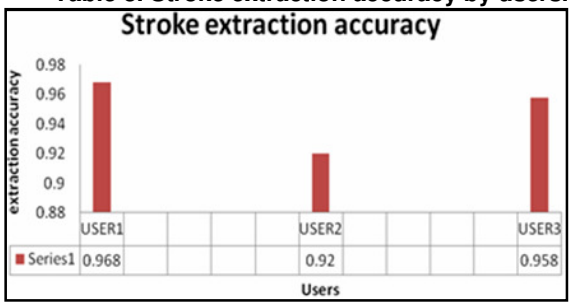
4.3. Verification User Study

This user study was conducted to verify how best our algorithm can match the extraction result compared to a human. Another objective of this user study was to validate that our system can be used in a large environment setting like educational domain.

In our user study, ten users participated. Out of ten users, seven were engineers whereas two were graduate students in our computer science department. The last user was from a non-engineering background with a graduate degree in mass communication. A 15-minute basic Finite State Machines (FSM) tutorial was conducted, where users were introduced to FSM through sketches and verbal explanations. They were also provided with some sample sketches to get familiar with the sketching conventions for drawing FSMs.

After this initial introduction, all the users were given some FSMs to copy on white paper for practice. After this, an analogous example was dictated to the users and they were asked to sketch the FSM solution for the problem. They were initially asked to sketch a rough solution and then copy their final solution on a new white paper with clarity. They were encouraged to sketch and draw freely, except that they were told that the start or end point of the arrow should not lie on the circumference of the circle. This constraint was imposed because the local conditions of the intersection favored the arrowhead component stroke to follow the circle during extraction in some cases, which caused misrecognition. However, we can argue that this constraint would not affect the natural sketching style of users. Considering the fact that many users had little knowledge about FSMs, they were helped in their logical doubts, but no attempt was made to affect their natural sketching style. Thus, we were able to capture the natural sketching style of the user and at the same time were able to collect some real world classroom data. In total, 4 such sketches of different FSM's were collected per user.

Table 3. Stroke extraction accuracy by users.



4.3.1 Stroke Extraction Evaluation

In order to evaluate that our algorithm is extracting correct strokes from the images, we conducted a small user study in which three users participated distinct from those ten that participated in the previous study. They were shown color-coded images of the extracted strokes that were obtained by running our extraction algorithm on 40 scanned in images. They were also asked to compute sketch extraction accuracy as:

$$\frac{Number\ of\ perceptually\ correct\ extracted\ stroke}{Total\ number\ of\ strokes\ in\ the\ sketch\ as\ seen}$$

This user study demonstrated that our extraction algorithm was able to achieve a sketch extraction accuracy of 94.8% on the FSM domain. Table 3 shows the average accuracy reported by each of the 3 users.

4.3.2 Stroke Recognition Evaluation

We also evaluated the performance of our system as input to LADDER. The extraction data was time-stamped and presented to LADDER as input for high-level recognition. The recognition accuracy for LADDER was computed as:

$$\frac{Number\ of\ recognized\ components\ in\ a\ clean\ sketch}{Total\ number\ of\ components\ in\ the\ sketch}$$

For example, in a given FSM with three states and three transition arrows, if three states are identified correctly and three transition arrows are not recognized, the accuracy of the system would be 0.5.

4.4. Near Miss vs. Perception

While evaluating our data from the user study, we observed that arrows accounted for almost all of the near miss cases. In order to determine if the arrows that were not recognized by the sketch recognition system do perceptually look like arrows to humans, we conducted a perceptual user study in which 5 users participated. In this user study, users were shown screen shorts of sketches processed by LADDER and asked to note the number of arrows they see on a given screen shot.

Table 4. Results from our algorithm for our dataset.

Users	Components	Recognized	Miss Recognized	Near Miss	Near Miss Accuracy	Actual Accuracy
1	31	21	10	9	.967	0.677
2	28	21	7	6	.928	0.75
3	30	17	13	9	0.866	0.566
4	33	24	9	2	0.787	0.727
5	33	25	8	4	0.878	0.757
6	33	22	11	6	0.848	0.667
7	31	17	14	12	0.935	0.548
8	32	26	6	3	0.9026	0.812
9	31	21	10	3	0.774	0.677
10	32	28	4	4	1	0.876
AVERAGE ACCURACY					0.888	0.706

After gathering input from the five users on the same set of data, the number of arrows associated with a given sketch was taken as the number of arrows reported by the maximum users. For example, if for a given sketch three users identified four arrows and two identified three arrows, the number of arrows associated with that sketch was taken as four. In case of all different responses, whole number closest to the median was taken as the number of arrows associated with that sketch. We computed near miss accuracy as:

$$\frac{\sum_{i=1}^{n}(NSI_i - NAS_i) + (NAI_i)}{\sum_{i}^{n} TS_i}$$

where, $NSI_i$is the total number of sketches identified in the given sketch *i*, $NAS_i$is the actual number of arrows present in the sketch *i*, NAI_i is the number of arrows identified by a human user in the cleaned sketch (obtained after processing through LADDER) and $TS_i$is the total number of sketch components in the sketch *i*. The number of images is given by *n*.

The main goal of computing this accuracy was to account for the limitation of recognizer, which if improved could also improve the overall accuracy of our paper based sketching interface. Near-miss examples were actually the shapes in the domain, which look very similar to the shapes being recognized but because some constraints were not met, recognizer failed to identify the shapes. In this user study, we obtained average actual recognition accuracy of 70.6% and the average near miss accuracy obtained of 88.88%.

5. Conclusion

In this work, we presented a summary of our stroke extraction algorithm explained in [16]. The main focus of this paper was our major user study that was conducted to verify our claim that our system can be used in large environment settings such as educational domain. We also evaluated extraction accuracy of our algorithm by presenting color-coded extraction data to three different subjects to identify the number of perceptually correct strokes in the sketch. This study demonstrated that the three users reported extraction accuracy of our algorithm as 96.8%, 92.5% and 95.8% respectively. The average accuracy was reported as 94.8%. This study demonstrated that our extraction algorithm was able to match human perception of correct stroke extraction to a great extent. Our integration analysis, in which we used our extracted data with artificially appended time stamps as input to LADDER, achieved an accuracy of 70.6% which can be improved to 88.88% or more by further improvements in the traditional sketch recognizers. We believe that artificial time-stamps that were generated by our algorithm to use our data in online sketch recognizers did seem to work; however, it did have small impact on recognition ability of LADDER. It should be noted that LADDER does use speed data as one of the features for recognizing low level primitives. We also conducted a T-Test, based on means to determine statistical difference between the tablet-based recognition and paper-and-pen based recognition results. The two-tailed P-value based on means equal 0.6433 proved that difference is statistically insignificant.

6. Future Work

Despite the cost-effectiveness of our interface compared to Tablet PCs, we admit that artificial time-stamp did affect the accuracy of our system. Speed data is used by most of the sketch recognition systems like [14] and [18] to identify corners or to parse two components of the stroke. As such, we believe that the inability to obtain the actual speed may affect the recognition ability of our system significantly. It is worthwhile to investigate if the information provided by the extracted stroke data can be combined with the recognition ability of offline vision-based sketch recognition algorithms [5].

Integrating other domain specific knowledge, such as shape description and characteristics, can also be used along with the sketch data to develop a more robust recognizer for a paper-and-pen-based interface. A recognizer, which can combine the power of statically pattern recognition, with stroke data-based sketch recognition systems, is worth investigating in the future.

The current preprocessing technique, based on morphological thinning would not provide best results for such sketches. As such, work by Bartolo in [4] can be used as a pre processing step for such sketches and thus make the proposed system more robust.

7. Acknowledgement

This work was funded in part by the National Science Foundation through IIS grant #0757557. Authors would also like to thank the current members of the TAMU Sketch Recognition Lab for their valuable comments and editing help for the paper.

8. References

[1] Ahuja, N., Tuceryan, M. "Extraction of early perceptual structure in dot patterns: integrating region, boundary, and component gestalt." Computer Vision, Graphics, and Image Processing 48.3 (1989): 304-356.

[2] Alvarado, C., Davis, R. "Sketch READ: A multi-domain sketch recognition engine." In Proceedings of 17th Annual ACM Symposium on User Interface Software and Technology. Santa Fe, NM, USA, 2004.

[3] Bartolo, A., Camilleri, K.P., Fabri, S.G., Borg, J.C., Farrugia, P.J. "Scribbles to vectors: preparation of scribble drawings for CAD interpretation." Proceedings of the 4th Eurographics Workshop on Sketch-Based interfaces and Modeling (SBIM '07). Riverside, CA, USA, 2007.

[4] Bartolo, A., Camilleri, K.P., Farrugia, P.J., Borg, J.C. "A New Sketch Based Interface using the Grey-level Co-occurence Matrix for Perceptual Simplification of Paper Based Scribbles." Eurographics Workshop on Sketch Based Interfaces. Ed. Sousa M. C., Jorge J. A. Stahovich T. 2006. 91—98.

[5] Gonzales, R.C., Woods, R.E. Digital Image Processing: Second Edition. 2002.

[6] Guy, G., Medioni, G. "Inferring global perceptual contours from local features." International Journal of Computer Vision 20 (1996): 113-133.

[7] Hammond, T., Davis, R. "LADDER, sketching language for user interface developers." Computers & Graphics August 2005: 518–532.

[8] Henzen, A., Ailenei, N., Fiore, F. D., Van Reeth, F., Patterson, J. "Sketching with a low-latency electronic ink drawing." Proceedings of the 3rd international Conference on Computer Graphics and interactive Techniques in Australasia and South East Asia. Dunedin, New Zealand, 2005.

[9] Juchmes, R., Leclercq, P., Azar, S. "A multi agents system for the interpretation of architectural sketches." Proceedings of the 1st Annual EUROGRAPHICS Workshop on Sketch Based Interfaces and Modeling. Grenoble, France, 2004.

[10] Lee, S., Pan, J.C. "Off-line tracing and representation of signatures." IEEE Transactions on Systems, Man, and Cybernetics 22 (1992): 755–771.

[11] Lipson, H., Msary, M. "A sketch based interface for the iterative design and analysis of the 3D objects." Proceedings of the 2nd Annual EUROGRAPHICS Workshop on Sketch Based Interfaces and Modeling. Dublin, Ireland, 2005.

[12] Lipson, H., Shpitlani, M. "Optimization based reconstruction of 3D object from a single freehand line drawing." Journal of Computer Aided Design 28.8 (1996): 651-663.

[13] Notowidigo M., Miller R. "Off-line sketch interpretation." Proceedings of AAAI Fall Symposium on Making Pen-Based Interaction Intelligent and Natural. 2004.

[14] Paulson, B. and Hammond, T. "PaleoSketch: accurate primitive sketch recognition and beautification." Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI '08). Gran Canaria, Spain, 2008.

[15] Pratt, W.K. Digital Image Processing. John Wiley & Sons, Inc., 1991.

[16] Rajan, P., Hammond, T. "From Paper to Machine: Extracting Strokes from Images for use in Sketch Recognition." 5th Annual EUROGRAPHICS Workshop on Sketch Based Interfaces and Modeling. 2008.

[17] Saund, E. "Finding Perceptually Closed Paths in Sketches and Drawings." IEEE Transactions on Pattern Analysis and Machine Intelligence 25.4 (2003): 475-491.

[18] Sezgin, T. M., Stahovich, T., and Davis, R. "Sketch based interfaces: early processing for sketch understanding." Proceedings of the 2001 Workshop on Perceptive User Interfaces (PUI '01). Orlando, FL, USA, 2001.

[19] Tarel, F.G. "Curve Finder Combining Perceptual Grouping and a Kalman Like Fitting." Proceedings of the international Conference on Computer Vision. Washington, DC, USA: IEEE Computer Society, 1999.

[20] Ullman, S., Sha"ashua, A. "Structural Saliency: the Detection of Globally Salient Structures Using a Locally Connected Network." 1988.

[21] Waranusast, R. "Perceptual-based region extraction from hand drawn sketches." Proceedings of the Third Conference on IASTED international Conference: Advances in Computer Science and Technology. Phuket, Thailand, 2007.

[22] Wong, Y.Y. "Rough and ready prototypes: Lessons from graphic design." Short Talks Proceedings of CHI '92: Human Factors in Computing Systems. Monterey, CA, 1992.

Sketch-a-Doc: Using Sketches to Find Documents

Manuel J. Fonseca Daniel Gonçalves
Dept. of Computer Science and Engineering
INESC-ID/IST/TU Lisbon, Lisbon, Portugal
{mjf,daniel.goncalves}@inesc-id.pt

Abstract

With the vast amount of documents that users tend to accumulate in their hard drives, it is natural that they often forget where a certain file is stored or even its name. However, sometimes they still recall a mental image of the document layout. To explore this, we propose a new approach to document retrieval, based on sketches, that capitalizes on human visual memory to help users find their personal documents. The users can sketch the layout of the desired document, using a calligraphic interface, and the system will present to them those that match the sketched query. Documents are processed to extract their relevant features, blocks are segmented and classified according to their contents and a description of the layout is created. A visual language is used to identify the blocks specified using sketches.

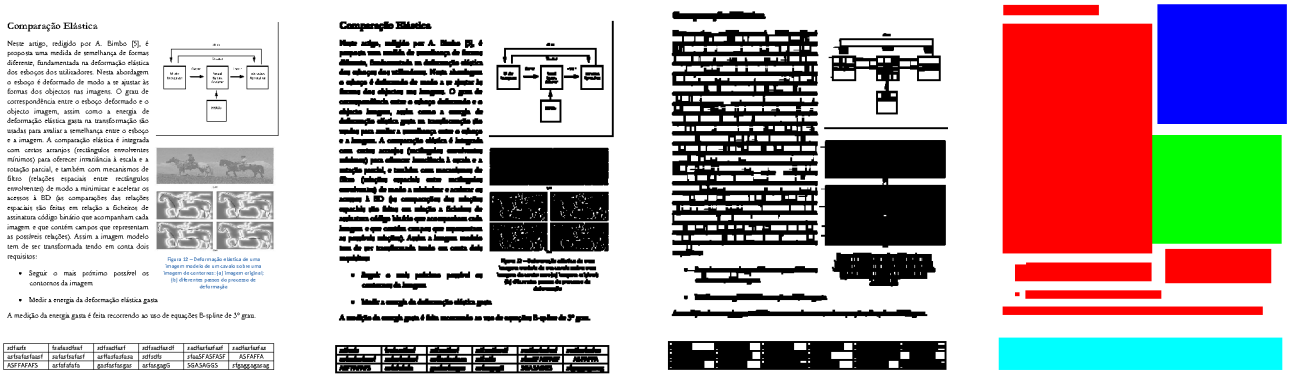
1. Introduction

Nowadays, we sometimes find it hard to locate that document that we know to be somewhere in our hard drive. Usually, this is not a trifling task, especially for older documents. Sometimes when looking for a document, we can not remember its name, location or the typical attributes used in common search tasks, to get satisfactory results. However, occasionally we can recall its appearance, how the first page looked like (it was written on two columns, it had a picture at the top of its rightmost column and a table at the bottom of the document, etc.). The visual memory of the user plays a crucial role in the recognition of objects and also documents. Human beings can easily remember and perceive images rather than words.

In the last decades, as an attempt to help users retrieve their documents based on their appearance, various systems were studied and developed. Some addressed only part of the problem (document processing and segmentation, image indexation, etc.) while others tried to produce complete document retrieval solutions. One system that tried to employ such a method was WISDOM++ [2], which extracts the semantic structure from document images and uses machine learning techniques to automatically label the different blocks identified. Hashimoto and Igarashi [6] presented an approach for web page retrieval based on its layout. Their solution compares a specified layout (using predefined icons) with the layout defined by HTML tags. This approach supports elements like text, images and tables, but does not allow users to freely sketch the page layout. Watai et al. [9] also presented an approach for web page retrieval using visual queries. In their approach users can find web pages by sketching colored regions, which are then compared to screenshots of the web pages using image retrieval techniques. More recently, Lecerf and Chidlovskii [7] presented a schema for querying large document collections by document layout. To speed up search in large databases, authors developed a technique based on clustering to identify a set of representative blocks (images and text) from all the collection and use these to compare with the query.

In this paper, we explore users' visual memory and the fact that one of the best and easiest ways to describe a visual representation of something, in this case a document, is by sketching it, to define a new document retrieval model based on sketches. Our approach combines a richer document layout description with the use of sketches to specify queries. We identify a set of blocks from the document according to their content, namely text areas, images, graphics, tables, horizontal lines and vertical lines. To describe the layout of the document page we developed two techniques. One uses a topology graph to code the spatial arrangement of the page, the distance between the blocks and their type; and another that uses a predefined grid. For the retrieval we developed a calligraphic interface, capable of supporting free hand sketches to specify the layout of the desired document. These sketches are then recognized and their semantic meaning identified through a visual language.

In the remainder of the paper, we explain how we analyze and describe the documents in Section 2, while in Section 3 we explain the use of sketches to specify the queries. Section 4 presents some preliminary results and Section 5



a) b) c) d)

Figure 1. Document analysis process. Left to right: a) the original document; b) the document after thresholding and applying the erosion filter; c) the result of the RSLA; d) the final result, with all blocks colored according to their type (red - text, blue - graphic, green - image, light blue - table).

concludes the paper and enumerates the future work.

2. Document Analysis and Description

Our solution is composed of two main processes. One is responsible for analyzing the document and extract the information about the different blocks (text, image, etc.) that constitute it. The second component produces descriptions of the document in terms of its layout and relationships between the various blocks.

2.1. Document Analysis

To extract the required features from a document we first transform it into a format more amenable for processing, abstracting from the underlying file format (pdf, doc, etc.). To that end, we grab the first page of the document and convert it into an image [5], taking advantage of the several techniques already available for image processing and block segmentation.

After grabbing the first page of the document we convert it into a black and white image, using a basic threshold algorithm. This led to an image reflecting the overall structure of the document page, as illustrated in Figure 1-b. To eliminate some noise caused by the thresholding process we applied an erosion filter with a simple cross flat structuring element (which essentially removes all black pixels not surrounded by other black pixels). This minimizes the number of spurious pixels and improves the efficiency of the block segmentation algorithm.

Many studies have already been made about image and document block segmentation and classification. Some of them are strict rule-based approaches and others more dynamic, resorting to machine learning techniques. For simplicity and effectiveness sakes, we use an improved two-step block segmentation version of the original RLSA (Run-Length Smoothing Algorithm) [8]. This algorithm is able to detect content blocks in a document image, while at the same time classifies those blocks according to their type. To detect content blocks, the RLSA algorithm starts by finding content lines (uninterrupted horizontal sequences of pixels), and then group those closer than a predefined threshold (see Figure 1-c). Although this algorithm is able to identify blocks of the type text, image, graphic, horizontal line and vertical line, it is not able to identify tables, which were one of the block types we needed to identify. To support this we changed RLSA rules and parameters. Now, our algorithm first identifies tables, images, graphics and lines, letting the text blocks for the "otherwise" rule, since the other blocks are easier to recognize than text. Figure 1-d shows the identified blocks.

2.2. Description using a Grid

The grid description is based in spatial organization features only. To that end, we divide the document layout according to a pre-defined 4x10 grid. We chose to partition the document page in 40 units (four columns and ten rows) because after some analysis we concluded that almost no document was formatted in more than three columns. The number of rows was chosen empirically, and ten was the number that held enough expressiveness.

Figure 2 shows a document with the grid overlaid on it. According to this type of description we would get

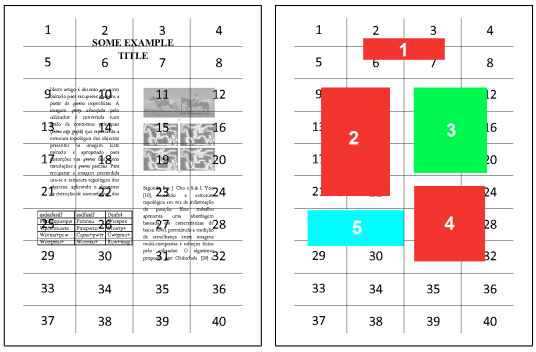


Figure 2. Grid of a document layout.

a specification like this: There is a text block in cells $\langle 2, 3, 6, 7\rangle$, another in $\langle 9, 10, 13, 14, 17, 18, 21, 22\rangle$ and another one at $\langle 23, 24, 27, 28, 31, 32\rangle$; there is an image at $\langle 11, 12, 15, 16, 19, 20\rangle$; and there is also a table located at cells $\langle 25, 26, 29, 30\rangle$. Although this description is very simple, it is good enough to portray the layout of the first page of a document. In this case there are no intersections on the same grid cell, but if there were the algorithm would use the same cell number on the description of both blocks.

These type of description allows us to know the type (or types) of blocks that are present on each of the grid cells. During retrieval, we are able to compare these descriptions with the query grid submitted by the user through a sketch.

2.3. Description using a Topology Graph

We also developed another description mechanism using a topology graph to code the spatial organization of the document page. We extract two types of spatial relationships between blocks (horizontal and vertical connections), the block type (text, image, etc.) and the relative sizes of the blocks to the page size. The block type and the relative size are stored in the graph nodes, while the type of relationship

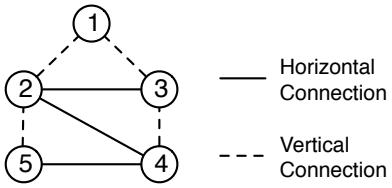


Figure 3. Topology graph describing the layout of the document page in terms of horizontal and vertical connections among blocks.

3. Sketches to Find Documents

To take advantage of the human visual memory, we devised a solution to retrieve documents where users define the layout and the type of blocks through sketches. Users can draw blocks at the desired position and use a simple visual grammar to specify their types.

3.1. Visual Grammar

We defined a visual grammar, based on the studies performed on our previous works [1, 3], to identify the six types of blocks (text, image, graphic, horizontal line, vertical line and table).

$$\texttt{Text} \rightarrow \{\texttt{Rectangle WavyLine}\} \quad if \ \text{Contains}(\texttt{Rectangle,WavyLine})$$

$$\texttt{Image} \rightarrow \{\texttt{Rectangle Circle}\} \quad if \ \text{Contains}(\texttt{Rectangle,Circle})$$

$$\texttt{Graphic} \rightarrow \{\texttt{Rectangle Triangle}\} \quad if \ \text{Contains}(\texttt{Rectangle,Triangle})$$

$$\texttt{Table} \rightarrow \{\texttt{Rectangle Cross}\} \quad if \ \text{Contains}(\texttt{Rectangle,Cross})$$

$$\texttt{HLine} \rightarrow \{\texttt{Line}\} \quad if \ \text{Horizontal}(\texttt{Line})$$

$$\texttt{VLine} \rightarrow \{\texttt{Line}\} \quad if \ \text{Vertical}(\texttt{Line})$$

Terminal symbols on the productions correspond to geometric shapes identified from the sketches drawn by users, using the CALI recognizer [4], as illustrated in Figure 4.

3.2. Matching

Currently we have two algorithms to compare the query with the documents in our database. One uses the grid description while the other uses the topology graph. Our aim is to evaluate both descriptions to figure out which is the best, or if we can obtain better results by combining them.

To perform the matching using the grid description we look for every document that have the same type of block from the query in the same grid cells. For each match the document awarded one point. After processing all blocks from the query we sort the documents by the number of points and present to the user those with the highest scores.

For the description using the topology graph, we compare the topology graph extracted from the sketched query with the topology graphs from the documents.

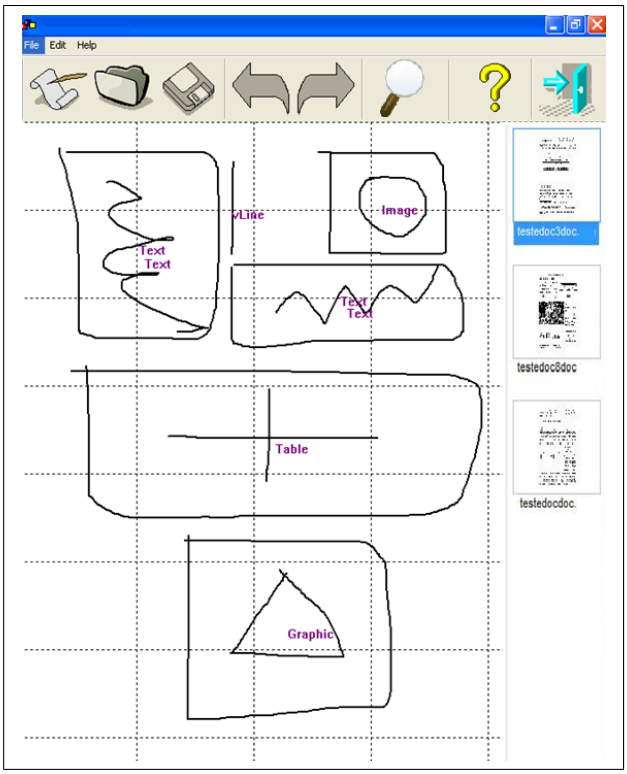


Figure 4. Example of a query specified using sketches and the visual grammar. Results are presented on the right.

4. Experimental Evaluation

Since this work is at a preliminary stage, we did not yet evaluate the two description and matching techniques. Currently we were only able to evaluate the modified RLSA algorithm for block identification.

The RLSA algorithm is sensitive to various parameters used in the identification rules. To choose the most appropriate values for these parameters we performed an experiment in which several values were tried and the quality of the results evaluated. We used a set of 46 documents, representative of different block types, sizes and combinations commonly found in personal documents.

At the end, we were able to infer the parameter values that produces, on average, the best results. Overall, our algorithm is able to correctly identify and classify 87.5% of all blocks.

5. Conclusions and Future Work

In this paper we presented an approach for document retrieval based on their layout and on sketches to specify queries, which takes advantage of the visual memory owned by humans. To that end we modified the RLSA algorithm to identify six types of blocks from the first page of a document, and we developed two approaches to describe the spatial organization of the blocks. One uses a grid while the other uses a topology graph. For query specification we rely on a calligraphic interface where the users describe the layout of the desired document through sketches. We also developed a simple visual grammar to interpret the users' sketches and identify the type of blocks drawn.

Our next steps will be the evaluation of both description approaches to see which one presents better retrieval results and the optimization of the matching algorithms to provide a more efficient solution. One possibility is the inclusion of an indexing mechanism to avoid the comparison of the query with all the documents in the database.

Acknowledgments

This work was supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program, project A-CSCW (PTDC/EIA/67589/2006) and project Crush (PTDC/EIA-EIA/108077/2008). We thank Filipe Alves for coding the prototype.

References

[1] M. P. Albuquerque, M. J. Fonseca, and J. A. Jorge. Visual Languages for Sketching Documents. In *IEEE Int. Symposium on Visual Languages (VL'00)*, 2000.

[2] M. Berardi, M. Lapi, and D. Malerba. An integrated approach for automatic semantic structure extraction in document images. In *Document Analysis Systems (DAS'04)*, 2004.

[3] A. Caetano, N. Goulart, M. Fonseca, and J. Jorge. Javasketchit: Issues in sketching the look of user interfaces. In *AAAI Spring Symposium - Sketch Understanding*, 2002.

[4] M. J. Fonseca and J. A. Jorge. Experimental Evaluation of an on-line Scribble Recognizer. *Pattern Recognition Letters*, 22(12), 2001.

[5] D. Gonçalves and J. A. Jorge. In search of personal information: narrative-based interfaces. In *Int. Conference on Intelligent User Interfaces (IUI '08)*, 2008.

[6] Y. Hashimoto and T. Igarashi. Retrieving web page layouts using sketches to support example-based web design. In *Eurographics Workshop on Sketch-Based Interfaces and Modeling (SBIM'05)*, 2005.

[7] L. Lecerf and B. Chidlovskii. Scalable indexing for layout based document retrieval and ranking. In *ACM Symposium on Applied Computing (SAC'10)*, 2010.

[8] F. Shih and S. Chen. Adaptive document block segmentation and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 26(5), 1996.

[9] Y. Watai, T. Yamasaki, and K. Aizawa. View-based web page retrieval using interactive sketch query. In *Int. Conference on Image Processing (ICIP'07)*, 2007.

Reviewers' Index

A
Giovanni Adorni
Tim Arndt

B
Arvind K. Bansal
Dorothea Blostein
Paolo Buono

C
Tolga Capin
Alfonso F. Cardenas
Augusto Celentano
Ing-Ray Chen
Shu-Ching Chen
William Cheng-Chung Chu
Yuan-Sun Chu
Kendra Cooper
Maria Francesca Costabile
Gennaro Costagliola
Philip Cox
Alfredo Cuzzocrea

D
Sergiu Dascalu
Davide De Chiara
Andrea De Lucia
Massimo De Santo
Vincenzo Del Fatto
Aidan Delaney
Vincenzo Deufemia
Pierpaolo Di Bitonto
Sergio Di Martino
Stephan Diehl
Jing Dong
David H. C. Du

E
Larbi Esmahi

F
Filomena Ferrucci
Andrew Fish
Paul Fishwick
Daniela Fogli
Manuel J. Fonseca
Farshad Fotouhi
Rita Francese
Vittorio Fuccella
Kaori Fujinami

G
Moncef Gabbouj
Ombretta Gaggi
Stefan Goebel
Dorian Gorgan
Jivesh Govil
Angela Guercio

H
Niklas Hallberg
Mao Lin Huang
Kao-Shing Hwang

I
Pedro Isaias

J
Hewijin Christine Jiau
Qun Jin
Joemon Jose
Erland Jungert

K

Egidijus Kazanavi ius
Yasir Khalifa
Wolfgang Klas
Lars Knipping
Yiannis Kompatsiaris
Jun Kong
Yau-Hwang Kuo

L

Robert Laurini
Fuhua Lin
Alan Liu
Jonathan Liu
Benjamin Lok

M

Paolo Maresca
Andrea Molinari
Piero Mussio

N

Satoshi Nakamura
Paolo Nesi
Kia Ng

O

Vincent Oria

P

Sethuraman Panchanathan
Luca Paolino
Joseph J. Pfeiffer
Antonio Piccinno
Fabio Pittarello
Giuseppe Polese

R

Steven P. Reiss
Teresa Roselli
Veronica Rossano

S

Nicoletta Sala
Giuseppe Santucci
Giuseppe Scanniello
Monica Sebillo
Buntarou Shizuki
Peter Stanchev
Gem Stapleton
Hui-Kai Su
Nik Swoboda
Mahbub Syed
Mahbubur Syed

T

Bjorn Thuresson
Christian Timmerer
Genny Tortora

U

Joseph E. Urban

V

Athena Vakali
Stefano Valtolina
Athanasios Vasilakos
Giuliana Vitiello

W

Ellen Walker

Y

Kazuo Yana
Chi-Lu Yang
Atsuo Yoshitaka

Z

Ing Tomas Zeman
Kang Zhang

Authors' Index

A

Andrea F. Abate, 261
Julia Albath, 277
Abdullah S. Alghamdi, 149
Francesca Arcelli, 23
C. Ardito, 206
Carmelo Ardito, 180
Ryan Armstrong, 251

B

Swapnil Barai, 289
Emanuele Bellini, 153
Pierfrancesco Bellini, 7, 110
J. Beyerer, 64
Tine Blomme, 98
Karel Braeckman, 98
Sermsak Buntha, 300
Paolo Buono, 180

C

Nicola Capuano, 222
Alfonso F. Cardenas, 289
Angela Carrillo-Ramos, 164
Rosanna Cassino, 271
Daniele Cenni, 7, 110
Shi-Kuo Chang, 33, 47
Yeim-Kuan Chang, 76
Peter Chapman, 283
Edward Chow, 132
Chih-Ping Chu, 70, 76
Carlos E. Cirilo, 168
Mauro Coccoli, 189
Francesco Colace, 33, 47
Kendra M.L. Cooper, 257, 297
Vladimiro Corsi, 110
Maria F. Costabile, 180
Gennaro Costagliola, 222, 315
Joao Bosco S. Cunha, 94
Salvatore Cuomo, 228

D

Doraid Dalalah, 88
Sergiu Dascalu, 300
D. De Chiara, 265
Mattia De Rosa, 222
Massimo De Santo, 33, 47
Rogéria Cristiane Gratão de Souza, 11
Wanderley L. de Souza, 168
Robbie De Sutter, 98
Vincenzo Del Fatto, 261
Aidan Delaney, 305
Yi Deng, xxii
Marcel Aime Deussom, 153
P. Di Bitonto, 211
Michele Di Capua, 315
Antonio F. do Prado, 168
Ellen Yi-Luen Do, 311
Tiansi Dong, 41
Jorg Dorflinger, 174
Gerald Draxler, 104
David Hung-Chang Du, xxi

E

Chris Ellison, 137
Lars Eriksson, 57

F

Clay Fenlason, 235
Mauricio G. V. Ferreira, 94
Manuel J. Fonseca, 327
G. Franco, 193
Vittorio Fuccella, 222, 228, 315
Alice Fuzier, 7

G

Aurona Gerber, 200
Richard Göbel, 143
Daniel Goncalves, 327
Christina Grönwall, 57
Angela Guercio, 216
Anupam Guha, 311

H

Charles Haiber, 216
Niklas Hallberg, 57
Kevin W. Hamlen, 257
Tracy Hammond, 321
Frederick C. Harris, Jr., 300
Kimiyoshi Hayashi, 243
Yoshiyuki Hino, 243
Yen-Chieh Huang, 70, 76

J

Todd Jensen, 232
Chad Jones, 251
Erland Jungert, 57

K

Hyungsin Kim, 311
Josh Knight, 114
Hisato Kobayashi, 243
Jeffrey Koch, 297
Yasushi Kodama, 239, 243
Zora Konjović, 17
E. Krempel, 64
Carsten Kropf, 143
Varun Kulkarni, 126
Chia-Hao Kuo, 70
Birgitta Kylesten, 57

L

Andr´e Luiz Peron Martins Lanna, 29
Fredrik Lantz, 57
R. Lanzilotti, 206
M. Laterza, 211
Petra le Roux, 200
Jennifer L. Leopold, 277
Juzheng Li, 82
Yu-Shih Lin, 70
Jonathan C.L. Liu, 126

M

Kwan-Liu Ma, 251
Antonio Maña, 120
Paolo Maresca, 189, 193
Mike Matton, 98
Michael J. McMahon, Jr., 300
Veselina Milanova, 174
Branko Milosavljević, 17
Muhanna Muhanna, 300
Aniello Murano, 228

N

Michele Nappi, 261
Paolo Nesi, 7, 110, 153
Antonio Marcos Neves Esteca, 11
Kia Ng, 137
Luis Carlos Niño-Tavera, 164
G. Nota, 193

O

Sohei Okamoto, 300
Bee Ong, 137

P

Gianni Pantaleo, 7
L. Paolino, 265
Michela Paolucci, 7
Aditi Patwardhan, 257
Ting-Chun Peng, 33, 53
Billy Pham, 243
Carlos Alberto Marques Pietrobon, 29
Beryl Plimmer, 305
Riccardo Porrini, 23
Gimena Pujol, 120

R

Pankaj Rajan, 321
Stefano Ricciardi, 261
Pierluigi Ritrovato, 222
Peter Rodgers, 305
Nenad Rogulja, 17
M. Romano, 265
T. Roselli, 211
V. Rossano, 211
Edgar Enrique Ruiz, 164

S

Chaman L. Sabharwal, 277
Johannes Sametinger, 104
Mireille Samia, 3
Adriana Barbosa Santos, 11
Kay-Uwe Schmidt, 174
Olaf A. Schulte, 232
Monica Sebillo, 265, 271
Sol M. Shatz, 82
Chia-Chun Shih, 33, 53
Zeeshan Siddiqui, 149
Axel Spriestersbach, 174
Lidia Stanganelli, 189
Gem Stapleton, 283, 305
Donald V Steward, 114
Kathleen A. Stirbens, 216
Thangaraj Karol Stuart, 159

T

Paul Taele, 321
Yuichi Takahashi, 239
Yuki Terawaki, 239
Susan Thomas, 174
Adriana Cursino Thome, 94
Yuji Tokiwa, 239
Francisco Torres, 132
Genoveffa Tortora, 261, 265
Maurizio Tucci , 271

V

H. Vagts, 64
Alta van der Merwe, 200
G. Vitiello, 265
Mattia Vivanti, 23

W

Andreas Wiesauer, 104
Joseph Williams, 216

X

Zhiwei Xu, 159

Y

Kazuo Yana, 239, 243
Chi-Lu Yang, 76

Z

Luciana A. M. Zaina, 168
Marco Zanoni, 23
Emilio Zegarra, 33

DMS 2011 Call For Papers

The Seventeenth International Conference on Distributed Multimedia Systems

www.ksi.edu/seke/dms11.html

Convitto della Calza (www.calza.it/it/) Florence, Italy
August 18 - August 20, 2011

Organized by

Knowledge Systems Institute, USA, University of Florence, Italy, University of Leeds. UK

Call for papers

The DMS conference is an international conference series, which covers a wide spectrum of paper presentations, technical discussions and demonstrations in the fields of distributed multimedia computing. The theme for DMS2011 is Multimedia and the Arts. The conference organizers seek contributions of high quality papers, panels or tutorials, addressing various aspects of distributed multimedia systems and applications, for presentation at the conference and publication in the proceedings. Both research and case study papers or demonstrators describing results in research area as well as industrial development cases and experiences are solicited. Demonstrators prototypes are welcome to support the submission and to be presented at the conference in the discussion sections.

Special issue of journals

The DMS conference is closely coordinated with the International Journal of Software Engineering and Knowledge Engineering, the Journal of Visual Languages and Computing and the International Journal of Distance Education Technologies. Usually the best ranked papers from the conference, after rigorous reviews, extensive revisions and further enhancements, will appear in one or more special issues of the above journals. Papers suggested for the special issue(s) will be reviewed by external reviewers following the standard procedure of review stipulated by the respective journal.

Topics

Distributed Multimedia Technology
- media coding and compression standards
- media streaming and QoS control
- digital rights management
- middleware and OS support
- privacy and security issues
- mobile devices and wireless networks
- sensor networks
- context sensing and processing

Distributed Multimedia Models and Systems
- human-computer interaction issues
- situated computing
- languages for distributed multimedia: Java, XML, X3D
- software engineering issues
- web services and multi-agent systems
- multimedia databases and information systems
- multimedia indexing and retrieval systems
- distributed multimedia systems evaluation
- multimedia authoring

Applications of Distributed Multimedia Systems
- distributed collaboration
- cultural heritage preservation, management and fruition
- distance and lifelong learning
- emergency and safety management
- environment control and management
- e-commerce and e-government applications
- health care management and disability assistance
- intelligent multimedia computing
- social networking
- virtual, mixed and augmented reality

The use of prototypes and demonstration video for presentations is encouraged

Information for authors

Papers must be written in English. An electronic version (Postscript, PDF, or MS Word format) of the full paper should be submitted using the following URL: http://conf.ksi.edu/dms2011/submit/SubmitPaper.php. Manuscript must include a 200-word abstract and no more than 6 pages of IEEE double column text (include figures and references).

All submissions must not be published or under consideration for publication in a journal or in a conference with proceedings. Papers will be evaluated based on originality, significance, technical soundness and clarify of exposition. Depending upon the results of evaluation a paper may be accepted as regular paper (6 pages), short paper (4 pages) or poster (poster presentation only).

Workshops and Special Sessions

This year, DMS 2011 will be held in conjunction with workshops, conferences and special sections. Papers submitted to workshops or special sessions are invited by the program committee chairs of the workshops/sections. This joint organization aims to collect research results from different perspectives. The following workshops and conferences are being planned:
International Workshop on Visual Languages and Computing
International Workshop on Distance Education Technology

Please contact the conference program co-chairs if you are interested in organizing a workshop or a special session. A one-page proposal with the following items is required:
Title of the workshop/special session
Name of the program committee chair(s)
A list of program committee members
E-mail address of the corresponding program committee chair
A brief description of the theme

Each special session should include at least 5 invited (or accepted) papers. Each workshop should include at least 2 special sessions (10 papers). Paper review process of the workshop/special session will be arranged by the individual program committee chair. It is the responsibility of the individual chairs to ensure the quality of papers in the workshop/special sessions. The schedule of workshops/special sessions will be the same as the main conference (See Important Dates below). Each individual program committee chair can set up the Web site of the workshop/special session. However papers will be printed in the same volume as the main conference.

Information for reviewers

Papers submitted to DMS'11 will be reviewed electronically. The users (webmaster, program chair, reviewers...) can login using the following URL: http://conf.ksi.edu/dms2011/review/pass.php.

If you have any questions please send e-mail to: dms11@ksi.edu.

DMS 2011 Conference Secretariat
Knowledge Systems Institute Graduate School
3420 Main Street
Skokie, IL 60076 USA
Tel: 847-679-3135
Fax: 847-679-3166
E-mail: dms11@ksi.edu

Important dates

Paper submission due: March 30, 2011
Notification of acceptance: June 20, 2011
Camera-ready copy: July 11, 2011
Early conference registration due: July 11, 2011

The conference is listed by:

DBLP
AllConferences.com, a directory of conferences and conventions.

Proceedings of the Sixteenth International Conference on Distributed Multimedia Systems

DMS 2010

Copyright © 2010
Printed by
Knowledge Systems Institute
3420 Main Street
Skokie, Illinois 60076
(847) 679-3135
info@ksi.edu
www.ksi.edu
Printed in USA, 2010
ISBN 1-891706-28-4 (paper)

