

Jiming Liu
Jinglong Wu
Yiyu Yao
Toyoaki Nishida (Eds.)

LNCS 5820

Active Media Technology

5th International Conference, AMT 2009
Beijing, China, October 2009
Proceedings

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Jiming Liu Jinglong Wu Yiyu Yao
Toyoaki Nishida (Eds.)

Active Media Technology

5th International Conference, AMT 2009
Beijing, China, October 22-24, 2009
Proceedings

Volume Editors

Jiming Liu

Department of Computer Science

Hong Kong Baptist University, Kowloon Tong, Hong Kong

E-mail: jiming@comp.hkbu.edu.hk

Jinglong Wu

Graduate School of Natural Science and Technology

Okayama University, Japan

E-mail: wu@mech.okayama-u.ac.jp

Yiyu Yao

Department of Computer Science

University of Regina, Regina, Saskatchewan, Canada,

E-mail: yyao@cs.uregina.ca

Toyoaki Nishida

Dept. of Intelligence Science and Technology

Graduate School of Informatics, Kyoto University, Japan

E-mail: nishida@i.kyoto-u.ac.jp

Library of Congress Control Number: 2009936115

CR Subject Classification (1998): H.5, H.3, H.4.3, I.2.11, D.2.2, C.2.4, I.6

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743

ISBN-10 3-642-04874-9 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-04874-6 Springer Berlin Heidelberg New York

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

springer.com

© Springer-Verlag Berlin Heidelberg 2009

Printed in Germany

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

Printed on acid-free paper SPIN: 12771516 06/3180 5 4 3 2 1 0

Preface

This volume contains the papers selected for presentation at The 2009 International Conference on Active Media Technology (AMT 2009), jointly held with The 2009 International Conference on Brain Informatics (BI 2009), at Beijing University of Technology, China, on October 22–24, 2009. Organized by Web Intelligence Consortium (WIC) and IEEE Computational Intelligence Society Task Force on Brain Informatics (IEEE TF-BI), this conference marked the fifth in the AMT series since its debut conference at Hong Kong Baptist University in 2001 (followed by AMT 2004 in Chongqing, China; AMT 2005 in Kagawa, Japan; and AMT 2006 in Brisbane, Australia).

Over the past years, Active Media Technology (AMT) and its applications have engulfed our daily lives, enhancing connectivity and interactivity in ways never before imaginable; today's examples include Facebook, Twitter, and Google Latitude. At the same time, AMT applications have redefined how business is being conducted by empowering consumer engagement and participation (e.g., ParkatmyHouse.com). Advertisers are tapping into social networks to create new business opportunities (e.g., social media marketing). Intelligent electric grids are enabling better energy-efficient distribution and storage, while fighting climate change (e.g., ecotricity.com and eco-metering).

All these initiatives have once again confirmed our vision back in 2001 to capture and document the evolution unfolding in our digital era. AMT 2009 continued to be a shared forum for researchers and practitioners from diverse fields, such as computer science, information technology, artificial intelligence, Web intelligence, cognitive science, conversational informatics, media engineering, economics, data mining, data and knowledge engineering, intelligent agent technology, human-computer interaction, complex systems and systems science. It offered new insights into the main research challenges and development of AMT by revealing the interplay between the studies of human informatics and the research of informatics on the Web/Internet, and mobile- and wireless-centric intelligent information processing systems.

Here we would like to express our gratitude to all members of the Conference Committee for their instrumental and unfailing support. AMT 2009 had a very exciting program with a number of features, ranging from keynote talks, special sessions, technical sessions, posters, workshops, and social programs. This would not have been possible without the generous dedication of the Program Committee members and the external reviewers in reviewing the papers submitted to AMT 2009, of our keynote speakers, John Anderson of Carnegie Mellon University, Jeffrey M. Bradshaw of Florida Institute for Human and Machine Cognition, Frank van Harmelen of Vrije Universiteit Amsterdam, Lynne Reder of Carnegie Mellon University, Zhongzhi Shi of the Chinese Academy of Sciences, and Zhi-Hua Zhou of Nanjing University, and organizers (Chen Li and Toyooki Nishida)

and invited speakers in the special session on Conversational Informatics, and the special session on Human-Web Interaction, in preparing and presenting their very stimulating talks. We thank them for their strong support.

AMT 2009 could not have taken place without the great team effort of the Local Organizing Committee and the support of the International WIC Institute, Beijing University of Technology. Our special thanks go to Boyuan Fan, Ze Zhang, Zhenyang Lu, Pu Wang, and Jianwu Yang for their enormous efforts in planning and arranging the logistics of the conference from registration/payment handling, venue preparation, accommodation booking, to banquet/social program organization. We would like to thank Shuai Huang, Jiajin Huang, Jian Yang, and Juzhen Dong of the conference support team at the International WIC Institute (WICI), the Knowledge Information Systems Laboratory, Maebashi Institute of Technology, and Web Intelligence Laboratory, Inc. for their dedication and hard work. We are very grateful to the AMT 2009 corporate sponsors: Beijing University of Technology (BJUT), Beijing Municipal Lab of Brain Informatics, the Chinese Society of Radiology, the National Natural Science Foundation of China (NSFC), State Administration of Foreign Experts Affairs, Shanghai Psytech Electronic Technology Co. Ltd, Shenzhen Hanix United, Inc. (Beijing Branch), Beijing JinShangQi Net System Integration Co. Ltd, and Springer Lecture Notes in Computer Science (LNCS/LNAI) for their generous support. Last but not least, we thank Alfred Hofmann of Springer for his help in coordinating the publication of this special volume in an emerging and interdisciplinary research area.

August 2009

Jiming Liu
Jinglong Wu
Yiyu Yao
Toyoaki Nishida

Conference Organization

Conference General Chairs

Jeffrey Bradshaw Institute for Human and Machine Cognition,
USA
Toyoaki Nishida Kyoto University, Japan

Program Chairs

Jiming Liu International WIC Institute,
Beijing University of Technology, China
Hong Kong Baptist University, HK
Jinglong Wu International WIC Institute,
Beijing University of Technology, China
Okayama University, Japan

Workshop Chairs

Yiyu Yao International WIC Institute,
Beijing University of Technology, China
University of Regina, Canada
Runhe Huang Hosei University, Japan

Organizing Chairs

Shengfu Lu International WIC Institute,
Beijing University of Technology, China
Yulin Qin International WIC Institute,
Beijing University of Technology, China

Publicity Chairs

Jian Yang International WIC Institute,
Beijing University of Technology, China
Jiajin Huang International WIC Institute,
Beijing University of Technology, China

WIC Co-chairs/Directors

Ning Zhong Maebashi Institute of Technology, Japan
Jiming Liu Hong Kong Baptist University, HK

Wenbin Li	Shijiazhuang University of Economics, China
Yuefeng Li	Queensland University of Technology, Australia
Mark Looi	Queensland University of Technology, Australia
Wenji Mao	Institute of Automation, CAS, China
Helen Meng	Chinese University of Hong Kong, Hong Kong
Pierre Morizet-Mahoudeaux	University of Technology of Compiègne, France
Sung Hyon Myaeng	KAIST, Korea
Yoshihiro Okada	Kyushu University, Japan
Eugene Santos	University of Connecticut, USA
Xijin Tang	Academy of Mathematics and Systems Sciences, CAS, China
Ruizhi Wang	Tongji University, China
Jian Yang	Beijing University of Technology, China
Tetsuya Yoshida	Hokkaido University, Japan
Mengjie Zhang	Victoria University of Wellington, New Zealand
Zili Zhang	Southwest University, China
William Zhu	University of Electronic Science and Technology, China

Table of Contents

Keynote Talks

Using Neural Imaging to Inform the Instruction of Mathematics	1
<i>John Anderson</i>	
Distributed Human-Machine Systems: Progress and Prospects	2
<i>Jeffrey M. Bradshaw</i>	
Large Scale Reasoning on the Semantic Web: What to Do When Success Is Becoming a Problem	3
<i>Frank van Harmelen</i>	
How Midazolam Can Help Us Understand Human Memory: 3 Illustrations and a Proposal for a New Methodology	4
<i>Lynne Reder</i>	
Research on Brain-Like Computer	5
<i>Zhongzhi Shi</i>	
A Framework for Machine Learning with Ambiguous Objects	6
<i>Zhi-Hua Zhou</i>	

Special Session on Conversational Informatics

Implementing a Multi-user Tour Guide System with an Embodied Conversational Agent	7
<i>Aleksandra Čereković, Hsuan-Hung Huang, Takuya Furukawa, Yuji Yamaoka, Igor S. Pandžić, Toyooki Nishida, and Yukiko Nakano</i>	
Actively Adaptive Agent for Human-Agent Collaborative Task	19
<i>Yong Xu, Yoshimasa Ohmoto, Kazuhiro Ueda, Takanori Komatsu, Takeshi Okadome, Koji Kamei, Shogo Okada, Yasuyuki Sumi, and Toyooki Nishida</i>	
Low-Overhead 3D Items Drawing Engine for Communicating Situated Knowledge	31
<i>Loic Merckel and Toyooki Nishida</i>	
A Method to Detect Lies in Free Communication Using Diverse Nonverbal Information: Towards an Attentive Agent	42
<i>Yoshimasa Ohmoto, Kazuhiro Ueda, and Takehiko Ohno</i>	
An Integrative Agent Model for Adaptive Human-Aware Presentation of Information during Demanding Tasks	54
<i>Andy van der Mee, Nataliya M. Mogles, and Jan Treur</i>	

Special Session on Human-Web Interaction

Consumer Decision Making in Knowledge-Based Recommendation 69
Monika Mandl, Alexander Felfernig, and Monika Schubert

Incremental Learning of Triadic PLSA for Collaborative Filtering 81
Hu Wu and Yongji Wang

Interactive Storyboard: Animated Story Creation on Touch
 Interfaces 93
Kun Yu, Hao Wang, Chang Liu, and Jianwei Niu

Comparative Evaluation of Reliabilities on Semantic Search Functions:
 Auto-complete and Entity-Centric Unified Search 104
Hanmin Jung, Mi-Kyoung Lee, Beom-Jong You, and Do-Wan Kim

Integrated Recommender Systems Based on Ontology and Usage
 Mining 114
Liang Wei and Song Lei

Active Media Retrieval and Sharing

Knowledge-Based Concept Score Fusion for Multimedia Retrieval 126
Manolis Falelakis, Lazaros Karydas, and Anastasios Delopoulos

Example-Based Query Analysis Using Functional Conceptual
 Graphs 136
Hui Liu and Yuquan Chen

Checking Satisfactions of XML Referential Integrity Constraints 148
Md. Sumon Shahriar and Jixue Liu

A Verification Method of Hyponymy between Chinese Terms Based on
 Concept Space 160
Lei Liu, Sen Zhang, Lu Hong Diao, Shu Ying Yan, and Cun Gen Cao

Sharing Mobile Multimedia Annotations to Support Inquiry-Based
 Learning Using MobiTOP 171
*Khasfariyati Razikin, Dion Hoe-Lian Goh, Yin-Leng Theng,
 Quang Minh Nguyen, Thi Nhu Quynh Kim, Ee-Peng Lim,
 Chew Hung Chang, Kalyani Chatterjea, and Aixin Sun*

Understanding Perceived Gratifications for Mobile Content Sharing
 and Retrieval in a Game-Based Environment 183
*Chei Sian Lee, Dion Hoe-Lian Goh, Alton Y.K. Chua, and
 Rebecca P. Ang*

Why We Share: A Study of Motivations for Mobile Media Sharing 195
*Dion Hoe-Lian Goh, Rebecca P. Ang, Alton Y.K. Chua, and
 Chei Sian Lee*

Active Support Systems and Intelligent Interfaces

The Layout of Web Pages: A Study on the Relation between Information Forms and Locations Using Eye-Tracking	207
<i>Mi Li, Yangyang Song, Shengfu Lu, and Ning Zhong</i>	
Human Characteristics on Length Perception with Three Fingers for Tactile Intelligent Interfaces	217
<i>Haibo Wang, Jinglong Wu, and Satoshi Takahashi</i>	
A Model and Environment for Improving Multimedia Intensive Reading Practices	226
<i>Thomas Bottini, Pierre Morizet-Mahoudeaux, and Bruno Bachimont</i>	
Study on Adaptive Computer-Assisted Instruction for In-Service Training	238
<i>Yu-Teng Chang, Chih-Yao Lo, and Ping-Chang Chen</i>	
Research on Recreational Sports Instruction Using an Expert System . . .	250
<i>Chih-Yao Lo, Hsin-I Chang, and Yu-Teng Chang</i>	

Smart Digital Media

Using 6LowPAN UPnP and OSGi to Implement Adaptable Home Ambient Intelligence Network Platform	263
<i>Zhang Hui-bing and Zhang Jing-wei</i>	
Low Frequency Domain Aided Texture Synthesis for Intra Prediction . . .	273
<i>Xiaowei Sun, Baocai Yin, and Yunhui Shi</i>	
A Novel Geometry Image Coding	281
<i>Yunhui Shi, Wen Wen, Baocai Yin, and Jijun Shi</i>	
Musical Style Classification Using Low-Level Features	288
<i>Armando Buzzanca, Giovanna Castellano, and Anna Maria Fanelli</i>	

Multi-agent Systems and Autonomy-Oriented Computing

Enterprise Cluster Dynamics and Innovation Diffusion: A New Scientific Approach	299
<i>Marco Remondino, Marco Pironi, and Paola Pisano</i>	
A Novel Application of Organic Plant Farming Analysis System – Using Game Theory and Multi-Agent Technique	311
<i>Chih-Yao Lo and Yu-Teng Chang</i>	
A Dynamic Trust Network for Autonomy-Oriented Partner Finding	323
<i>Hongjun Qiu, Jiming Liu, and Ning Zhong</i>	

Modeling an Educational Multi-Agent System in MaSE 335
Izabela Salotti Braga Gago, Vera M.B. Werneck, and Rosa M. Costa

Enhancing Decentralized MAS-Based Framework for Composite Web Services Orchestration and Exception Handling by Means of Mobile Agents Technology 347
Mounira Ilahi, Zaki Brahmi, and Mohamed Mohsen Gammoudi

Multi-objective Analysis on Optimization of Negotiation Support 357
Yu-Teng Chang, Chih-Yao Lo, Ping-Chang Chen, and Shu-Huei Han

Data Mining and Ontology Mining in Active Media

Rough Set Based Personalized Recommendation in Mobile Commerce 370
Lei Shi, Li Zhang, Xinming Ma, and Xiaohong Hu

SpamTerminator: A Personal Anti-spam Add-In for Outlook 376
Wenbin Li, Yiyang Cheng, Ning Zhong, TaiFeng Liu, and Xindong Zhang

Classifying Images with Image and Text Search Clickthrough Data 385
Gavin Smith, Michael Antunovic, and Helen Ashman

A Novel Fast Inter Mode Decision Algorithm in H.264/AVC for Forest Fire Prevention Surveillance 397
Chen Chen, Ning Han, Chunlian Yao, and Yuan Li

A Method for Analyzing Software Faults Based on Mining Outliers' Feature Attribute Sets 409
Jiadong Ren, Changzhen Hu, Kunsheng Wang, and Di Wu

Web Intelligence

Unifying Web-Scale Search and Reasoning from the Viewpoint of Granularity 418
Yi Zeng, Yan Wang, Zhisheng Huang, and Ning Zhong

The Quest for Parallel Reasoning on the Semantic Web 430
Peiqiang Li, Yi Zeng, Spyros Kotoulas, Jacopo Urbani, and Ning Zhong

A Model for Personalized Web-Scale Case Base Maintenance 442
Jingyu Sun, Xueli Yu, Ruizhi Wang, and Ning Zhong

X3D-Based Web 3D Resources Integration and Reediting 454
Zhoufan Zhou, Hisao Utsumi, and Yuzuru Tanaka

Providing Relevant Answers for Queries over E-Commerce Web Databases	467
<i>Xin Li, Jun Zhang, and Liping Li</i>	
Detecting News Event from a Citizen Journalism Website Using Tags . . .	478
<i>Alton Y.K. Chua, Dion Hoe-Lian Goh, and Khasfariyati Razikin</i>	
Networks and Security	
A New Mechanism for Job Scheduling in Computational Grid Network Environments	490
<i>Nandagopal Malarvizhi and V. Rhymend Uthariaraj</i>	
Efficient and Provably Secure Self-certified Signature Scheme	501
<i>Jianhong Zhang, Hua Chen, and Qin Geng</i>	
A Reversible Watermarking Scheme for 3D Meshes	513
<i>Dan Wu and Guozhao Wang</i>	
Neighbor-List Based Pairwise Key Management Scheme in Wireless Sensor Networks	522
<i>Xing Zhang, Jingsha He, and Qian Wei</i>	
Author Index	529

Using Neural Imaging to Inform the Instruction of Mathematics

John Anderson

Department of Psychology
Carnegie Mellon University, USA
ja+@cmu.edu

I will describe research using fMRI to track the learning of mathematics with a computer-based algebra tutor. I will describe the methodological challenges in studying such a complex task and how we use cognitive models in the ACT-R architecture to interpret imaging data. I will also describe how we can use the imaging data to identify mental states as the student is engaged in algebraic problems solving.

Distributed Human-Machine Systems: Progress and Prospects

Jeffrey M. Bradshaw

Florida Institute for Human and Machine Cognition, USA
jbradshaw@ihmc.us

Advances in neurophysiological and cognitive science research have fueled a surge of research aimed at more effectively combining human and machine capabilities. In this talk we will give an overview of progress and prospects for four current thrusts of technology development resulting from this research: brain-machine interfaces, robotic prostheses and orthotics, cognitive and sensory prostheses, and software and robotic assistants. Following the overview, we will highlight the unprecedented social ethics issues that arise in the design and deployment of such technologies, and how they might be responsibly considered and addressed.

Large Scale Reasoning on the Semantic Web: What to Do When Success Is Becoming a Problem

Frank van Harmelen

AI Department
Vrije Universiteit Amsterdam, The Netherland
`Frank.van.Harmelen@cs.vu.nl`

In recent years, the Semantic Web has seen rapid growth in size (many billions of facts and rules are now available) and increasing adoption in many sectors (government, publishing industry, media). This success has brought with it a whole new set of problems: storage, querying and reasoning with billions of facts and rules that are distributed across different locations. The Large Knowledge Collider (LarKC) is providing an infrastructure to solve such problems. LarKC exploits parallelisation, distribution and approximation to enable Semantic Web reasoning at arbitrary scale. In this presentation we will describe the architecture and implementation of the Large Knowledge Collider, we will give data on its current performance, and we will describe a number of use-cases that are deploying LarKC.

How Midazolam Can Help Us Understand Human Memory: 3 Illustrations and a Proposal for a New Methodology

Lynne Reder

Department of Psychology
Carnegie Mellon University, USA
reder@cmu.edu

Midazolam is a benzodiazepine commonly used as an anxiolytic in surgery. A useful attribute of this drug is that it creates temporary, reversible, anterograde amnesia. Studies involving healthy subjects given midazolam in one session and saline in another, in a double-blind, cross-over design, provide insights into memory function. Several experiments will be described to illustrate the potential of studying subjects with transient anterograde amnesia. This talk will also outline how this drug can be used in combination with fMRI to provide more insights about brain functioning than either method in isolation.

Research on Brain-Like Computer

Zhongzhi Shi

Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
Beijing 100190, China
shizz@ics.ict.ac.cn

After more than 60 years of development, the operation speed of computer is up to several hundred thousand billion (10^{14}) times, but its intelligence level is extremely low. Studying machine which combines high performance and the people's high intelligence together becomes the effective way with high capacity and efficiency of exploring information processing. It will bring the important impetus to economic and social sustainable development, promotion of the information industry and so on to make breakthrough in the research of brain-like computer.

Mind is all mankind's spiritual activities, including emotion, will, perception, consciousness, representation, learning, memory, thinking, intuition, etc. Mind model is for explaining what individuals operate in the cognitive process for some thing in the real world. It is the internal sign or representation for external realistic world. If the neural network is a hardware of the brain system, then the mind model is the software of the brain system. The key idea in cognitive computing is to set up the mind model of the brain system, and then building brain-like computer in engineering through structure, dynamics, function and behavioral reverse engineering of the brain. This talk will introduce the research progress of brain-like computer, mainly containing intelligence science, mind models, neural columns, architecture of brain-like computers.

Intelligence Science is an interdisciplinary subject which dedicates to joint research on basic theory and technology of intelligence by brain science, cognitive science, artificial intelligence and others. Brain science explores the essence of brain, research on the principle and model of natural intelligence in molecular, cell and behavior level. Cognitive science studies human mental activity, such as perception, learning, memory, thinking, consciousness etc. In order to implement machine intelligence, artificial intelligence attempts simulation, extension and expansion of human intelligence using artificial methodology and technology. Research scientists coming from above three disciplines work together to explore new concept, new theory, new methodology. Intelligence science is a essential way to reach the human-level intelligence and point out the basic principles for brain-like computer.

A Framework for Machine Learning with Ambiguous Objects

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@nju.edu.cn

Machine learning tries to improve the performance of the system automatically by learning from experiences, e.g., objects or events given to the system as training samples. Generally, each object is represented by an instance (or feature vector) and is associated with a class label indicating the semantic meaning of that object. For ambiguous objects which have multiple semantic meanings, traditional machine learning frameworks may be less powerful. This talk will introduce a new framework for machine learning with ambiguous objects.

Implementing a Multi-user Tour Guide System with an Embodied Conversational Agent

Aleksandra Čereković¹, Hsuan-Hung Huang², Takuya Furukawa², Yuji Yamaoka³,
Igor S. Pandžić¹, Toyoaki Nishida², and Yukiko Nakano³

¹ Faculty of Electrical Engineering and Computing, Zagreb, Croatia

² Graduate School Informatics, Kyoto University, Japan

³ Dept. of Computer and Information Science, Faculty of Science and
Technology - Seikei University, Tokyo, Japan

{aleksandra.cerekovic, igor.pandzic}@fer.hr,
{huang, furukawa, nishida}@ii.ist.i.kyoto-u.ac.jp,
{50007646208, y.nakano}@st.seikei.ac.jp

Abstract. In recent years, computer-generated interactive virtual characters, called Embodied Conversational Agents (ECAs), are subjects of considerable ongoing research. Nevertheless, their conversational abilities are mediocre compared to real human behaviors. Among limitations, most of ECAs are incapable of participating in natural conversations in which the number of participants can change dynamically. In the ongoing work we investigate principles of integrating a multi-user support in an ECA system. We present experiments and implementation approach of a prototype system in which a tour guide ECA interacts with one or two users. The system combines different technologies to detect and address the system users and draw their attention. Experimental interaction with the system produces encouraging results. The system can address the user's appearance, departure, decreased level of interest and identify his conversational role.

1 Introduction

In recent years, computer-generated interactive virtual characters, called Embodied Conversation Agents (ECAs) [14], are subjects of considerable ongoing research. ECAs have potential to behave like humans, and thus an opportunity to achieve naturalness in human-machine interaction. It has been proven that the presence of ECAs can improve a human's engagement and significantly increase his positive perception of learning experiences [10].

Nevertheless, conversational abilities of present-day ECAs are mediocre compared to real human behaviors. Among other limitations, most ECAs are incapable of participating in natural conversations in which the number of participants can change dynamically. In the theory of multi-party human-agent communication there are several bases which discuss multi-party dialogue issues and human's nonverbal conversational signals during interaction. Still, multi-user ECA systems are experimental and scarce. Nowadays technology is too immature to overcome challenges of bringing multi-party support to human-agent conversational systems; e.g. requirements to track

and update user state, detect and resolve their requests in real-time, handle complex conversations...

In the ongoing work we investigate principles of integrating multi-user support in an ECA system. We present implementation approach of a prototype system in which a tour guide ECA interacts with one or two users. The system solves some of the multi-party dialogue issues presented in the next section. It can detect a user's arrival, departure, decreased level of interest and identify his conversational role. It can also recover from system failures, e.g. failure of speech recognition.

The paper is organized as follows. First, we present related work and discuss key challenges and the conversation model. The system design and architecture are explained in the fourth section. We conclude the paper with a summary of future research directions and brief discussion.

2 Related Work

In the theory of multi-party dialogue interaction between agents and humans, most of the work has been done by Traum [1] [9] [12]. For the purposes of the Mission Rehearsal Exercise (MRE) Project [12], Traum investigated multi-party dialogue issues and grouped them into three parts [9]:

- **Participants' roles.** The issue refers to identification of participants' local roles and responsibilities which shift during interaction (who is addressee, listener, and speaker). In addition to that, it also refers to participant's social roles and their effect on interaction behaviors (e.g. status, relationship).
- **Interaction management.** Managing the communication flow in a multi-party system is far more difficult than in a dyadic system. Some of the difficult issues are how to give and recognize a turn and how to handle participants' channels (and backchannels). Besides, conversations can be easily split and merged together and attention can be paid to several persons.
- **Grounding and obligations.** are notions commonly used to model local state of dialogue. In multi-party communication usage of these models can become very complex; e.g. if there are more addressees, it can become unclear what a proper grounding should be.

Traum's paper only discusses issues which arise in multi-party human-agent systems, but does not give practical solutions to it.

Vertegaal et al. [7] focus on human gaze behaviors in multi-party conversation. They studied a real three-party conversation and discovered that the subjects look about seven times more at the individual they listen to than at others, and about three times more at an individual they speak to than at others. The conclusion is that gaze is an excellent predictor of conversational attention in multi-party conversations.

Similar to the functionality of our system, Rehm and Wissner [8] developed a gambling system in which an ECA plays a dice game with two humans. As game rules define conversational turns, their system has a simple dialogue model in which turn-taking mechanism and participants' roles are round-based. The system lacks the active gaze model which follows human users. In further study, Rehm and Andre [11] investigate human gazing patterns in interaction with a real human and an agent. They

note that people spend more time looking at an agent that is addressing them than at a human speaker. This phenomenon can be explained by the fact that prolonged eye contact in a social interaction can be considered impolite and rude; hence, the agent in this game may have been regarded as an artifact rather than a human being.

3 Key Features and Initial Experiments

In addition to the findings from multi-party dialogue theory, at the very beginning of development we identified key features which a multi-user ECA system should possess. Afterwards, we proceeded with several experiments with the system equipment to determine its final configuration, which is shown in Figure 1.

- **Speech processing.** If the system users are standing next to each other the microphones should be localized to avoid an overlap between their verbal channels. Conducted experiments showed that Loquendo ASR [18] meets our needs very adequately; it is stable in noisy environments, speaker-independent, there is no significant speech overlap at reasonable distance (0,5 meters approx.), and has keyword spotting feature absent from the free Microsoft SAPI engine. However, Loquendo sometimes reacts to voices which are not from the users. We think such errors can be prevented by using the Speaker Verification function, which we have not yet tested.
- **Nonverbal data acquisition.** During interaction it is extremely important to detect and locate the users so the ECA can gaze at them. For that purpose we installed two cameras and used image processing techniques to detect the user's arrival, position and departure.

To process image from one camera we decided to use Okao library [16] which provides accurate face detection and extra features such as face orientation, gaze direction, the positions and openness of eyes and mouth, gender detection, age identification and face identification from a single image. In preliminary tests with a 960x720@15fps web cam, accuracy of face detection was sufficiently high and undoubtedly usable, but most of the other functions were not reliable and could only be treated as an extra bonus. Since Okao Vision does not require stereo cameras, this result was acceptable for our system.

The second camera recognizes moving objects and detects the user's arrival and departure by comparing the differences between sequential images coming from the camera's input.

- **The conversation model.** The final conversation model we designed simplifies participants' conversational roles and is based on narration. In the system scenario, an ECA named Dubravka takes visitors on a tour of the city of Dubrovnik and talks about its cultural heritage, history, and monuments. She maintains the initiative in the conversation, though users can interrupt her and ask questions about the current topic. We predict topic-related questions and define it by using specific keywords in the speech recognition engine (such as "where", "when", "how"). To hold the users' attention during the session, Dubravka also asks the users simple "yes/no" questions.
- **ECA appearance.** The users should perceive the ECA's gaze direction, which has proven to be very important in regulation of conversation flow in multi-party

communication. Since the system environment is displayed on a 2D screen, the ECA's size and an appropriate gaze model are important to avoid the Mona Lisa effect (an impression perceived with 2D pictures of humans, that "the eyes follow the observer across the room"). The ECA's final position on the screen was derived from initial experiments with the ECA's size and its attention towards the users. We concluded that users can distinguish the direction in which the ECA is gazing only if the ECA's size on the screen is very large.

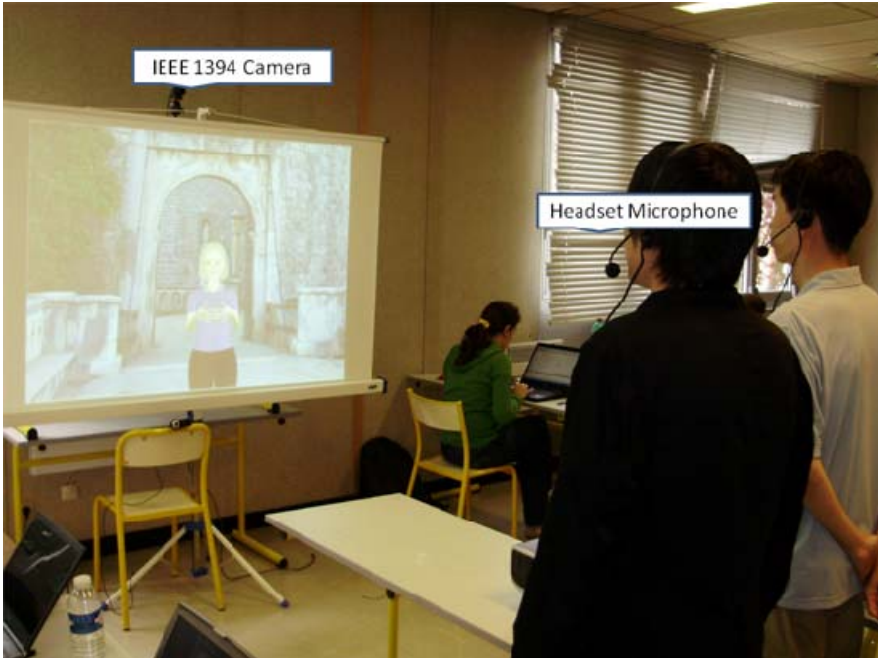


Fig. 1. Final system configuration

4 The Conversation Model

4.1 Multi-party Features

The conversation model is designed to handle situations typical for multi-party interaction. When designing the model we predicted specific situations, such as speech collision and mutual conversation between the users, and defined the system's response. We discuss multi-party features of the conversation model and system with regard to issues identified in Traum's paper [9]:

- **The user's appearance.** During interaction, the system can dynamically detect the user's arrival, his position in front of the system and his departure. In its initial state, Dubravka waits for someone to appear. When that happens, she approaches and engages the potential user. In case that user departs, she turns back and reverts to the

initial state. At most two users can stand in front of the system. If there are less than two users, Dubravka can also invite another person to join the conversation.

- **Channel management.** The system combines users' verbal and nonverbal channels to resolve their utterances. Nonverbal behaviors taken into account are face orientation and gazing. By combining those modalities the system is able to define one of the following situations: decreased level of attention, making requests of the system, departure, speech collision.
- **Speech collision.** During interaction it might happen that the users ask questions simultaneously. The system handles this situation by having the ECA address one of the speakers and give him turn, e.g. Dubravka says "You can speak one by one (gesture "calm down"), I'll respond to you both. You can be the first one (addressing one user by pointing with an open hand)"
- **Identification of conversational roles.** Our conversation model simplifies the rules of how participants' local roles shift during the session. System scenario defines the communication workflow and several specific situations:
 1. When talking about cultural heritage of the city, Dubravka gazes at both users with the same frequency. In the case when one of the users asks a question, the system identifies him by using a localized microphone, and when Dubravka responds to him, he becomes an addressee, and the other user becomes an overhearer. Following the findings from studies on gazing [7, 11] we developed a computational model of gaze in which the agent gazes at the addressee more than he does at the overhearer.
 2. During the session, Dubravka may ask the users simple questions, e.g. "Is this your first time in Dubrovnik?" In this case she waits for both reactions and responds to each user separately. The reason why we added a questionnaire into the story is to hold the users' interest in the session.
 3. As described, most of the time the agent is the speaker and one or both users are addressees. Unfortunately, natural conversation between three parties in which each individual can become speaker, listener or addressee is not handled in a natural way. Since our agent is capable of understanding only simple questions related to the current topic, conversation between users is not welcome.

4.2 Additional Features

We are also concerned with features which can make interaction with the system more fluid. It is important to recover the system from failures, such as failure of speech recognition. In the case when a user's speech is not recognized, we propose two-stage recovery. First Dubravka asks the user to repeat his question. If his speech is not identified the second time around, she responds to him: "I'm sorry. I don't know a response to your question".

We also use Okao vision to recognize a situation in which the user is paying less attention to the session. If this occurs, Dubravka speaks until the end of the planned utterance, turns to the user and says "Seems you are not interested in this topic. Would you like to hear something else?"

5 System Overview

The system is constructed from several computers and hardware equipment which communicate through TCP/IP-based OpenAir routing protocol [20]. Implementation of the OpenAIR server and plugs is used to set up the GECA Framework, a software framework which enables rapid integration of different ECA modules. Apart from the server and plugs, GECA framework also defines communication rules between ECA components. More details about GECA can be found in the work [5].

5.1 System Design

System design is depicted in Figure 2, which depicts the software components connected to the GECA platform/server. The components communicate through the server using GECA messages. As part of the message content we introduce a variable system which describes interaction between the agent and humans. For example, *SpeechInput* represents the most recent result from one of the speech recognition components, *Speaker* represents the id of the speech recognition component, *UserNumber* represents the number of users who are standing in the user area, *UserStatus* represents the availability of the users, *UserAttention* represents how much the users are paying attention to the system, *Addressee* specifies the addressee of the agent's next utterance, etc. During the session, depending on the situation in the environment, the system components update the values of variables and exchange data to define a final response.

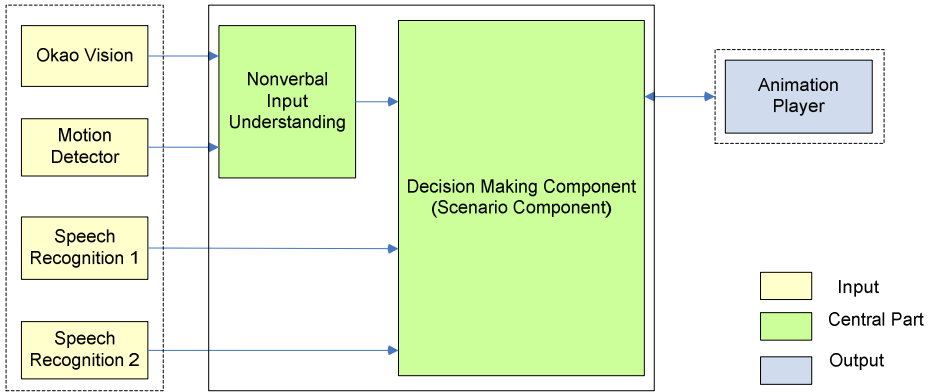


Fig. 2. Design of the system components. Blue arrows represent message workflow.

5.2 System Components

The system components can be roughly divided into three categories: input, central part, and system output.

System Input

Input components detect and process human speech, appearance and face orientation. Combination of this data maps to one of several distinct situations (Table 1) in the system environment.

Table 1. Situations in the system environment and explanations how each situation is detected

Situation/Feature	Situation Description	Components which detect situation	Explanation how components work
User's arrival	System is in idle state until it detects potential users	Motion Detector	Motion Detector detects movements in the area which activate the system.
Number of users	During interaction system tracks the number of users	Okao Vision & Motion Detector	Okao vision detects number of faces. If it fails, Motion Detector processes users' areas on the image and detects motions in each area.
Interruptions	User interrupts Dubravka to make a request	Speech Recognition	For each utterance we predict topic-related questions defined with keywords as "how", "where"...
Unrecognized speech	Subtype of interruption; request is not recognized	Speech Recognition	Speech is detected, but no keywords are triggered. Loquendo SAPI has additional features for this purpose
Decreased level of attention	Users are not paying attention to the system, e.g. they gazing around	Okao Vision & Motion Detector	Okao Vision processes user's facial orientation and Motion Detector handles their movements.
Leaving the system	User(s) is(are) leaving the system	Speech Recognition, Okao Vision, Motion Detector	User can say "bye, bye" or "I'm leaving" which triggers speech recognition component. If there are no motions or faces on the input images, users have departed.

Speech recognition. Component is based on Loquendo ASR [18]. Once the component is started, it awaits users' speech input. When it receives a speech start signal, it starts recognizing the speech. The results of recognition, timing and speaker's id, are then sent to the server and passed to Decision Making Component.

Nonverbal Behaviors. To detect user's nonverbal behaviors we developed two input components: Okao Vision and Motion Detector.

- **Okao Vision** component is based on Okao Vision library [16]. The system uses the following features of Okao: face detection, face orientation tracking and eye-mouth openness. Face orientation is used to approximately determine the users' gaze direction. This is not very accurate but should be sufficient when we only need to distinguish rough directions like the screen or another user. Results of the component are sent to Nonverbal Input Understanding component through server.
- **Motion Detector**, the component which we use to detect moving objects in distinct areas, is based on OpenCV library [17]. The component divides the viewed image region coming from the camera into two distinct areas (which we refer to as *user areas*) and detects motion inside each area. The user areas are surrounded with a blue and a red square, respectively. To recognize moving objects we calculate a

sum of pixels in one frame and compare this value to the value from the previous frame. If threshold is exceeded, moving objects are detected.

Central Part of the System

Nonverbal Input Understanding component uses information received from users' nonverbal channels to detect some of the situations defined in the system scenario.

First, it combines input data coming from Motion Detector and Okao Vision and uses simple heuristic methods to resolve the number of users. Okao Vision fails in detection when users rotate their head beyond 60 degrees, so it is not sufficient to track the number of users. However, we can determine the user's presence by detecting movement in the user area with Motion Detector. For example, during the system session two users listen to the agent and the left user turns his head to see who entered the room behind him. In this situation Okao Vision sets `UserNumber` variable to one, as if there is just the right user in the system, and sends it to the Understanding component. At the same time, Motion Detector detects motions in the left area and notifies the Understanding component. It then sets `UserNumber` value to two and sends it to the Decision Making Component.

Nonverbal input understanding component also tracks users' attention. We identify two meaningful patterns the users tend to look at for a significant proportion of time – the agent and the other user. Okao Vision is used to detect face orientation which we assume is the user's gaze direction. Since two patterns (agent, other user) are placed in different directions this approach is satisfying to efficiently track the user's attention. By combining results of Okao Vision and Motion Detector, the system can smoothly respond to situations in which a user is not interested in the system anymore.

Decision Making Component. This component is implemented based on Information State Theory [2] [3], which is a general theory of human-agent dialogues. A dialogue based upon this theory is defined as a set of variables (or information) that describe the current state of the dialogue. The implemented prototype is based on information state dialogue move engine [4] and capable of handling multi-modal, multi-party conversations, dynamically changing behaviors accompanying the emotion dynamics simulating component, etc.

To support the concepts proposed in the theory, we developed a script language [5] based on AIML [22]. In AIML, possible human-agent interactions are defined as one-to-one pairs. Compared to AIML which merely matches recognized speech inputs and non-verbal inputs with predefined patterns, we introduce a variable system to describe interaction between the agent and humans. Values of these variables are updated with the agent system's internal status and perception events sent from the speech recognition and nonverbal input interpretation components. Script designers can also specify variable update logic as effects of particular input patterns. `Effect` element is introduced into the `Template` element for this purpose. An input event can cause the values of particular variables to be bound to, added with, or subtracted from certain values.

The pattern syntax is also extended. Predicate element is introduced to represent a test of variable values. It is possible to test if the value of a variable is equal to, lesser or greater than another value.

The chatbot-like ECA system is then extended to a more powerful rule-based autonomous system. The agent or the script execution engine updates its internal status variables based on perception of the outside world or users, and picks for execution the first valid template for which all conditions (predicates) evaluate as true. Therefore, rules such as the one that specifies what the agent should do when a user appears in or departs from a user area can be specified in scripts.

States limit possible patterns that will be used in matching in the current conversational situation and thus isolate interference from other states which may happen to have the same triggering patterns. Due to the absence of a context management mechanism in the agent's behavior control, there is no way to determine whether a user's answer is related to the last question asked by the agent. However, for example, when the agent is going to ask a yes/no question such as "Do you need a tour guide?", transition to a specific state corresponding to the question can isolate it from other yes/no questions.

`GlobalState` is introduced for error and interruption handling. When a failed or unknown recognition occurs, appropriate response will be sought among the categories defined in the global state. When interruptions from the user such as "excuse me" or "pardon" occur, they are also matched against the patterns defined in this state.

The disadvantage of this approach is that, in absence of a full dialogue-managing central component, the agent does not conduct a plan that contains multiple steps to achieve a certain goal. The agent's behaviors are driven by the events that occurred in the outside world. The management mechanism for information like grounding or topics is not included in the script execution kernel.

In this version of the system which is still a work in progress, only a few issues specific to multi-party conversation are addressed. The gaze direction of the agent, which is essential in three-party dialogue, is not controlled by the central component but by the player. Control is done by using the `Addressee` attribute introduced in the `Utterance` element to specify if the addressee of an utterance is located to the left, right or both. Addressee specification is done by the rules in the scenario script by the script programmer; e.g. a specific utterance of the agent can be directed at the last speaker, and so on.

System Output

Animation Player. The role of the Animation Player is to produce a final output which represents the response to situations in the system environment. It displays the city of Dubrovnik where the tour guide ECA Dubravka stands and communicates with the users. During the session she walks through the city, shows them landmarks and reacts to their interruptions, questions and departures.

The animation system is built to the following requirements:

- **Virtual environment for the ECA.** The Player is based on `visageSDK`, an MPEG-4 compatible character animation framework [15]. As this product does not have support for virtual environments, the player uses 2D images of the city of Dubrovnik as background. Position of the agent is calculated using `ARToolkit` software, which tracks a marker to position the ECA [6]. Final output gives an impression of the agent standing within the city.

- **Ability to run multimodal behaviors described with GSML syntax.** During the session Animation Player receives messages from the Decision Making component. Messages contain GSML description of the ECA's multimodal behavior which should be run in real time. For that purpose we developed a thread-based mechanism which schedules and synchronizes running nonverbal behaviors with speech. The agent's nonverbal behaviors are produced and tested with GSML syntax before the system is integrated. Animations for the agent are either modeled manually in 3ds Max or procedurally in Visage. At the moment the system uses 25 modeled hand gestures in total (points, beats, symbolic). Procedural animations are head movements (nod, shake, tilt), facial expressions (joy, anger, fear, surprise, sadness), gazing and walking.
- **Responding to interruptions and departures.** The Animation Player should have smart scheduling so it can immediately respond to situations from the environment, even if it is running behaviors for the agent. We created a simple scheduler which immediately stops running behaviors when the Decision Making component sends a `Stop` message and resets the Player to initial state when users leave.
- **Gaze at the user.** Gazing model runs on the utterance level and is controlled by the `Addressee` attribute. E.g. if the addressee is the left user, the agent will gaze first at the left side, and then it will glance to the right for a while and gaze back to the left user again. Since we cannot predict the duration of the utterance, we repeat the patterns and stop when the utterance is finished.

6 Conclusions and Future Work

In this paper we presented an implementation approach of a prototype system in which an Embodied Conversational Agent communicates with two human users. Although no system evaluation has yet been performed, the initial results are encouraging. The system can correctly detect the user's arrival and departure, distinguish local roles of conversation participants and use ECA's gazing direction to specify an addressee. It can react to interruptions such as user requests and it can shut down automatically after users' departure. The implemented features execute smoothly, except when the user is leaving, when a certain delay is observed.

As future directions we plan to address two implementation additions which should make the interaction with the system more fluent. Limitations of Speech Recognition component make the agent capable of understanding only topic-related questions. As the agent cannot process discussion between the users, we aim to detect this situation and handle it in a polite way. If this happens, the agent will try to draw attention and continue narration about current topic. For example: "(waving) Alright, let's continue with the tour." Furthermore, we aim to extend the system to dynamically invite observers to join the session, which should make the interaction interesting. Example is the situation when one user leaves and there is one observer standing in the background. In this situation, the agent will look at the observer and say "We have one free place, please come closer and join the tour". This feature demands more complex image processing than what we currently perform on our camera input. After we implement these features we intend to perform system evaluation. In particular we are interested in the level of the interaction naturalness and how it can be increased.

In parallel, to make the system components open to the ECA community, we are upgrading the Animation Player with support for Behavior Markup Language (BML) [13][21][20]. BML is a language for describing physical realizations of multimodal human behaviors and it appears to have been well-received by the research community.

The weakest point in the system is the rather simple conversation model which limits the ECA's abilities to respond to questions we did not take into account. Nevertheless, we find this work useful for further studies and experiments in the field of multi-party interaction between ECAs and humans.

Acknowledgement

This research is partially supported by the Ministry of Education, Science, Sports and Culture of Japan, Grant-in-Aid for Scientific Research (S), 19100001, 2007, "Studies on Construction and Utilization of a Common Platform for Embodied Conversational Agent Research" and the Ministry of Science Education and Sports of the Republic of Croatia, grant nr. 036-0362027-2028 "Embodied Conversational Agents for Services in Networked and Mobile Environments." We would also like to thank Dr. Lao Shihong from OMRON Corporation Research & Development Headquarters for licensing the OKAO Vision.

References

- [1] Traum, D., Rickel, J.: Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. In: AAMAS 2002, vol. 2
- [2] Traum, D., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., Poesio, M.: A model of dialogue moves and information state revision (1999)
- [3] Larsson, S., Berman, A., Gronqvist, L., Kronlid, F.: TRINDIKIT 3.0 Manual. Trindi Deliverable D6.4 (2002)
- [4] The MITRE Corporation: Midiki User's Manual, version 0.1.3 beta edition (2005)
- [5] Huang, H., Cerekovic, A., Pandzic, I., Nakano, Y., Nishida, T.: The Design of a Generic Framework for Integrating ECA Components. In: Proceedings of 7th International Conference of Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 2008, pp. 128–135 (2008)
- [6] Huang, H., Cerekovic, A., Tarasenko, K., Levacic, V., Zoric, G., Pandzic, I., Nakano, Y., Nishida, T.: An Agent Based Multicultural Tour Guide System with Nonverbal User Interface. The International Journal on Multimodal Interfaces 1(1), 41–48 (2007)
- [7] Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2001, Seattle, Washington, United States, pp. 301–308. ACM, New York (2001)
- [8] Rehm, M., André, E., Wissner, M.: Gamble v2.0: social interactions with multiple users. In: Proceedings of the Fourth international Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2005, The Netherlands, July 25 - 29 (2005)
- [9] Traum, D.: Issues in multiparty dialogues. In: Dignum, F.P.M. (ed.) ACL 2003. LNCS (LNAI), vol. 2922, pp. 201–211. Springer, Heidelberg (2004)

- [10] Lester, J.C., Converse, S.A.: The Persona Effect: Affective Impact of Animated Pedagogical Agents. In: Pemberton, S. (ed.) *Human Factors in Computing Systems: CHI 1997 Conference Proceedings*, pp. 359–366. ACM Press, New York (1997)
- [11] Rehm, M., Andre, E.: Where do they look? Gaze Behaviors of Multiple Users Interacting with an Embodied Conversational Agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) *IVA 2005. LNCS (LNAI)*, vol. 3661, pp. 241–252. Springer, Heidelberg (2005)
- [12] Traum, D., Marsella, S., Gratch, J., Lee, J., HartHolt, A.: Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-Modal Virtual Agents. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) *IVA 2008. LNCS (LNAI)*, vol. 5208, pp. 117–130. Springer, Heidelberg (2008)
- [13] Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., Vilhjálmsón, H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) *IVA 2006. LNCS (LNAI)*, vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
- [14] Cassell, J.: *Embodied Conversational Agents*. MIT Press, Cambridge (2000)
- [15] <http://www.visagetechologies.com>
- [16] http://www.omron.com/r_d/coretech/vision/okao.html
- [17] <http://sourceforge.net/projects/opencvlibrary/>
- [18] <http://www.loquendo.com/en/>
- [19] <http://en.wikipedia.org/wiki/OpenAIR>
- [20] <http://wiki.mindmakers.org/projects:bml:main>
- [21] Vilhjálmsón, H.H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S.C., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: The behavior markup language: Recent developments and challenges. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) *IVA 2007. LNCS (LNAI)*, vol. 4722, pp. 99–111. Springer, Heidelberg (2007)
- [22] A.L.I.C.E. AI Foundation: *AIML, Artificial Intelligence Markup Language* (2005), <http://www.alicebot.org/TR/2005/WD-AIML>

Actively Adaptive Agent for Human-Agent Collaborative Task

Yong Xu¹, Yoshimasa Ohmoto¹, Kazuhiro Ueda², Takanori Komatsu³,
Takeshi Okadome⁴, Koji Kamei⁵, Shogo Okada¹, Yasuyuki Sumi¹,
and Toyoaki Nishida¹

¹ Department of Intelligence Science and Technology,
Graduate School of Informatics, Kyoto University, Yoshida-Honmachi,
Sakyo-ku, Kyoto 606-8501, Japan

² Department of System Sciences, the University of Tokyo

³ International Young Researcher Empowement Center,
Shinshu University

⁴ School of Technology and Science, Kwansei Gakuin University

⁵ Innovative Communication Laboratory,
NTT Communication Science Laboratories

Abstract. Active interface is one of critical characteristics of agents who have to interact with human users to achieve human-agent collaboration. This characteristic is especially important in beginning phase of human-agent interaction when an ordinary human user begins to interact with an adaptive autonomous agent. In order to investigate principal characteristics of an active interface, we developed a human-agent collaborative experimental environment named WAITER. Two types of experiment: WOZ agent experiment and autonomous agent experiment were conducted. Objective of the experiment is to observe how human users change their instructions when interacting with adaptive agents with different degree of freedom. Experimental results indicate that participants can recognize changes of agent's actions and change their instruction methods accordingly. It infers that changes of instruction method depend not only on waiter agent's reactions, but also on human manager's cognitive models of the agent.

1 Introduction

In researches on human-agent interaction(HAI), many rapid scientific and technological developments in human-centered interfaces are drawing a lot of attention. There are mainly two types of agents: interface agents and autonomous agents [1]. Active interface is considered to be one of critical characteristics of agents who need to interact with human users to achieve human-agent collaboration. This characteristic is especially important in beginning phase of human-agent interaction when an ordinary human user begins to interact with an adaptive autonomous agent. In [2], Yamasaki defined *Active Interface* as a type of human-robot(agent)

interface that does not only wait for users' explicit input but also tries to get information from users' implicit input, and external environment. Based on the gathered information, it acts spontaneously and keeps the system in an advantageous condition for users. Maes [3] argued that an adaptive autonomous agent need to solve two problems: action selection and learning from experience. In order to enable an agent to make a rational decision, a popular solution is utilization of a BDI (Belief-Desire-Intention) model [4]. BDI model supposes that an autonomous agent should have its own intention, and make its own decision when interacting with its environment. In order to investigate how human users interact with an adaptive autonomous agent with an active interface, we designed a specific experimental task that includes an adaptive autonomous agent with different degree of freedom. We implemented a human-agent collaborative experimental environment named WAITER (waiter agent interactive training experimental restaurant). Two types of agents: WOZ(Wizard of OZ) agent and autonomous agent are implemented. A WOZ agent has less degree of freedom, and a human WOZ operator partly controls its action selection policy. While an autonomous agent can choose its actions by itself. Experiments are conducted to investigate how human users change their instructions when they interact with agents having different degree of freedom.

2 Active Interface and Mutual Adaptation

2.1 Definition and Hierarchical Structure

Since goal of our study is to develop an adaptive autonomous agent with an active interface, it is necessary to clarify the meaning of the phrase "adaptive autonomous" at first. The word "autonomous" means that the agent has high degree of freedom in decision-making, and word "adaptive" - high degree of freedom in self-organizing. It is difficult to decide how much degree of freedom or initiative (leadership) to endow an autonomous agent. As deficiency of degree of freedom may restrict agent's ability of adapting to environment, however, redundant degree of freedom can produce unnecessary movement and instability. In addition to the issue of degree of freedom, people's adaptation is another problem. Utility of an adaptive interface includes: it can (1) make people easily recognize the change of the agent, (2) help people feel easy to adapt to the agent, and (3) facilitate invoking mutual adaptation phenomenon. In research field of HAI, since human factors make the environment more complicated, it is necessary to take account of people's adaptation ability during the procedure of human-agent interaction. Therefore, an agent with an active interface can not only help it adapt to human users, but also help people feel easy to adapt to the agent. In other words, *mutual adaptation* is considered as one of competences that are useful for designing an active interface.

Suppose that there are two agents A and B who need to perform a collaborative task. Neither A nor B has complete information about the task. Each agent only has different partial information and shares part of common information with each other. In order to achieve a common purpose, each agent has to build a

model for the other, try to adapt to the other by changing its action, predict the other's next action, and draw inferences about the other depending on various situations. In general, the agents have to complete the task by gradually adapting to each other. In this paper, the above-mentioned phenomenon is termed *mutual adaptation*.

2.2 Related Works

Various of researches have been done in the area of HRI(human-robot interaction) or HAI(human-agent interaction). For example, [5] took a view of point from human-teacher and robot-learner and [6] tried a method of animal training to teach a pet robot. While [7] studied adaptation between agents. However, these researches have not yet provided a solution to the problem that how human user can finish a collaborative task with an adaptive autonomous agent. Mutual adaptation phenomenon has been observed in human-human interaction experiment [8], results of this experiment indicates that mutual adaptation is helpful for establishing communication protocols and conveying intentions between human users in a collaborative task. Regarding people as advanced autonomous agent, we expect that mutual adaptation phenomenon will also occur in human-agent interaction.

3 Experimentation

3.1 Objective

This study aims to verify if an agent's actively adaptive behavior can cause the human users change their instruction methods, and if this change can cause agent's change so that mutual adaptation phenomenon can be observed. For this purpose, an experiment was conducted by considering a waiter training task.

3.2 Task

A waiter (agent) performs its task by collaborating with a manager in a virtual restaurant which is called WAITER (waiter agent interactive training experimental restaurant.) The layout of the virtual restaurant is illustrated in Fig. 1. Entrance has up to two states: "has no customer," and "has new customer(waiting for guide)." Kitchen has up to two states: "no dish," and "has dish." Each of nine tables has up to four states: "vacant," "waiting for order," "eating," and "need to clear." The manager can issue an instruction by pressing one of eleven buttons (including entrance, kitchen and nine tables. The waiter agent has up to three states:"free," "guiding (customer)," and "carrying dish."

A human manager (instructor) is asked to issue instructions toward a waiter agent(actor) when he feels necessary by pressing buttons. If the waiter changes its reaction according to managers' instructions, it may affect the managers' way of instruction. On the other hand, if the manager changes his way of instruction

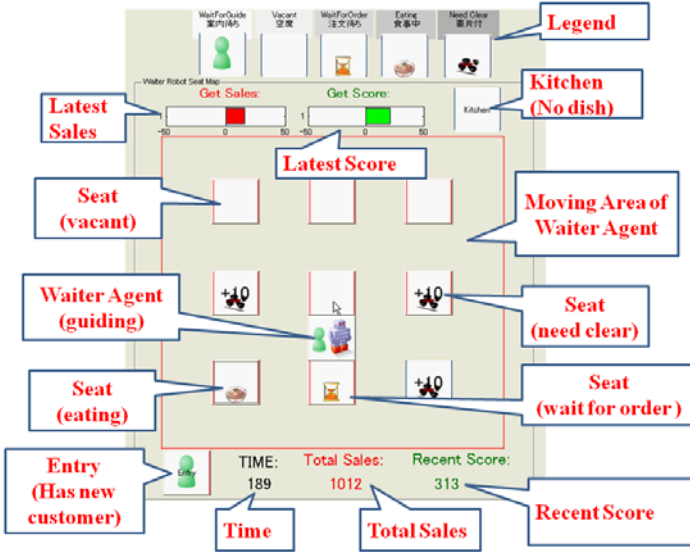


Fig. 1. Graphical User Interface of WAITER (waiter agent interactive training experimental restaurant) system

by observing the waiter’s responses, it may affect the waiter’s action as well. In order to achieve a better score, both of them have to collaborate with each other. Therefore, mutual adaptation phenomena are expected to occur.

In this manager-waiter (human-agent) collaborative task, if waiter agent’s active adaptation can cause human manager’s adaptation, in other word, if a waiter agent changes its action by adapting to a human manager, it is expected that the manager will accordingly change his instruction methods. This is the first loop of mutual adaptation circulation. If the agent can learn by improving its competence, it is hopeful that the adaptation circulation will continue, and long term human-agent interaction will be possible. In this paper, we will focus on the first circulation of the mutual adaptation, and investigate its relation to the active interface of an adaptive autonomous agent.

3.3 Implementation

A waiter agent(robot) simulator software system [9] with a graphical user interface has been developed using MATLAB(2007), as shown in Fig. 1. Two modes are designed for the waiter agent: manual mode and autonomous modes. In manual mode, the waiter agent follows every instruction from the manager. In contrast, in autonomous mode, the agent can move autonomously by making its decision to take its own action or to follow the manager’s instruction.

Human participants who play the role of manager can instruct agent by pressing buttons on the GUI(Graphical User Interface) of the WAITER system. The

goal of the experiment is to investigate if active changes of the agent's behaviors can cause participants change their instructional methods, and invoke mutual adaptation in human-agent interaction. Firstly, whether changes of an active adaptive agent can be recognized by a human instructor need to be examined. Secondly, whether an active adaptive agent can encourage the human instructors to change their instruction way needs to be investigate as well. Finally, if both are proved effective, establishment of the first circulation loop of mutual adaptation can be confirmed. In other words, it can be confirmed that an agent adapts to a manager by changing its action, and the manager adapts to the agent by changing his instructions accordingly.

Two types of experiments: WOZ agent experiment and autonomous agent experiment were conducted. There are two types of WOZ agent experiment and three types of autonomous agent experiment. In the first type of WOZ experiment, number of dishes that the agent can carry to kitchen increases gradually for three stages. In the second type of WOZ experiment, the agent changes its action for six stages. In addition to the number of dish the agent can carry to kitchen each time, the capability of automatically find a path to the kitchen is appended. The stage-switching is controlled by a human WOZ operator or by a timer. If the WOZ operator does not give any command, the agent automatically switches its stage at every one third moment respectively in the first WOZ experiment, and switches its stage at every one sixth moment in the second WOZ experiment.

In the WOZ agent experiment, the agent changes its behavior timely or by following the WOZ operator(expert user)'s command. In contrast, in the autonomous agent experiment, the agent changes its behavior according to the manager(participants who are ordinary user)'s instructions.

A typical scenario for autonomous mode is as follows. When a new customer enters the restaurant, if a seat is vacant and the agent is idle (in "free" state), the waiter agent will move to the entrance and guide the customer to the vacant seat. When the waiter is guiding a customer, its state will change to "guiding customer" automatically. If there is neither an available vacant seat nor a seated customer waiting to place an order, the agent needs to find a "need to clear" table, move there, clear the table, and change its state to "carrying dish" while carrying dishes to the kitchen. However, if any customer is waiting to place an order, the agent will place the order. After the customer has been seated at a vacant seat, the table state will be changed to "waiting for order." It will be kept unavailable for placing orders for a specific period. In order to prevent customers from leaving the restaurant because they have waited too long time, which in turn reduces the sales and scores, the agent always tries to place an order as soon as possible when operating in the autonomous mode. When the agent finishes placing an order, it receives tips, and the restaurant's sales and scores are updated simultaneously.

In the beginning of WOZ agent experiment, the agent is designed to purposefully ignore "need to clear" table at high possibility. Since the agent's capability of carrying dish is gradually increased by a timer or by the operation of a WOZ

operator, so that the manager is expected to be aware of the agent’s changes easily.

In autonomous agent experiment, three types of switching mode of the agent: “linear prediction mode,” “random mode,” and “bayesian network mode” are designed to switch its priority among three types of specific task: “guide a customer,” “place an order,” and “clear a table.” As autonomous agent always tries to adapt to the manager by changing its actions, the manager is expected to be aware of the adaptation from the agent.

In linear prediction mode, the agent calculates the recent frequency of three types of instruction (“guide a customer,” “place an order,” and “clear a table.”) and changes its priority according to the most frequently used instruction. In random mode, except to finish current task automatically, the agent switches its priority randomly at fixed intervals. While in Bayesian network mode, a Bayesian network, which was generated based on previous experiment results, was used to choose next action.

Think aloud method was utilized to facilitate analysis of instruction intention of human manager (instructor). All participants were required to speak out intentions of their instruction while issuing instructions.

As a result, three types of data were collected. Voice data were collected by using a digital voice recorder. The intermediate processing data of WAITER system were recorded in log files. These data include movements of the waiter, instruction (pressing button actions) of the manager, and value of sales and score etc. The screen output data of WATIER system were captured by a screen recording software and recorded as video files. Behaviors and voice of participants were also videotaped by a video camera.

It is expected that the manager’s instruction method will be affected by the agent’s autonomous function. This in turn will enable the agent to learn and improve its performance and adapt to the manager by trying various actions. Therefore, it is desirable to invoke mutual adaptation.

3.4 Participants and Procedure

Totally 12 students (5 male and 7 female, age: 20-26 years, average age: 22.8) participated in the WOZ agent experiment. We grouped them into 2 groups. For convenience, we used alphabets to represent two groups of participants in WOZ agent experiment: A through F for the first group, and G through L for the second group.

As for the autonomous agent experiment, 25 students (16 male and 9 female, age: 18-37 years, average age: 21.7) participated in the experiment. We grouped them into 3 groups. The first group, 6 participants (5 male and 1 female, age: 19-23 years, average age: 20.1), participated in the first experiment. The second group, 11 participants (6 male and 5 female, age: 19-37 years, average age: 22.5) participated in the second experiment. The third group, 8 participants (5 male and 3 female, age: 19-28 years, average age: 21.8) participated in the third experiment. For convenience, we used combination of alphabets and numbers to represent three groups of participants in autonomous agent experiment respec-

tively : N1 through N6 for the first group, N7 through N17 for the second group, and N18 through N25 for the third group.

First group participants were asked to complete the first trial by instructing waiter agent in manual mode. Then, they were asked to complete the second trial in random mode. Finally, they were asked to complete the last trial in linear prediction mode. The result data of all trials ,namely, scores, sales, agent states, table states, together with manager's press-button instructions were recorded in log files. In the second group, in addition to manual mode, linear prediction mode, and random mode, a Bayesian Network mode is appended. In the third group, in addition to the above mentioned four modes, an observation-phase mode was added.

In random mode, the agent switched its priority by random. In linear prediction mode, the agent changed its priority by adapting to human user's instruction. Bayesian Network mode was established basing on previous experimental results, and was expected to predict human instructor's instruction by learning from experience. Observation-phase mode was added, since it was expected to enforce the human managers to stop their instructions so that they had a chance to recognize the merit of autonomous function of the agent. This aimed to prevent the human instructor to keep giving instruction in all time.

3.5 Results and Discussions

In the first group (3-stage) of WOZ agent experiment, each one participated for 4 10-min trial respectively with agent working in four different modes. The four modes were manual mode, autonomous mode with timely stage-switching, autonomous mode with operator stage-switching and autonomous mode with timely or operator stage-switching. Caused by recording problems, two trials of first participants were not recorded correctly, an additional trial was added to the last participant. Totally 23 log files, 230 min voice and videotaped data for 6 participants were recorded in the first group of WOZ agent experiment. In the second group (6-stage) of WOZ agent experiment, each participant participated for 6 trials with one for manual mode (10 min), one for autonomous timely switching mode (10 min), and three for autonomous operator switching mode (10min, 20min and 30min respectively). Three people participated in an additional autonomous operator switching mode. Totally 33 log files, 510 min screen captured video files, and voice and videotaped data with the same time length were recorded.

As for the autonomous agent experiment, from the first group of 6 participants, second group of 12 participants and third group of 8 participants, 25 manual mode trials, 25 linear prediction mode trials, 24 random mode trials and 8 observation mode trials, 19 Bayesian Network mode trials were recorded.

In both group of WOZ agent experiment, there were significant differences in the instruction interval time between the first half and the second half of the same trial. For the first group, 16 out of 23 trails used longer time interval in second half than in the first half ($p=0.047$, one-tailed test). For the second group, 24 out of 33 trails used significantly longer time interval in the second half than

in the first half ($p=0.007$, one-tailed test). Although in the autonomous agent experiment, there was no significant difference for general comparison in the instruction interval time between the first half time and the second half time of the trials (29 out of 51 trails, $p=0.201$, one-tailed test), there was significant difference in the Bayesian Network mode (13 out of 19 trails, $p=0.084$, one-tailed test).

The statistical analysis results indicate that for a WOZ agent, more degree of freedom is useful to improve the human participants' awareness on the changes of the agent. As for the difference between WOZ agent and autonomous agent, since the agent's behavior was controlled to change immediately to respond to human user's instructions, but current version of autonomous agent may have a time lag between the changes of the agent's action and the changes of the participant(manager)'s instruction, this time lag may make it more difficult for the participants to recognize changes of the autonomous agent. Although linear prediction function enable the agent adapt to the user by switching its priority among "guiding a customer firstly," "placing an order firstly" and "clearing a table firstly." Since Bayesian Network mode was implemented basing on the results of previous experiment results, it seems work better to encourage participants to change their instruction behaviors.

Fig. 2 shows a typical example of changing behaviors of one manager's instruction. It indicates that the same manager spend shorter time in the first half of the trial than in the second half for switching between consecutive instructions. This change of instruction method infers that the manager may adapt to the agent by observing the actions of the agent before giving instructions. As long as the agent takes a proper action, the manager prefers to observe and give initiative(leadership) back to the agent. Only when the manager finds the agent takes some wrong or unexpected actions, he gives instructions to change the agent's actions.

Fig. 3 illustrated the changing procedure of the table states with all types of instructions. From this figure, it is obvious that the numbers of two types of instruction (vacant-guiding instructions that a manager uses to instruct an agent to guide a new customer to a vacant table, and kitchen-guiding instruction that is used to tell the agent where it should put dishes. Decrease of these instructions indicates that the manager may adapt to the waiter by observing if the waiter agent can finish these two types of task automatically. If the agent can work well, they just give the initiative and let it work by itself. Therefore, initiative transferring happens in this case.

Fig. 4 shows the usage times of three types of instruction: guidance instruction, order instruction and clear instruction are all decreased along time. The trend that all usage times of instruction decreases along the time shows similar result as mentioned above.

Fig. 5 shows another example where the manager took another different method of instruction. In this case, the manager observed the agent's autonomous movement without giving any instruction in the beginning. Since the agent's actions were not satisfying, he began to instruct, it caused an increase of instruction

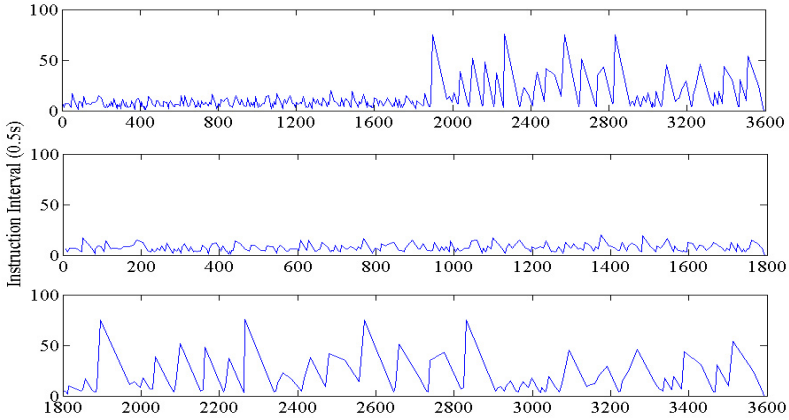


Fig. 2. Changes of Instruction Interval, Participant J, 30min, WOZ agent experiment, top is full Trial, middle is first half and bottom is second half, time unit is 0.5s

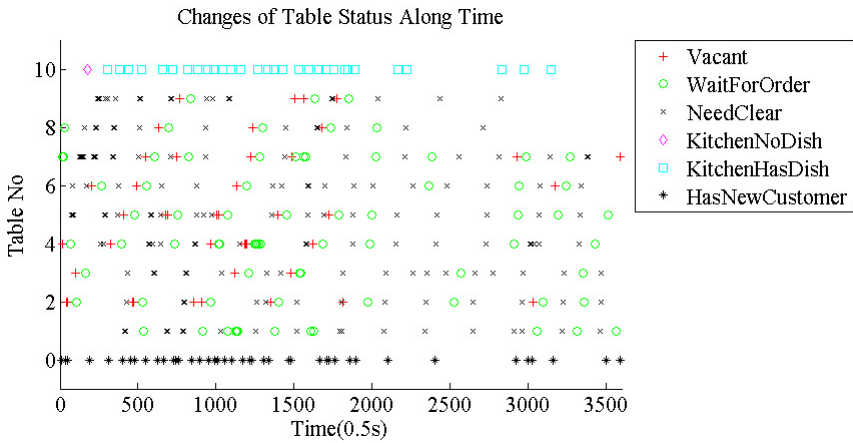


Fig. 3. Changes of Table States Along Time (Participant J, 30min, WOZ agent experiment, Axis Y is table No: 0,Entry; 1-9,Table 1-9;10, Kitchen;Axis X is Time, time unit is 0.5s)

interval after a short period as shown in the top figure. The lower figure shows that the most frequent instruction interval is around 8.0 (std=4.2, time unit is 0.5s), which means that the manager averagely spent about 4s to issue next instruction after preceding one. It infers that the manager might build a model for the agent after observing its autonomous actions in the beginning of the first trial. The fact that no obvious changes during the trial infers that the manager might finish building a stable mental model for the agent in the beginning and keep using it in remain time.

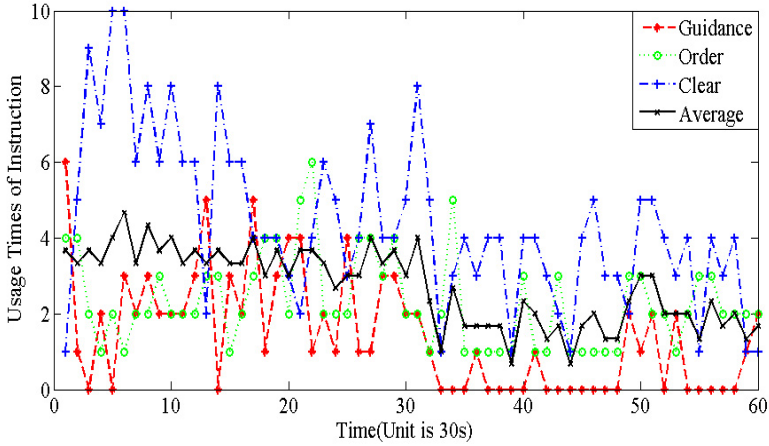


Fig. 4. Usage Times of Instruction (Participant J, 30min, WOZ agent experiment, width of time window is 30s)

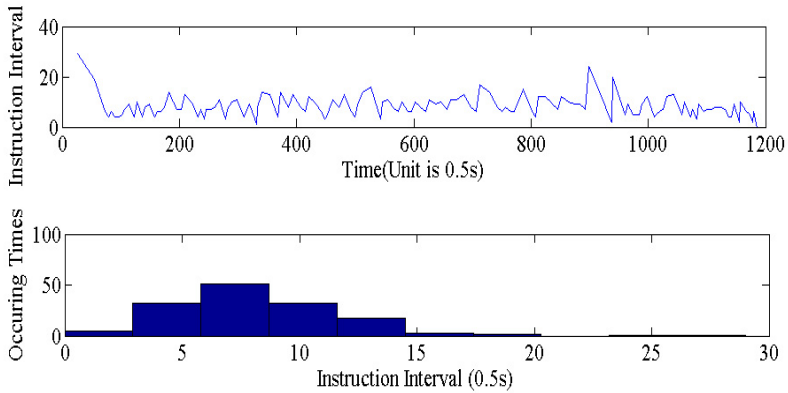


Fig. 5. Instruction Intervals ((Participant H, 10min, WOZ agent mode, top figure, time unit is 0.5s) and histogram of Occurring Times(bottom figure)

Subjective evaluation on agent’s adaptation ability was performed by requiring participants to fill questionnaires before and after experiment, and one specific questionnaire after each section. For WOZ agent experiment, answers to the question “How intelligent is the robot do you expect(felt) in this experiment?” show that 67% (8 out of 12) participants gave a positive evaluation by rating same or higher score after the experiment than before the experiment.

Table [1](#) shows participants’ subjective evaluation results about the changes of agent’s behavior and the participant’s instructions during break time of the experiment between different trails. Answering to question Q1: “How much do you think the agent changes its behavior?” and Q2: “How much different do you

Table 1. Subjective Evaluation for the Behavior Changes

Experiment Type	Rate Item	Rate Score
Woz agent	Change of Agent’s behavior(Question1)	5.3(std=1.23)
Auto Agent	Change of Agent’s behavior(Question1)	5.08(std=1.24)
Woz Agent	Change of Instruction behavior(Question2)	4.6(std=1.5)
Auto Agent	Change of Instruction behavior(Question2)	4.96(std=1.4)

instruct the waiter at the beginning and at the end?” the participants of WOZ agent experiment gave relatively higher scores (mean= 5.3, std=1.23, 7 stage evaluation score with lowest evaluation 1 to highest evaluation 7) to Q1 and a slightly high score (mean=4.6, std=1.5, 7 stage evaluation) to Q2. This result suggests that many participants recognized the changes of agent’s behavior and also changed their instructions method by themselves. As for the reason why they changed their instruction method, 9 out of 12 of participants of WOZ agent experiment chose the answer “because the agent changes its actions.” Rest participants chose the answer “because I could not predicate the agent’s action.” As for the answers to the same question, 44% (11 out of 25) participants of autonomous agent experiment chose the same answer, and other 6 participants chose the answer “because I could predicate the agent’s behavior.”

Although all three utilities of an adaptive interface have not yet completely achieved in current experiment, at least part of them were achieved effectively. The experimental results indicate that current system can (1) make most of human users recognize the change of the agent with relatively high degree of freedom, (2) cause some of human users change their instruction method to adapt to the agent, and (3) seems hopeful to finish at least the first circulation loop of the mutual adaptation phenomenon. Since mutual adaptation is considered as a very general phenomenon that occurs when human users facing any adaptive artifacts. Although current environment uses a specific waiter training task, it can be considered as an effective platform to study this topic. By implementation of various machine learning algorithms, this environment is potentially to be used to elucidate the essences of interactive learning.

4 Conclusions

In order to investigate principal characteristics of an active interface, we develop a human-agent collaborative experimental environment named WAITER. Two types of experiment: WOZ agent and autonomous agent experiment were conducted. Objective of these experiments is to disclose how human users change their instructions when interacting with adaptive agent with different degree of freedom. As a result, some adaptive behaviors were observed. Experimental results indicate that human participants can recognize changes of agent’s actions and change their instruction methods accordingly. It infers that changes of instruction method depend not only on agent’s reactions, but also on human user’s

cognitive models of the agent. Further analysis of result data is needed to extract participants' instruction patterns and to improve the agent's capability. The results of experiment also suggest that active adaptation may play an important role in human-agent collaborative task.

References

1. Lieberman, H.: Autonomous interface agents. In: CHI 1997: Proceedings of the SIGCHI conference on Human factors in computing systems, USA, pp. 67–74 (1997)
2. Yamasaki, N., Anzai, Y.: Active interface for human-robot interaction. In: IEEE International Conference on Robotics and Automation, Japan, vol. 3, pp. 3103–3109 (1995)
3. Maes, P.: Modeling adaptive autonomous agents. *Artificial Life* 1(1–2), 135–162 (1994)
4. Wooldridge, M.: Reasoning about Rational Agents. MIT Press, Cambridge (2000)
5. Andrea, L., Thomaz, C.B.: Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 716–737 (2008)
6. Kaplan, F., Oudeyer, P.Y., Kubinyi, E., Miklosi, A.: Robotic clicker training. *Robotics and Autonomous Systems* 38(3-4), 197–206 (2002)
7. Goldman, C.V., Rosenschein, J.S., Rosenschein, J.S.: Incremental and mutual adaptation in multiagent systems. Technical report, Insitute of Computer Science, The Hebrew University (1996)
8. Xu, Y., Ueda, K., Komatsu, T., Okadome, T., Hattori, T., Sumi, Y., Nishida, T.: Woz experiments for understanding mutual adaptation. *Journal of AI & Society* 23(2), 201–212 (2009)
9. Xu, Y., Ohmoto, Y., Ueda, K., Komatsu, T., Okadome, T., Kamei, K., Okada, S., Sumi, Y., Nishida, T.: A platform system for developing a collaborative mutually adaptive agent. In: International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2009). LNCS(NGAI), vol. 5579, pp. 576–585. Springer, Heidelberg (2009)

Low-Overhead 3D Items Drawing Engine for Communicating Situated Knowledge

Loic Merckel^{1,2} and Toyoaki Nishida¹

¹ Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

loic@ii.ist.i.kyoto-u.ac.jp,
nishida@i.kyoto-u.ac.jp

² Dept. of Scientific Systems R&D, HORIBA Ltd., Kyoto 601-8510 Japan

Abstract. This paper presents a low-overhead 3D items drawing engine for a situated knowledge medium aiming at improving the knowledge communication between experts and end-users of scientific instruments. The knowledge representation is based on the concept of Spatial Knowledge Quantum that associates a 3D geometry structure with knowledge. Our engine attempts to provide an effective means for creating those 3D structures. It consists of a hand-held Augmented Reality (AR) system and allows to directly draw, in the context of a subject instruments, free 3D lines. A line is recorded as a set of sample points. From these points a volume can optionally be interpolated by performing a Delaunay tetrahedralization. During the drawing operations, a physically concrete version of the instrument is not required. The AR interface shows the real-world combined, not only with the created 3D items, but also with a virtual representation of the instrument. A virtualized instrument offers many advantages such as availability, mobility, scalability, spatial freedom (e.g., we can draw through it) and rendering options.

1 Introduction and Background

The recent progress of the Augmented Reality (AR) on embedded/portable devices makes this technology realistic to be utilized in real-world environment with the aim of developing efficient support systems for the users of complex instruments and machineries [1][2][3][4]. Consequently, a number of systems have been designed as an efficient alternative to the documentation/manual paradigm to support the knowledge communication, between experts and end-users, in real-world situations. Among these initiatives, we can distinguish several approaches. For instance, 3D computer-generated data are superimposed onto the images so as to guide a user through performing a sequence of 3D tasks (e.g., an early prototype is described by Feiner et al. [5]), whereas some other contributions focus on techniques to annotate the real-world [6][7][8]. This latter approach is hampered by the fact that text information, even situated, may limit the possibilities of expression (especially in an international context in which the different actors do not share the same native language).

A support system (e.g., [5]) that enhances the real-world using complex geometry structures requires a certain amount of preliminary work to construct and integrate those

structures, making this approach falling into what Fischer and Ostwald [9] refer to as “*traditional KM*” that, among other flaws, lacks of interactivity (knowledge integration is performed at “*design time, before system deployment*”). Moreover, a CAD software is employed to build those structures. Although numerous of such a software are commercially available, they are all complex and requiring a long period of apprenticeship (in addition to the fact they are usually expensive). As a consequence, many experts on a subject complex device may either encounter some serious setbacks to express their knowledge (to the point of giving up), or simply not consider using such media to formalize their knowledge.

The long-term goal of this research is to develop a situated knowledge medium, aiming at supporting the users of complex scientific instruments, that consists in grounding knowledge in real-world situation [10][11]. The goal of the current research is to build an engine (as a part of the entire system) that allows the user to easily draw free 3D lines and items in the context of the instrument in order to generate virtual 3D structures as a vehicle for expressing/communicating knowledge. The engine consists of a tablet PC equipped with a video camera and a 6-DOF orientation and position sensor. It provides the user with a hand-held AR interface showing a superimposition of a virtual representation of the instruments as well as the drawn items onto the real-world image. To draw a line, the user just moves the tablet PC so that the trajectory of a virtual pointer, displayed with a fix position with regard to the tablet PC, is recorded (as a set of sample points). From this set of points, a Delaunay tetrahedralization can be performed to interpolate a volume item.

The remainder of this paper is organized as follow. In the next section, we introduce the entire framework we use for implementing a situated knowledge medium. In section 3, we briefly present our knowledge representation that combines a virtual 3D item with some additional information (multimedia files). The engine for 3D lines and items drawing is described in section 4. Finally, in section 5, we summarize our key points.

2 Framework for Situated Knowledge Management

Our situated knowledge medium encompasses the knowledge management process as shown in Fig. 1. This approach follows the design perspective proposed by Fischer and Ostwald [9] in which the knowledge is created through a collaborative process, integrated at use time (and not at design time) and disseminated on demand. When a user needs to get some information about a particular component, she or he just asks for it by selecting the item of interest (via the proposed interfaces). If the system fails to obtain an adequate answer for the selected position, the user is invited to ask a question about the position. The question is stored in the database and forwarded to experts. When the experts answer the question, a new entry will be created in the database and associated with the position for which the question was given.

The interface allows the user to create/edit and retrieve knowledge. We propose a computational framework that integrates AR and Augmented Virtuality (AV) as interface. The knowledge representation is independent from those interfaces. Both AR and AV can be used to manipulate the same knowledge base. Both AR and AV interfaces possess a similar functionality with each other in the sense that these interfaces allow

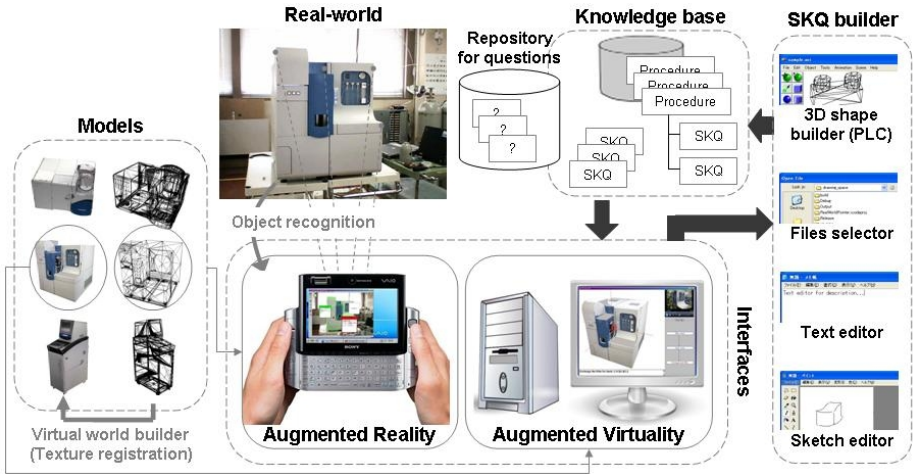


Fig. 1. Framework overview. The current work addresses the “3D shape builder”.

to visualize the subject instrument enhanced with adequate knowledge representation (see the next section). The main difference lies in the fact that the situated knowledge does not augment directly the real-world instrument, but augments a virtual environment representing the instrument. Figure 1 presents this approach. The AV interface is easy to produce and to distribute via the software dedicated to the instrument without additional costs (allowing knowledge dissemination at a larger scale), whereas these statements are untrue regarding the AR approach.

A major difficulty to build such interfaces is the pose computation of the subject instrument (that consists of its position and its orientation with respect to the camera coordinate system). Indeed, AR needs the pose (at run time) for registering virtual objects with the real world image. In order to build a virtual instrument, we augment the 3D model with real-world textures that are acquired from a set of images. For each of those images, a precise pose is required to extract a correct texture patch. We have proposed a solution to the pose calculation problem in previous works and we have introduced a novel object recognition algorithm [12][13] as well as a low-overhead technique for texturing 3D models [10][14].

A three-dimensional pointing system is implemented by the both approaches, allowing the user to select a 3D position from the 2D screen. Concerning the AR interface, we introduced this low-cost three-dimensional pointer based on pose calculation and mesh generation of the instrument model in a previous work [8]. This approach has been adapted to the AV based interface. Since we are in a virtual environment, the pose of the instrument is already precisely known, and consequently, only the mesh generation of the model needs to be performed.

The initial inputs of the overall system are only the CAD models corresponding to the instruments. As set forth above, the knowledge base does not need to be preliminary crafted by specialists. The required knowledge will be dynamically created after deployment.

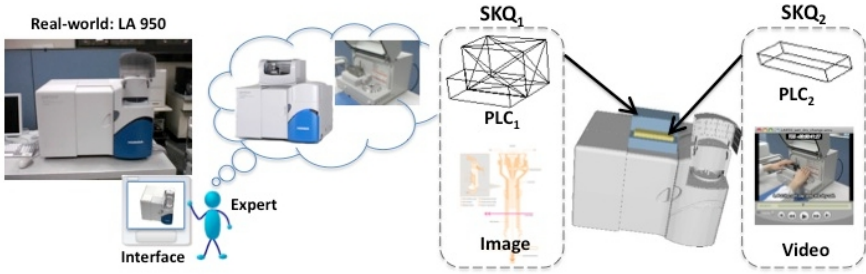


Fig. 2. The expert shares his or her knowledge about the PowderJet by creating two SKQs that encapsulate PLCs to suggest the shape and to indicate the location of the additional equipment

3 Situated Knowledge Representation

Preliminary version. A preliminary conceptualization of Spatial Knowledge Quantum (SKQ) was introduced in [8]. In SKQ, a spatial position in the environment is associated with knowledge. SKQ consists of two components: a three-dimensional point in the real world $p \in \mathbb{R}^3$ and the associated knowledge, which can be potentially stored in many forms: text, drawing, video and audio (it can be a combination of different items). Ordered sequences of SKQs (referred as *procedures*) can be defined and stored in the knowledge base.

Let us consider the situation in which the expert want to provide information about optional equipments and accessories for the HORIBA Partica LA-950 Particle Size Analyzer¹ which comes with several options including the PowderJet Dry Feeder². Suppose that the expert would like to explain the advantages and utilization of this option to users who already own the basic instrument. Text explanations, pictures or video files might be unsatisfactory, especially if a user is interested by asking a question about a particular position in the optional equipment (as those media are not interactive). The expert may prefer draw an interactive 3D sketch of the additional device integrated with the instrument (Fig. 2).

Extended spatial knowledge quantum. We extend the concept of SKQ to include arbitrary 3D shapes. The point p introduced in the previous definition can be replaced by a three-dimensional piecewise linear complexes. Piecewise Linear Complexes (PLC) were introduced by [15]. A simpler description, limited to three dimensions, has also been given [16,17]. A PLC is a general boundary description for three-dimensional objects and can be simply defined as a set of vertices, segments, and facets. Each facet is a polygonal region; it may have any number of sides and may be non-convex, possibly with holes, segments, or vertices. PLCs have restrictions. For PLC X , the elements of X must be closed under intersection. For example, two segments can intersect only at a common vertex that is also in X . Two facets of X may intersect only at a shared segment

¹ <http://www.jobinyvon.com/LA-950V2>

² <http://www.jobinyvon.com/PowderJet>

or vertex or in a union of shared segments and vertices. Another restriction is that the set of points used to define a facet must be coplanar.

Our choice to restrict the 3D structure to a PLC lies in the fact that such a structure exhibits adequate properties to perform a powerful mesh generation algorithm (Refinement of Constrained Delaunay Triangulation [18]). As a result, the three-dimensional pointing system introduced by [8] can be directly employed to select a precise position on a given SKQ.

Regarding the new SKQ definition, the CAD model defining the instruments can be interpreted as a particular SKQ, which is hidden in the rendered scene. It serves as a root structure that provides the reference coordinate system in which all other SKQs will be defined. We can associate new SKQs with any existing SKQ.

This extended definition of SKQ is compatible with the previous one presented by [8]. Indeed, A previous SKQ can be interpreted as a new SKQ_i such that the associated PLC_i is reduced to a point $p \in \mathbb{R}^3$.

4 Toward Building Spatial Knowledge Quantum

4.1 Overview

The difficulty in building SKQs resides in the generation and the integration of the associated three-dimensional PLCs from a 2D interface with low overhead. Although numerous 3D modeling softwares allow to produce precise and complex shapes of various kinds and are used for a wide range of applications, the user interfaces remain generally complex enough to prevent a non-experienced user from quickly grasping even the basics functionalities. The fact that those interfaces are based on the WIMP paradigm has been criticized and considered as unnatural for 3D tasks [19]. Alternatively, many works have proposed new types of interfaces, such as gestural and suggestive (e.g., [20,21,22,23]), in order to draw rough 3D objects in an *effortless* manner. However, these approaches may lack functionalities for registering on-the-go the created virtual items in the current scene (video image under perspective projection). Furthermore, they require the user to learn a gestural semantic that may not always be intuitive and natural.

The purpose of drawing is not to manufacture the object, but to help the user to express his or her idea so as to share it with others. In this context, the precision is not a main concern. Complex shapes can be approximated by a combination of simple primitives. From our experiences, the WIMP based 3D modeling approach can be efficient to draw and to assemble primitives such as cube, sphere, cylinder and so on.

We have tried *Art of Illusion*³. Although it does not provide as much functionalities as commercial alternatives, it is significantly easier to operate and still appears to be enough for drawing fairly complex PLCs. The root model (PLC_0) as well as the selected PLC_i (if different from PLC_0) are exported (currently STL format) and imported in the CAD software with which the new PLC_j is created, relatively to PLC_0 and PLC_i , before being exported and re-imported in our system.

³ <http://www.artofillusion.org/>

However, this approach exhibits several flaws. First, to manipulate a 3D component through only a 2D multi-panels view (e.g., top, front, sides, that is a typical configuration for WIMP based 3D modeling software) can be conceptually difficult for users who are not well prepared. Second, the texture information cannot be easily communicate to the CAD software (it would required a huge amount of additional work). Finally, the 3D structure generated using the CAD software will most likely not conform the PLC definition. Consequently, additional efforts must be done so as to “manually” constrain the geometry (lot of skills on utilizing the CAD software as well as a deep understanding of the PLC structure are necessary). Furthermore, this approach is not well designed for the AR based interface, that should be run on a tablet PC using a stylus. Our proposed 3D drawing engine attempts to address those limitations.

There is a vast body of literature on 3D user interaction techniques and interfaces for 3D drawing (e.g., [24][25][26][27]). Most of them rely on cumbersome hardware components and settings (e.g., specific room settings, head-mounted displays, numerous wearable sensors and so on), and usually apply to virtual environments rather than the real world. Our present approach has some similarities to the collaborative AR system discussed in [26]. The both systems attempt to provide an easy way to draw 3D lines in real-world. In [26], a complex hardware settings is employed and the focus is set on collaborative drawing in an empty space involving several users simultaneously. In contrast, our system is handy and our target is to draw not only lines, but also other sort of items, in the context of a subject instrument.

4.2 3D Items Drawing Engine

Our 3D items drawing engine is a hand-held AR system (Fig. 3) that allows to *naturally* draw 3D lines, defined as a set of sample points (that is equivalent to a set of subsequent line segments), directly in the real-world by moving the system (as if it was a sort of *3D pen*). From the set of points defining a given line, we can interpolate a volume by performing a Delaunay tetrahedralization (Fig. 4). Note that the interpolated volume conforms to the definition of a PLC.

The AR interface shows the real-world combined with a virtual representation of the instrument, called a *virtualized instrument*. To build this virtualized instrument we have, in a previous work, introduced a texture mapping apparatus [14], which is also a part of our situated knowledge medium. Working with a virtual instrument rather than an actual version presents several advantages. First, numerous scientific instruments are expensive and difficult to move (e.g., they may be voluminous, heavy and/or requiring particular conditions that are not fulfill in common rooms). Consequently, even for an expert working on manufacturing the unit it might be an issue to access to a final/assembled version. Second, to draw some items over a voluminous instruments may not be easy, whereas a virtual instrument can be scaled. Third, physical constraints are set by the actual instrument while we can draw through the virtual one. The engine can render the virtualized instrument in wireframe view as well as textured view.

The current prototype requires a stylus (which is a common input tool for nowadays tablet PC) to select the few commands in order to operate the engine, thus the both hands are employed. However, we can envisage in a future work to get rid of it. For example, a voice recognition system (which is realistic for the number of commands is

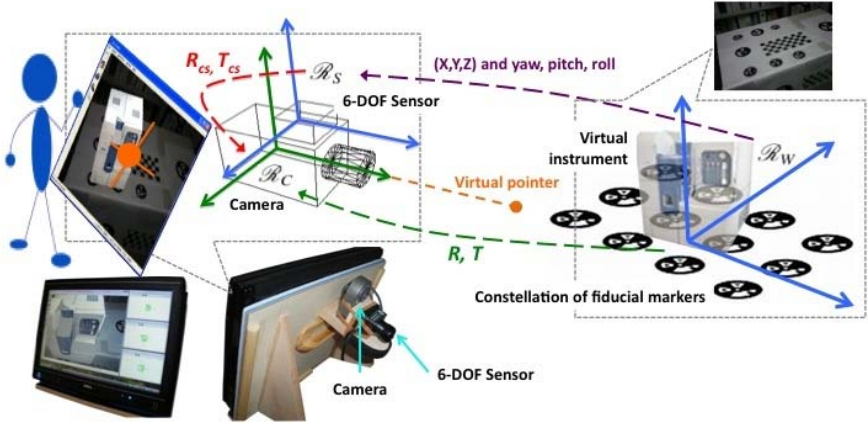


Fig. 3. 3D drawing engine

quite limited) can be implemented. Another solution is an interaction techniques such as the one described in [28], using a tilt sensor affixed to the Tablet PC.

System description. The software was created using wxWidgets (a cross-platform C++ GUI library) for the graphical interface, OpenCV for image manipulation and camera device management, OpenGL for perspective projection and 3D view manipulation, and the SQLite database for storing the 3D items. The software was tested on both Mac OS X and Windows XP/Vista operating systems; it should also compile and work as designed on the Linux OS.

The hardware consists of the Dell Latitude XT tablet PC, the Logicool QCam webcam and the IS-1200 VisTracker system (a precision 6-DOF tracker manufactured by Intersense Inc.). A constellation of fiducial markers are required to utilize the VisTracker. The inputs are the movement of the tablet PC (captured by the VisTracker) and the user's finger on the touch-screen (the current tablet PC requires a stylus). The outputs and feedbacks are given by the screen.

Note that the purpose of the VisTracker is two-fold. First, the position and orientation of the system with regard to the real-world reference frame are a necessary information to maintain the registration of the virtualized instrument, virtual pointer (see next paragraph) and 3D items in the current scene. Second, the position is used to define the 3D lines' sample points (as detailed in the next paragraph).

For registering the virtual objects in the current scene the camera pose (R and T) are needed (Fig. 3). The VisTracker measures the coordinates $(X, Y, Z)^T$ and the orientation (yaw, pitch and roll) of the sensor reference frame \mathcal{R}_S in the world reference frame \mathcal{R}_W . The relative position T_{CS} and orientation R_{CS} of the VisTracker (\mathcal{R}_S) and the video camera (\mathcal{R}_C) need to be known to compute the camera pose. To determine this transformation, we perform a calibration process that consists in computing n poses (R_i, T_i) from different viewpoints using a purely optical method (chessboard pattern recognition), simultaneously, recording the sensor data $(R_{s,i}, T_{s,i})$, and finally, finding the R_{CS} and T_{CS} that minimized the cost function $\sum_{i=1}^n (\|R_i - R_{CS}R_{s,i}\| + \|T_i - (R_{CS}T_{s,i} + T_{CS})\|)$.

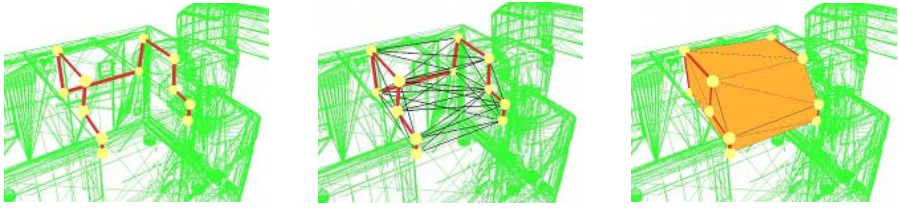


Fig. 4. 3D line defined by the sample points. Few corrections have been performed via the interface (left). Delaunay tetrahedralization mesh (center). Corresponding interpolated volume (right).

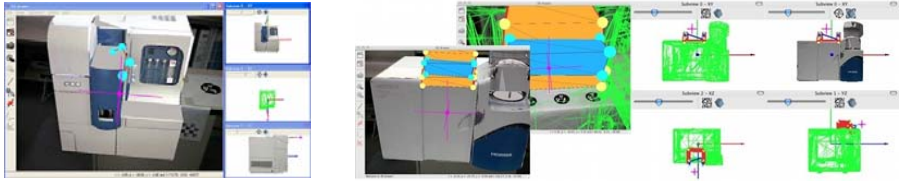


Fig. 5. Screenshot of the tablet PC (left). We can observe the pointer and a line being edited. The main view shows the current viewpoint, whereas the three other sub-views show a fixed viewpoint (e.g., front, top, left). Some additional screenshots (right), in which we can see the pointer (note that the pointer appears in the sub-views as well).

3D drawing. In order to draw a 3D line in the context of the instrument, the user moves the tablet PC in the real-world, in which the virtualized instrument is shown through the screen (Fig. 3). The 6-DOF VisTracker affixed to the tablet PC continuously measure its position, which can be recorded as a line’s sample point. As mentioned above, once a line is drawn, the user can select to interpolate a volume from the line’s sample points (Fig. 4). A Delaunay tetrahedralization is performed, which ensures that the obtained volume is a PLC.

There are two different modes for drawing a line, i.e., to acquire the sample points. The first mode is a “continuous” acquisition. While the user is moving the tablet PC, a new sample point is automatically recorded at each frame. The second mode requires the user to select for each position whether or not it should be recorded (in the current prototype, the selection is perform by pushing a button on the interface). We found that this second mode is of great convenience, especially when drawing a 3D volume, for the user does not need to concentrate in *real-time* so as to keep correct trajectory.

A virtual pointer corresponding to the current position (i.e., drawing position) is located a few dozens of centimeters behind the screen, on the line passing through the center of the screen and directed by the normal vector to the screen. It allows the user to see in the context of the instrument the position where he is drawing.

The interface consists of a main view that shows the current viewpoint and three additional sub-views showing the scene from a fix viewpoint, e.g., front, top and left (Fig. 5). This is a typical configuration adopted by common CAD software. The purpose of these sub-views are two-fold. First, to facilitate the perception of the 3D position of

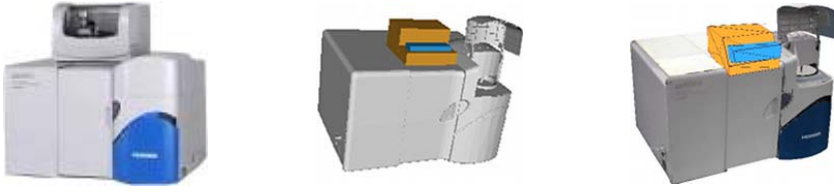


Fig. 6. Example of 3D items creation for the scenario depicted in Fig. 2. A photo of the actual combination of devices (LA950 equipped with the PowderJet) is shown (left). Comparing the result of our engine (right) against the result obtained using the CAD software *Art of illusion* (center).

the pointer. Second, to provide the user with a means of dragging a sample point so as to correct its position.

5 Preliminary Results

A preliminary evaluation of our engine consists in building the same 3D items by employing both our proposed engine and *Art of Illusion*. We consider the realistic scenario described in section 3 and depicted in Fig. 2. The goal is to create two related items to communicate knowledge about an optional device (the PowderJet) for the particle sizer LA950. Figure 6 shows a photo of the actual unit and a comparison of the both results. Although that the items created using the CAD software appear to be neater, they are the results of a consequential number of mouse/menu operations (including the realization of the items and the export/import of the CAD model in the appropriate formats) that are time-demanding and requiring the adequate skills for utilizing the CAD software. Moreover, these operations must be performed on a desktop-type computer (i.e., equipped with a multi-button mouse and a keyboard). Yet another flaw is that one of the item is not a PLC (even if the modifications required for conforming the PLC definition should be straightforward, it involves additional efforts).

In contrast, the items created utilizing our engine do not look as neat as the one created using the CAD software, but they are created from a limited number of natural movements in real-world and dragging operations to correct some of the sample points. Furthermore, no prerequisite skills are required to successfully utilize our engine. A few minutes should be enough to create the both items for a beginner, whereas several dozens of minutes may be spent when utilizing the CAD software. It is realistic to use our engine without real training. This is untrue considering the CAD software.

6 Conclusion

We have presented our low-overhead 3D items drawing engine for a situated knowledge medium. A preliminary evaluation have been conducted during which a set of 3D items are created using both an existing CAD software and our engine. Our target is not to manufacture objects, but to create 3D items in the context of a subject instrument for

communicating knowledge, hence the precision is not a main concern and rough shapes are sufficient. In this regard, our engine outperforms the CAD software in terms of required user's interactions and handiness as well as in terms of prerequisite skills.

A limitation is that only convex items can be created. Nevertheless, a combination of several items can provide an effective means of elaborating concave structures. Another drawback is the perception of the perspective through the AR interface. Although the user can visualize a virtual pointer, the distance from the pointer to another virtual structure (instrument or drawn items) can be unclear (note that it is, however, easy to notice a contact/collision). A solution to address this issue could be to use lighting effects, and particularly a virtual shadow (as suggested in [29]).

References

1. Azuma, R.: A survey of augmented reality. *Presence: Teleoperators and Virtual Environments* 6, 355–385 (1997)
2. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. *Computer Graphics and Applications* 25, 24–35 (2001)
3. Goose, S., Guven, S., Zhang, X., Sudarsky, S., Navab, N.: PARIS: Fusing vision-based location tracking with standards-based 3d visualization and speech interaction on a pda. In: *International Conference on Distributed Multimedia Systems*, San Francisco, USA, pp. 75–80 (2004)
4. Riess, P., Stricker, D.: AR-on-demand: A practicable solution for augmented reality on low-end handheld devices. In: *Workshop Virtuelle und Erweiterte Realitat der GI-Fachgruppe VR/AR*, pp. 119–130 (2006)
5. Feiner, S., Macintyre, B., Seligmann, D.: Knowledge-based augmented reality. *Commun. ACM* 36(7), 53–62 (1993)
6. Rose, E., Breen, D., Ahlers, K.H., Crampton, C., Tuceryan, M., Whitaker, R., Greer, D.: Annotating real-world objects using augmented reality. In: *Proceedings of Computer Graphics International*, Leeds, UK, pp. 357–370 (1995)
7. Reitmayr, G., Eade, E., Drummond, T.: Semi-automatic annotations in unknown environments. In: *Proc. ISMAR*, Nara, Japan, November 13–16, 2007, pp. 67–70 (2007)
8. Merckel, L., Nishida, T.: Enabling situated knowledge management for complex instruments by real-time reconstruction of surface coordinate system on a mobile device. *AI & Society* 24(1), 85–95 (2009)
9. Fischer, G., Ostwald, J.: Knowledge management: Problems, promises, realities, and challenges. *IEEE Intelligent Systems* 16(1), 60–72 (2001)
10. Merckel, L., Nishida, T.: Multi-interfaces approach to situated knowledge management for complex instruments: First step toward industrial deployment. In: *The 7th International Workshop on Social Intelligence Design*, San Juan, Puerto Rico (2008)
11. Merckel, L., Nishida, T.: Management of situated knowledge for complex instruments using 3D items creation. In: *The 22th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems*, Tainan, Taiwan (2009) (to be presented)
12. Merckel, L., Nishida, T.: Evaluation of a method to solve the perspective-two-point problem using a three-axis orientation sensor. In: *IEEE the 8th International Conference on Information Technology (CIT)*, Sydney, Australia, pp. 862–867. IEEE Computer Society Press, Los Alamitos (2008)
13. Merckel, L., Nishida, T.: Accurate object recognition using orientation sensor with refinement on the lie group of spatial rigid motions. *IEICE Transactions on Information and Systems* E91-D(8), 2179–2188 (2008)

14. Merckel, L., Nishida, T.: Low-overhead texture mapping on 3D models. In: IEEE the 9th International Conference on Information Technology (CIT), Xiamen, China. IEEE Computer Society Press, Los Alamitos (2009) (to be presented)
15. Miller, G., Talmor, D., Teng, S., Walkington, N., Wang, H.: Control volume meshes using sphere packing: Generation, refinement and coarsening. In: Fifth International Meshing Roundtable, Pittsburgh, Pennsylvania, pp. 47–61 (1996)
16. Shewchuk, J.R.: Tetrahedral mesh generation by delaunay refinement. In: Symposium on Computational Geometry, pp. 86–95 (1998)
17. Si, H., Gaertner, K.: Meshing piecewise linear complexes by constrained delaunay tetrahedralizations. In: Proceedings of the 14th International Meshing Roundtable, pp. 147–163 (2005)
18. Si, H.: On refinement of constrained delaunay tetrahedralizations. In: Proceedings of the 15th International Meshing Roundtable (2006)
19. van Dam, A.: Post-wimp user interfaces. *Communications of the ACM* 40(2), 63–67 (1997)
20. Zeleznik, R.C., Herndon, K.P., Hughes, J.F.: Sketch: an interface for sketching 3d scenes. In: SIGGRAPH 1996: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pp. 163–170. ACM Press, New York (1996)
21. Igarashi, T., Matsuoka, S., Tanaka, H.: Teddy: A sketching interface for 3D freeform design. In: ACM SIGGRAPH 1999, pp. 409–416 (1999)
22. Igarashi, T., Hughes, J.F.: A suggestive interface for 3D drawing. In: 14th Annual Symposium on User Interface Software and Technology, Orlando, Florida, pp. 173–181 (2001)
23. Masry, M., Kang, D.J., Lipson, H.: A freehand sketching interface for progressive construction of 3D objects. *Computers & Graphics* 29(4), 563–575 (2005)
24. Grossman, T., Balakrishnan, R., Kurtenbach, G., Fitzmaurice, G., Khan, A., Buxton, B.: Creating principal 3D curves with digital tape drawing. In: Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves, p. 121. ACM Press, New York (2002)
25. Fleisch, T., Brunetti, G., Santos, P., Stork, A.: Stroke-input methods for immersive styling environments. In: Proceedings Shape Modeling Applications, pp. 275–283. IEEE, Los Alamitos (2004)
26. Osaki, A., Taniguchi, H., Miwa, Y.: Direct-manipulation interface for collaborative 3D drawing in the real world. In: ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication, pp. 793–798. IEEE, Los Alamitos (2006)
27. Keefe, D., Zeleznik, R., Laidlaw, D.: Drawing on air: Input techniques for controlled 3d line illustration. *IEEE Transactions on Visualization and Computer Graphics* 13, 1067–1081 (2007)
28. Rekimoto, J.: Tilting operations for small screen interfaces. In: Proceedings of the 9th annual ACM symposium on User interface software and technology - UIST 1996, pp. 167–168. ACM Press, New York (1996)
29. Rekimoto, J.: Transvision: A hand-held augmented reality system for collaborative design. In: Proceedings of Virtual Systems and Multi-Media, Gifu, Japan, pp. 18–20 (1996)

A Method to Detect Lies in Free Communication Using Diverse Nonverbal Information: Towards an Attentive Agent

Yoshimasa Ohmoto¹, Kazuhiro Ueda², and Takehiko Ohno³

¹ Kyoto University, Graduate school of Infomatics
ohmoto@i.kyoto-u.ac.jp

² University of Tokyo, Graduate school of Arts and Sciences
ueda@gregorio.c.u-tokyo.ac.jp

³ NTT, Technology Planning Department
t.ohno@hco.ntt.co.jp

Abstract. We usually speculate partner's mental states by diverse nonverbal information. Agents need the ability for natural communication with people. In this paper, we focused on a lie as one of the typical behavior in which we often express our mental states unconsciously. The purpose of this study is to experimentally investigate the possibility of automatic lie detection in communication. We proposed an experimental setting in which participants could spontaneously decide whether or not to tell a lie. We then conducted an experiment to record participants' behavior in this setting. After that, we investigated, by discriminant analysis, that we could achieve 68% accuracy in classifying the utterances into lies and the rest without taking account of individual features by using the nonverbal behavior data. We would detect participants' stresses when they told a lie. The suggestions in this paper are useful to an agent which pays attention to user's mental states.

Keywords: Communication, nonverbal information, and lies.

1 Introduction

We have made noticeable progress in developing robots and computer agents. Humanlike robots and agents would be a reality (e.g. [15]). The abilities to communicate with people is not only the abilities to understand symbolic information such as speech and gestures. We usually speculate other people's intentions and feelings by their nonverbal expressions in free communication, which means natural communication whose partners can choose the contents of conversation spontaneously such as small talk in this study. We also nonverbally express mental states unconsciously in free communication. Those speculations and expressions are necessary for smooth communications and interactions. Therefore, agents need to understand people's mental states which are nonverbally and unconsciously expressed in many cases. In this paper, we will address possibility of automatically detecting mental states in free communication.

Some agents tried to understand user's mental states by using physiological indices. For example, [11] developed an interface application which took physiological data of a user in real-time, interpreted them as emotions, and addressed the user's affective states in the form of empathic feedback. Moreover, they also presented results from an exploratory study that aims to evaluate the impact of the Empathic Companion by measuring users' physiological data. Then, the outcome of the experiment suggests that empathic feedback has a positive effect on the interviewee's stress level while hearing the interviewer question.

It is important for smooth communication to understand user's mental states. We cannot, however, attach contact sensors on users' bodies in daily communication. Therefore, we investigated how to speculate user's mental states by using nonverbal information. Nonverbal information is a clue to speculate mental states of communication partners (e.g. [2]). In addition, we could measure nonverbal behavior in communication without contact sensors.

There are many situations which we have to speculate partners' mental states. In this paper, we focused on a lie as an expression of deceptive intention. Telling a lie is one of the typical behavior in which we often unconsciously express our stresses, such as anxieties, tensions, and so on. We do not disguise our mental states in common communication. However, we have to disguise that when we tell a lie. Therefore, detecting lies is more difficult than speculating common mental states. The understanding hidden mental states such as detecting lies is the ultimate requirement for a system in human-machine cooperation ([7]). The investigation of the method to detect hidden mental states by nonverbal information is useful to develop agents with social interaction ability. Therefore, we attempted to detect lies in free communication by nonverbal behavior which measured automatically (we call this measured data "nonverbal data" below).

It is difficult for people to detect lies ([1]). There are many previous studies on detecting lies. Polygraphs (e.g. [12]) and/or encephalometers (e.g. [8]) were often used to accurately detect lies. On the other hand, some researchers investigated how people detect lies (e.g. [5]). They mainly investigated nonverbal cues, such as gaze direction, pitch and power of a voice and so on (e.g. [4]; [3]). In these studies, participants were asked to tell lies in a controlled environment; for example, the participants first watched videos showing that a person was stealing goods (e.g. [13]) and then they were asked what they watched in interviews. In the interviews, they told lies or truths as previously decided by an experimenter and they could say nothing except direct answers to questions. However, in free communication, communication partners could spontaneously decide whether they tell lies or not, and what lies they tell when they do. The spontaneity is also one of the most significant features in general free communication. Since we could not investigate spontaneous lies in controlled environment, we proposed an experimental setting to investigate a method to detect lies in free communication, at first.

In controlled environment, Vrij et al [13] found that people could detect lies in the controlled situation by using diverse verbal and nonverbal information. However, participants could not spontaneously tell lies in the experiment. Moreover,

verbal and nonverbal cues were scored by people. Therefore, we experimentally investigated whether we could automatically detect lies in free communication by using diverse nonverbal data.

The purpose of this study is experimental investigation of the method to detect lies in free communication to find a clue to speculate mental states of communication partners. Specifically, we investigated experimentally whether or not we could automatically detect lies in free communication by using nonverbal data. Lying means to communicate something that you yourself consider untrue, with the intent to deceive another person (see [6]) in this paper. We conducted an experiment using a game in which participants could tell a lie spontaneously and intentionally, if necessary. The rest of the paper is organized as follows. Section 2 explains the experimental setting. Section 3 describes the method of analysis and results. We then discuss how to apply the suggestions to human-agent interactions in section 4. Finally, section 5 contains conclusions.

2 An Experiment for Recording People' Behavior When They Tell Lies in Free Communication

In order to analyze participants' lies statistically, we propose an experimental setting in which participants could tell lies repeatedly as well as spontaneously. We used revised Indian poker for the setting. We then conducted an experiment to record people's behavior in free communication by videos and a system to measure gaze direction and facial features.

2.1 Task

Indian poker is a kind of a card game. One card is first dealt to each player, who cannot look at its face side. Then each player has to place the card on his/her forehead, so that all the other players can see the face side of the card. This means that the players can see every card except their own. Finally all the cards are turned face up on the table (we call this operation "call a card"). The winner is the player who has the highest card. One (A) is counted as 14.

We added the following three rules in order to encourage players to communicate with other players during the game; 1) each player decides whether or not he/she "call" a card at the end of the game, 2) each player can change his/her

Table 1. List of Points

Result	Points
Defeated	-5 points
Quit the game	-3 points
Winner	+(points which the other players lose)
All players quit the game	highest card holder: -10 points other player: +5 points

card in the middle of the game, 3) players lose the fixed points when they quit the game or they are defeated (Table 1).

If each player think that his/her own card is lower than other players' cards, he/she can quit the game. When the player quits the game, he/she also lose points. In this case, however, he/she loses more points than when he/she is defeated (Table 1). The winner gets points which other players lose.

A basic strategy is that each player tries to make the other players with high-cards quit the game, and to make the other players with low-cards stay in the game. The basic strategy encourages players to communicate with others to get their own cards' information. In this communication, players tell a lie or truth.

2.2 Experimental Setting

Participants played Indian poker in the environment using Display Units (DUs). The experimental setting using DUs is shown in Fig. 1. Other participants' cards and faces were displayed on the DUs (Fig. 1(c)). The reason why we used the DUs was that we could accurately measure their nonverbal behavior in free communication. By using DUs, participants could look others' faces and their cards without the large movement of their heads. The DU was made by a half-mirror and two cameras (Fig. 1(b)). The cameras were set behind each displayed face (Fig. 1 (c)) so that participants could catch their eyes. They could also recognize where other participants looked and whether or not their eyes met. Participants were allowed to make free communication. We could thus regard that they could naturally communicate to a certain degree in this setting.

Participants were paid 1000 Yen for the two sessions they performed. In addition, they could earn a bonus prize depending on the points which they got in the game. Since good players' reward was about twice as much as bad players' one, participants could get profit if they could tell lies cleverly. Participants knew this before starting the game.

2.3 Measured Nonverbal Information

Previous studies summarized more than 150 verbal and nonverbal cues to deception ([3]). From those cues, we selected nonverbal cues which we measured in

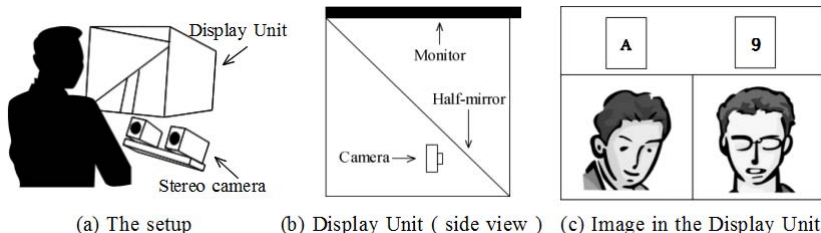


Fig. 1. Experimental setting

the experiment. First, we eliminated features which were strongly related with the meaning of a word (e.g. number of specific words and so on). The reason was that we did not focus on that in this study. Second, we also eliminated features which were strongly related with communication environment (e.g. body movements, eye blinks and so on). The reason was that communication environment was continually changed in accordance with game situations because we allowed participants to communicate freely with each other.

After that, we found typical features when participants tell lies from videos of preliminary experiments. Those features were gaze direction, prosody, and facial expression (especially, forced smile). Therefore, we measured the gazing points in a DU image (gaze direction), pitch and power of voices (prosody), 3D positions of upper and lower eyelids and corners of a mouth (facial expression).

We measured the above-mentioned nonverbal cues by using a measurement system which we made (see [10]). The accuracy of the measurement of the facial feature points was approximately ± 2 mm in 3D translation, ± 1 degree in head pose rotation. The accuracy of the gaze direction was approximately ± 2 degree.

2.4 Participants

Three participants formed a group, which will be referred as a "triad" in the rest of the paper. Each triad included one experimenter and two undergraduate or postgraduate students who were acquainted with each other. The experimenter introduced himself as a good lie detector. The reason was that participants changed their nonverbal expressions caused by stresses, such as anxieties and tensions, when they told lies. The experimenter behaved like a normal player.

Finally, we analyzed data of 18 participants, 9 males and 9 females. The total number of utterances which all participants told was 2214; lies were 653 and the rest were 1561.

2.5 Procedure

The experiment was conducted in the following manner.

1. The experimenter briefly provided instructions on the rules and strategies of the game, and the triad played a game for practice.
2. The experimenter dealt a card to each participant using the game software.
3. Each participant communicated spontaneously in order to win the game.
4. Each participant decided whether or not he/she "called" his/her card, and he/she told that he/she had decided.
5. The experimenter showed all participants' cards.
6. After the winner was decided, the losers paid their points to the winner.
7. The steps 3-7 were repeated.

Participants played Indian poker for about 80 minutes ($M = 79.9$ minutes, $SD = 8.28$), and 16.7 games ($SD = 3.88$), in an experiment. We conducted the same experiment again with the same participants after a month in order to confirm whether their nonverbal cues were consistently useful to detect lies after they got used to playing the game under this setting.

3 Statistic Analysis by Using Nonverbal Data

In this section, we investigate whether we can classify utterances into lies and the rest by using nonverbal data. For the purpose, three discriminant analyses were conducted to determine the accuracy of lie detection by using the nonverbal data which were recorded in the experiment.

3.1 Method of Analysis

Each utterance, a unit of analysis, was elicited from the voice data, which was recorded during the experiment. We call each utterance an "utterance unit."

Nonverbal data in every utterance unit was measured and recorded by using our system and a voice recorder in the experiment. The kinds of nonverbal data were the gaze direction, pitch and power of prosody, and the 3D positions of upper and lower eyelids and the corners of the mouth. The variables used to classify utterances into lies or the rest in discriminant analyses were elicited from the nonverbal data. The variables are shown in Table 2. Below, we explain how to elicit those variables from the data.

The seven variables of gaze row in Table 2 were related with the direction of gaze. In previous studies, variables of gaze direction were often encoded to "gaze aversion" (e.g. [16]). However, "gaze aversion" was not enough to describe the feature of gaze direction. We could not decide whether participants averted their eyes from communication partners face or purposefully referred to other cards because participants had to refer to such kinds of useful information for communication as others' face and others' cards in our setting. Therefore, duration of time during which they were gazing at each kind of useful information was encoded as a variable. The definitions of "the rate of gazing at the partner's face," "the rate of gazing at the other's face," "the rate of gazing at the partner's card," "the rate of gazing at the other's card," "the rate of gazing at the other

Table 2. List of variables

Nonverbal information	Code	Independent variables
gaze (seven)	G-1	The rate of gazing at the partner's face
	G-2	The rate of gazing at the other's face
	G-3	The rate of gazing at the partner's card
	G-4	The rate of gazing at the other's card
	G-5	The rate of gazing at the other place
	G-6	The number of gaze shift
	G-7	The frequency of gaze shift
prosody (six)	P-(1-3)	Pitch (the first half, the second half, change)
	P-(4-6)	Power (the first half, the second half, change)
facial expression (three)	F-1	Whether a mouth moved earlier than eyelids
	F-2	Whether eyelids moved
	F-3	Whether a mouth moved

place” are the proportion of the time, during which a participant gazed at the corresponding targets, to the total time of an utterance unit. Here ”other place” means any place except participants’ faces and cards. The definition of ”the numbers of gaze shifts” is the number of gaze shifts in an utterance unit. The definition of ”the frequency of gaze shifts” is the number of gaze shifts divided by the total time of an utterance unit.

Each utterance was divided into two parts, a first half and a second half. The averages of pitch and power in each utterance were classified into the following three categories; high (+1), middle (0), and low (-1), as variables of prosody. The procedure of coding ”pitch” as a variable in the first half of an utterance was as follows: First, the average and standard deviation (SD) of pitch in the first half of all the utterances were calculated. If a pitch average in the first half of an utterance unit was 1SD larger than the total pitch average in the first half, the ”pitch” variable in the first half of the utterance was coded as +1. If the pitch average was 1SD less than the total pitch average in the first half, the ”pitch” variable in the first half of the utterance was coded as -1. Otherwise it was coded as 0. In the same way, the variable of pitch in the second half was coded as +1, 0 or -1. The variables of power in the first half and the second half were also coded as +1, 0 or -1 as well. One of the reasons to apply this procedure was to normalize individual differences in the pitch and power of the voice. The variables of pitch and power change were coded as follows: When the value of the variable in the second half was higher than of that in the first half, the value of the change in this variable was set to +1; when lower, the value was set to -1; and when equal, the value was set to 0 (a prosody row in Table 2).

Ekman (1985) reported that subtle changes in facial expressions were good clues to detect lies. However, we could not statistically analyze whether we could use the subtle changes as clues to detect lies in the preliminary experiments because small numbers of those subtle changes were observed. Instead, we had noticed that people gave a forced smile while telling a lie in many cases. It is reported that there is a time difference between the beginning of reaction of eyes and that of a mouth in a forced smile (14). Therefore, ”whether a mouth moved earlier than eyelids” was regarded as a typical feature of facial expressions (a facial feature row in Table 2). The 3D positions of upper and lower eyelids and of the corners of a mouth were used to identify whether a smile is a forced one or not. The value was set to 1 (truth) when a mouth moved earlier than eyes by five frames in the video. Otherwise, it was set to 0 (false).

3.2 Procedure

Discriminant analyses were conducted to determine the accuracy of lie detection by using the encoded 16 variables in Table 2.

For discriminant analyses, we classified utterances into two groups. One was ”an utterance which is a lie” (hereafter, ”lie utterance” for short) and the rest was ”the rest utterances.” We defined an ”equivocal utterance” as an utterance that is neither a truth nor a lie; for instance, an ambiguous statement and a noncommittal answer. ”Equivocal utterances” accounted for 10-20% of the

whole utterances. An "equivocal utterance" was only classified into "lie utterances" when it contained a statement which was contradictory to the fact. Other "equivocal utterances" were classified into "the rest utterances."

We then arranged the nonverbal data into two data sets; 1) the whole utterances in the experiment and 2) the whole utterances of each participant. After that, a linear discriminant analysis was applied to each data set. After the analyses, we evaluated the results by 10-fold cross validation.

3.3 Results

The whole utterances in the experiment. Table 3 shows the results of a discriminant analysis by using the whole utterances in the experiment. The table shows the top five results in which a discriminant function could achieve more than 65% accuracy both in classifying utterances into lies or the rest. "Number of utterances" column shows the total number of lies and the rest. "Variables of the discriminant function" column shows the variables contained in the discriminant function, which were depicted in the form of the codes in Table 2. The signs of coefficients of the discriminant function are represented in round bracket. (+) means that high values of the variable tended to classify an utterance into a lie and (-) does vice versa. "Hit rate" column shows "hit rate of lies" and "hit rate of the rest." "Cross validation" column shows the average of the results that were obtained when 10-fold cross validation was executed three times.

The results in Table 3 show that we could achieve 68% accuracy on an average. The accuracy rates found in this study are higher than those found in the vast majority of previous deception studies by using nonverbal cues.

Compared to related works, for instance, Meservy et al [9] reported a method for deception detection through automatic analysis of nonverbal behavior. Since they only paid attention to movements of hands and the head, they failed to detect which utterance was a lie; they could only detect who was a liar. In addition, their results of cross validation in discriminant analyses reached only

Table 3. The results of discriminant analysis by the data of whole experiments

Number of utterances	Variables of the discriminant function	Hit rate	Cross validation
lie: 653 rest: 1561	G-1(-), G-3(-), G-4(-), G-5(+), G-6(-), P-1(-), P-2(+), P-3(+), P-5(-), F-1(-)	lie: 65.7% rest: 73.4%	lie: 66.9% rest: 70.2%
	G-1(-), G-2(+), G-3(-), G-6(-), P-1(-), P-2(+), P-3(+), F-1(-), F-2(-), F-3(+)	lie: 65.7% rest: 72.6%	lie: 65.2% rest: 71.8%
	G-1(-), G-2(+), G-3(-), G-6(-), P-1(-), P-2(+), P-3(+), P-5(-), P-6(+), F-1(-)	lie: 65.1% rest: 73.2%	lie: 65.4% rest: 71.4%
	G-1(-), G-2(+), G-3(-), G-6(-), P-1(-), P-2(+), P-3(+), F-1(-), F-3(+)	lie: 65.3% rest: 72.3%	lie: 64.9% rest: 71.9%
	G-1(-), G-2(+), G-3(-), G-6(-), P-1(-), P-2(+), P-3(+), P-4(-), P-6(-), F-1(-)	lie: 65.4% rest: 73.3%	lie: 64.5% rest: 72.2%

55.3%. We can say that our method could detect which utterance was a lie and achieve a certain amount of versatility to detect lies in free communication.

Bond & DePaulo [1] reported that "lie hit rate" by people who did not trained to detect lies was 48%. Ekman [5] reported that CIA agents who were best lie detectors could achieve 73% accuracy in detecting lies in a controlled environment. This result thus shows that we could correctly classify utterances in free communication automatically by using diverse nonverbal data.

We can predict a person's behavior when he/she shows the characteristics listed in the Table 3 in telling lies; "he/she gazes around except his/her partner's face and card," "he/she speaks with low and flat pitch," and "he/she gives a forced smile."

In addition, we performed SVM algorithm to classify the utterances. As a result, we could also achieve 68% accuracy in the SVM classification.

All utterances of each participant. Table 4 shows the results when utterances of every participant were individually analyzed by a discriminant analysis. The table shows the top results in which the discriminant function could achieve 65% accuracy in classifying into lies or the rest. "Hit rate" raw shows "hit rate of lies" and "hit rate of the rest." "Cross validation" raw shows an average of the results that were obtained when 10-fold cross validation was executed three times.

Table 4. The results by using all utterances of each participant (%)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Hit rate	lie	89	72	86	77	80	73	75	70	76	72	78	80	81	74	83	78	79	74
	rest	81	71	83	78	81	78	72	75	72	74	79	82	87	77	86	79	72	75
Cross validation	lie	81	68	82	74	77	71	70	51	71	68	79	83	86	68	79	79	77	73
	rest	78	71	83	76	77	76	66	61	68	68	77	80	85	68	81	77	70	74

The results in Table 4 show that we could achieve 70% accuracy in all the participants' cases. At the same time, the results of cross validation reached almost the same level except the case of Participant 8. It is, therefore, suggested that each participant consistently showed their nonverbal behavior. On the other hand, the "variables useful in the discriminant analysis" were different among all the participants. Therefore, the clues to detect lies were slightly different among the participants, and we could improve hit rates by taking account of individual features of participants.

Analysis by using data of a single modality. A linear discriminant analysis could be applied to each data set for a single modality, gaze and prosody, to confirm whether multi-modal nonverbal cues were necessary for detecting lies. Table 5 shows the result; that we conducted discriminant analyses only by using the gaze or prosody variables. The table shows the top results in which the discriminant function could correctly classify utterances into lies and the rest. "Hit rate (gaze)" and "Hit rate (prosody)" rows show "hit rate of lies" and "hit

Table 5. The results by single-modal data of each participant (%)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Hit rate (gaze)	lie	67	69	73	54	64	58	48	63	64	57	48	80	53	56	83	58	76	64
	rest	58	69	79	57	66	69	54	55	57	58	57	71	57	52	64	55	50	56
C. V. (gaze)	lie	64	69	71	50	56	58	46	63	65	56	46	77	54	54	56	54	72	63
	rest	62	63	78	57	66	68	52	56	52	56	56	75	57	51	69	53	50	57
Hit rate (prosody)	lie	64	41	73	62	50	56	31	30	44	49	69	48	72	56	58	62	59	53
	rest	94	80	78	87	85	91	97	90	84	90	79	83	84	75	86	79	82	88
C. V. (prosody)	lie	65	44	69	61	64	55	24	32	56	56	66	48	72	56	54	56	56	56
	rest	89	71	78	84	70	88	97	73	65	67	80	72	84	63	56	75	72	75

rate of the rest,” when using the gaze or prosody variables respectively. ”C. V. (gaze)” and ”C. V. (prosody)” rows show the average of the results that were obtained when 10-fold cross validation was executed three times. Bold numbers are the cases in which we could achieve more than 70% accuracy.

The results in Table 5 show that we could achieve more than 70% accuracy in classifying utterances into both of lies and the rest only in four cases out of 36 cases; Participant 3 (gaze and prosody), 12 (gaze) and 13 (prosody). In all the four cases, we could achieve more than 80% accuracy when discriminant analyses were conducted by using multi-modal variables; the gaze, prosody and facial expression variables. In the case of Participant 1 and 5, on the other hand, we could obtain only low accuracy when using single-modal variables while we could do more than 80% accuracy when using multi-modal variables. These results show that it is necessary to pay attention to multi-modal nonverbal cues.

4 Discussions and Future Work

As mentioned above, we could obtain some suggestions for lie detection. In this section, we discuss how to apply the suggestions to human-agent interactions.

One of the reason why we could detect lies by using nonverbal cues is that participants change their nonverbal expressions under stresses, such as anxieties, tensions, and so on, when people tell a lie. In other words, we would detect participants’ stresses when they told a lie. In fact, we could observe that participants often got flustered and nervous in the experiment. It is helpful for smooth interactions with agents to detect user’s stresses (e.g. [11]). For example, an agent may detect user’s stress when they can detect the nonverbal cues to deception in communication in which the user does not need to tell a lie. The agent can then abort and/or change his/her action to relieve user’s stress. We could achieve 68% accuracy in this study. The ratio was not enough to automate lie detection in real critical situations such as criminal investigations. However, we expect that the accuracy rate is enough to detect user’s stress in communication.

We confirmed that the hit rates of lie detection were improved by taking account of individual features which were expressed by diverse nonverbal behavior. A user’s body, furnitures, and/or other objects prevent an agent from measuring part of user’s nonverbal behavior in real situations. The agent would be able

to measure some nonverbal cues out of diverse nonverbal cues. Therefore, the agent can improve his/her judgement through long-term interactions by storing the individual features. We may then realize an agent which pay careful attention to users by detecting his/her stress accurately. However, the individual features were very different even in this study. Therefore, the agent which can pay attention to user's stress should communicate verbally in real situations.

The suggestions in this paper are useful to an attentive agent which pays attention to user's stress. As a matter of course, there are some issues that need to be solved to realize that. For example, we have to store the data of individual behavior over a long time in order to detect user's mental states accurately. However, it may have ethical and/or technical problems to store the data over a long time. We also have to investigate what is appropriate behavior of an agent when user feels stress. Furthermore, we may have to confirm that the suggestions in this paper can be applied to user's stress detection. These are future works.

5 Conclusions

In this paper, we investigated whether our proposed method could automatically detect lies by using nonverbal data. The results showed that our method could achieve 70% accuracy in classifying into utterances into lies and the rest without taking account of individual features of participants.

The significant variables for the discrimination included these which were not assumed significant in the previous studies. It means that people express different nonverbal behavior between in free communication and in a controlled environment when people tell lies. We could improve the discrimination ratio by taking account of individual features. It is also necessary to pay attention to multimodal nonverbal behavior to detect lies in most cases.

Therefore, we can say that we could automatically classify utterances into lies and the rest in free communication by using diverse nonverbal data. The suggestions in this paper are useful to an agent which pays attention to user's stress. We think there are some problems still. The suggestions in this paper can be considered a first step toward realizing an automated system to detect lies in free communication.

References

1. Bond Jr., C.F., DePaulo, B.M.: Accuracy of Deception Judgments. *Personality and Social Psychology Review* 10(3), 214–234 (2006)
2. Daibou, I.: The social meaning of interpersonal communication. *Japanese Journal of Interpersonal and Social Psychology* 1, 1–16 (2001)
3. DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to deception. *Psychological Bulletin* 129(1), 74–118 (2003)
4. Ekman, P.: *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W.W. Norton & Company, New York (1985)
5. Ekman, P., O'Sullivan, M., Frank, M.G.: A few can catch a liar. *Psychological Science* 10, 263–266 (1999)

6. Hopper, R., Bell, R.A.: Broadening the deception construct. *Quarterly Journal of Speech* 70, 288–302 (1984)
7. Kanno, T., Keiichi, N., Kazuo, F.: A method for team intention inference. *International Journal of Human-Computer Studies* 58, 393–413 (2003)
8. Kozel, F.A., Johnson, K.A., Mu, Q., Grenesko, E.L., Laken, S.J., George, M.S.: Detecting Deception Using Functional Magnetic Resonance Imaging. *Biological Psychiatry* 58(8,15), 605–613 (2005)
9. Meservy, T.O., Jensen, M.L., Kruse, J., Twitchell, D.P., Tschepnakis, G., Burgoon, J.K., Metaxas, D.N., Nunamaker, J.F.: Deception Detection through Automatic, Unobtrusive Analysis of Nonverbal Behavior. *IEEE Intelligent Systems* 20(5), 36–43 (2005)
10. Ohmoto, Y., Ueda, K., Ohno, T.: Real-time system for measuring gaze direction and facial features: towards automatic discrimination of lies using diverse nonverbal information. *AI & Society* 23(2), 187–200 (2007)
11. Prendinger, H., Ishizuka, M.: The empathic companion: a character-based interface that addresses users' affective states. *Applied Artificial Intelligence* 19(3/4), 267–285 (2005)
12. Raskin, D.C., Honts, C.R.: The comparison question test. In: Kleiner, M. (ed.) *Handbook of Polygraph Testing*. Academic Press, London (2002)
13. Vrij, A., Akehurst, L., Soukara, S., Bull, R.: Detecting Deceit via Analyses of Verbal and Nonverbal Behaviour in Children and Adults. *Human Communication Research* 30(1), 8–41 (2004)
14. Yotsukura, T., Uchida, H., Yamada, H., Akamatsu, S., Tetsutani, N., Morishima, S.: A Micro-Temporal Analysis of Facial Movements and Spontaneously Elicited and Posed Expressions of Emotion Using High-Speed Camera. *IEICE technical report. Image engineering*. 101(300), 15–22 (2001)
15. Walters, M.L., Syrdal, D.S., Dautenhahn, K., te Boekhorst, R., Koay, K.L.: Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots* 24(2), 159–178 (2008)
16. Zuckerman, M., DePaulo, B.M., Rosenthal, R.: Verbal and nonverbal communication of deception. In: Berkowitz, L. (ed.) *Advances in experimental Social Psychology*, vol. 14, pp. 1–59 (1981)

An Integrative Agent Model for Adaptive Human-Aware Presentation of Information during Demanding Tasks

Andy van der Mee¹, Nataliya M. Mogles², and Jan Treur²

¹ Force Vision Lab, Barbara Strozziilaan 362a, 1083 HN Amsterdam, The Netherlands

² VU University Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam

andy@forcevisionlab.nl, mogles1@yahoo.com, treur@few.vu.nl

Abstract. This paper presents an integrative agent model for adaptive human-aware information presentation. Within the agent model, meant to support humans in demanding tasks, a domain model is integrated which consists of a dynamical model for human functioning, and a model determining the effects of information presentation. The integrative agent model applies model-based reasoning methods to the domain model to analyze the state of the human and to determine how to adapt the presentation of information to this state.

1 Introduction

When a human is performing a demanding task, often support can be offered by presenting information that is relevant for the task. For example, a naval officer or air traffic controller may be offered visualized information on location and speed of several objects in the environment, and of the characteristics of such objects. Other cases are when information is presented on the workflow that is being followed, and on the status of the different tasks in a workflow, or relevant task information such as manuals for systems used. In all of such cases the human may take the initiative, for example, by activating certain options using menu structures. However, especially when tasks require a high level of attention and concentration, it is more beneficial when the human does not need to bother about such presentation aspects, by giving the system itself an active role in offering information of appropriate types and forms.

Adaptive information presentation can provide a useful type of support in a number of application contexts, varying from tourists in a museum (e.g., [17, 20]) and users in hypermedia and Web contexts (e.g., [22]), to students using educational systems (e.g., [13]) and humans in demanding tasks (e.g., [8, 10]). A main requirement for an adaptive information presentation system is that it presents information in types and forms that are strongly depending on these circumstances. Here circumstances may involve a number of aspects, for example (see also, e.g., [7, 19]): (1) the characteristics of the task, (2) the characteristics of the human involved, such as expertise level with respect to the task, (3) the state of the environmental context (4) task status and task progress, and (5) the cognitive, affective or functional state of the human. Here (1) and (2) may be considered static over longer time periods, but (3), (4) and (5) usually have a highly dynamic nature. To take such aspects into account an adequate presentation system has to be highly adaptive and has to be constantly aware of them.

Awareness of the state of the human, the task and the environment can in part be based on observation and sensing information acquired. However, often awareness is required on aspects for which information cannot be acquired in a direct manner, for example, the level of anxiety, stress and exhaustion of the human, or the progress on the task. In such cases dynamical process models can be used to relate information that is directly acquired to information about aspects that are not directly accessible. In this paper an integrative agent model for an adaptive human-aware presentation system for humans in demanding tasks is presented that makes use of a dynamical model of human functioning, in particular to monitor the human's functional state (covering aspects such as exhaustion and experienced work pressure), combined with a model to determine the effects of information presentation. In Section 2 first the context is described in some more detail. A computational domain model is introduced in Section 3. Section 4 introduces the overall architecture of the integrative agent model. Section 5 presents simulation results based on one example scenario. Finally, Section 6 is a discussion.

2 On Adaptivity in Information Presentation

The context of the research reported in this paper is the domain of naval operations. An operator on a naval ship has to decide on and perform actions within limited time. The results of these actions can be critical for the result of the entire operation and can even be critical for self preservation, so besides timeliness, quality of human task performance is also essential.

Given this context and the inherent fallibility of human task performance, automated support for operators in strenuous situations is an interesting topic of research that is likely to be beneficial. This kind of support cannot only be provided at the team level, but also on an individual level. An example of support at the individual level is *adaptive information presentation*, in which information presented to an operator is personalized and adapted to his specific circumstances. This last kind of support is explored in this paper.

The main principles of design of information presentation in displays are extensively described in literature on human information processing and human factors; e.g., see [12, 25]. It is well established in this literature that a good display design can enhance information processing and improve human performance. However, this conventional display design is based on the general characteristics of human information processing and aims to serve an average person performing a particular type of a task. It usually does not consider personal characteristics and dynamic, constantly changing environments and human functional states. The goal of the research reported here is to incorporate principles of information presentation in a dynamic model along with such factors as operator's functional states, environmental and task characteristics. The integrative model presented in this article will represent the relations between these factors and human functioning while performing a task.

Cognitive performance is affected by the human's activation state, or alertness. Alertness is a physiological state that affects the attentional system and varies depending on internal and external factors [23]. Besides alertness, cognitive performance is also influenced by human information processing aspects, such as perception and

working memory [25]. It is well-established that bright light rapidly increases alertness [18]. Therefore one of the assumptions underlying the work reported here is that the level of brightness, or *luminance*, may have an effect on alertness of an operator. Another characteristic of a display that may affect alertness is the *background colour* [21]. The *time of the day* is an environmental aspect that can also influence alertness according to numerous findings that relate alertness and performance to circadian rhythms. It is found that the activation of central nervous system passes through different stadia during the day according to the inner clock in a brain [25]. Fatigue, the physiological and psychological state of tiredness and dislike of present activity, is one of the aspects that influence a person's functioning [23]. It may be assumed that *exhaustion* has also negative influence on the alertness level as exhaustion is placed on a higher level of tiredness-fatigue-exhaustion continuum. Exhaustion as a factor that affects a person's functioning while performing a critical task is also mentioned in the functional state model presented in [2]. It is also found that motivation and alertness are correlated [11].

The findings below describe the relations between different factors of information presentation and processing demands. Display *luminance* affects visual search performance with monitor displays without affecting detection performance significantly [14]. According to Badderley's theory about the working memory, if the visuo-spatial sketchpad buffer of working memory is totally occupied by the processing of visuo-spatial information during the execution of a task, no more visual information can be perceived and processed [1]. In this case presenting information in another modality, auditory for instance, will lead to less processing demand if a task being performed requires predominately visuo-spatial resources, but will lead to more processing demand if a task is predominantly auditory. This principle is applied in the PACE (Performance Augmentation through Cognitive Enhancement) system architecture developed for the conditions of high stress and workload and presented in [16]. The *grouping* of numerous objects imposes less processing demand because attention resources are applied on the groups of objects at certain locations rather than on the whole field of a display with the isolated objects [25]. Symbol size plays a role in processing demand too. The larger the symbols are, the easier it is to process them, but after a certain threshold there is no gain in processing anymore [6]. It may be hypothesized that the processing of objects is performed in the same way: the larger the objects, the easier it is to process them. On the other hand, it is obvious that the more objects occur in a display and the larger they are, the more processing demand may be imposed as the objects become less distinct and more difficult to perceive.

3 A Domain Model for Functional State and Presentation Aspects

In this section the domain model used is presented, which consists of two interacting dynamical models, one to determine the human's functional state and one to determine the effects of the chosen type and form of information presentation. The approach used to specify the domain model is based on the hybrid dynamical modeling language LEADSTO [4]. In this language, direct temporal dependencies between two state properties in successive states are modeled by *executable dynamic properties*. The LEADSTO format used here is defined as follows. Let α and β be state properties. In the LEADSTO language the notation $\alpha \rightarrow_D \beta$, means:

If state property α holds at some time t , then state property β will hold at time $t+D$

Here, state properties can have a qualitative, logical format, or a quantitative, numerical format, and may consist of conjunctions of atomic state properties.

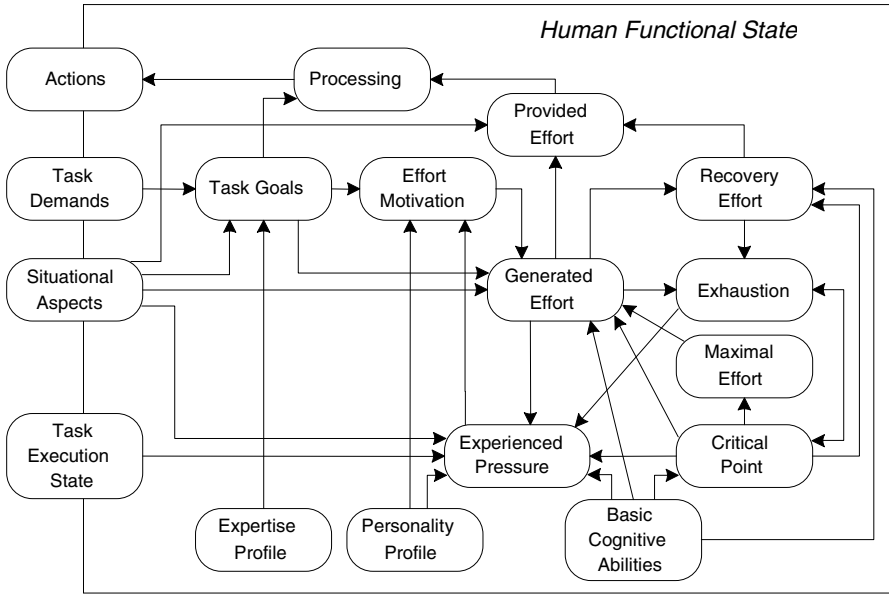


Fig. 1. Functional state domain model

The dynamic model for the functional state used was adopted from [2]; for a global picture, see Figure 1. Here the functional state is defined as the combination of exhaustion (fatigue), motivation, experienced pressure, and effort. These are determined by external factors such as task demands and the state of the environment, and by personal factors such as experience, cognitive abilities and personality profile. Originally the model was implemented in MatLab. For the work reported here it was remodeled in LEADSTO and integrated within the agent model, as discussed in Section 4. On the one hand this model is based on the (informal) cognitive energetic framework [9], that relates effort regulation to human resources in dynamic conditions. On the other hand, the model is based on literature on fatigue in exercise and sports as formalized by a computational model in [24], in particular on the concept *critical power*, which is the maximal effort level a person can (constantly) maintain over a longer time period without becoming (fully) exhausted.

The arrows in Figure 1 denote causal dependencies; note that cycles occur. For example, generated effort is affected by the person's motivation level (*effort motivation*), the amount of effort the task requires (*task level*) and the effort the human is able to contribute (*critical point* and *maximal effort*). When generated effort is above the critical point, the exhaustion is increased. When generated effort is below the critical point, some recovery takes place (*recovery effort*), thus decreasing *exhaustion*. Effort contributed to cope with noise in the environment (*noise effort*) is extracted from the

generated effort, so that the effort that can effectively be contributed to the task (provided effort) is less. The motivation is taken proportional to the task level, but also depends on the *experienced pressure*. An *optimal experienced pressure* is assumed which depends on the *personality profile*. The dynamical model has been formalized as a system of differential equations. For more details of this model, see [2].

The interaction from the model for information presentation to the model for functional state takes place by affecting the task demands. Conversely, a number of aspects of the functional state are used as input by the information presentation model: effort motivation, exhaustion, experienced pressure and provided effort. Figure 2 shows an overview of the information presentation model.

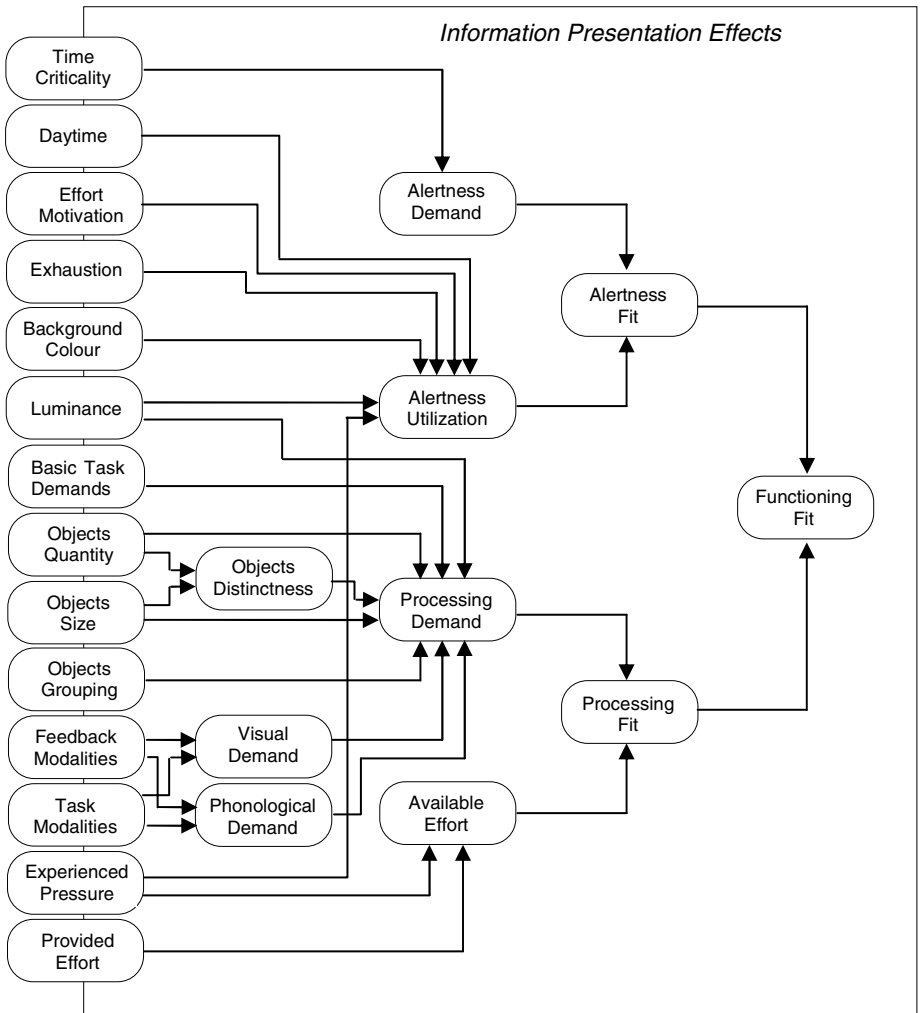


Fig. 2. Information presentation effect domain model

The general paradigm of the relations within the presentation model is partially based on the existing models on workload that consider the fit between individual factors, such as coping capacity, effort, motivation, on one side and work demands on the other side. One example of such a model can be found in [15]. This paradigm has been applied to the fit between the effort that a human is willing to invest while performing a task and demand. Effort is determined by internal and external factors while demand is imposed externally.

Presentation format aspects can be seen as a part of task demands that are imposed on a person because a form of a presentation may change processing demands. On the other hand, some presentation aspects, for example, background colour and luminance, can be seen as available resources that help a person to perform a task. Luminance is regarded both as a part of demands and as a part of resources in this model. All types of aspects are converged into two more global internal factors that influence the task performance: physiological state of *alertness* and mental *information processing* state of an operator. Among these concepts a distinction is made between the states of available and used recourses of alertness and information processing, *alertness utilization* and *provided effort* respectively, and the states of demand for alertness and information processing, *alertness demand* and *processing demand*. The fit between the usage of these capacities and the demands determines the functioning of a human while performing a task, the *functioning fit*. Two specific types of fit are considered: *alertness fit* and *processing fit*.

If the usage of capacities and demands are at the same level, the fits will be high. If the levels of capacities and demands differ much, then the fits will be low. If both *alertness fit* and *processing fit* are high, then the *functioning fit* will be high.

All inputs for the model are represented by numbers between 0 and 1. The same holds for the concepts *objects distinctness*, *visual demand*, *phonological demand*, *alertness demand*, *alertness utilization*, *processing demand*, and *available effort*. The concept *alertness fit* indicates the difference between alertness demand and alertness utilization and is represented by a number between -1 and 1. The same holds for *processing fit* which is the difference between available effort and processing demand. This was expressed in LEADSTO as can be found in Appendix A.

4 Overall Architecture of the Information Presentation System

For the overall architecture of the integrative agent model, principles of component-based agent design have been followed, as, for example, used within the agent design method DESIRE; cf [5]. Within the agent model two main components have been distinguished: the analysis component and the support component (see Figure 3). Accordingly, two different ways to integrate the domain models within the agent model have been used; see Figure 3.

- *analysis component*
To perform analysis of the human's states and processes by (model-based) reasoning based on observations and the domain model.
- *support component*
To generate support actions for the human by (model-based) reasoning based on observations and the domain model.

Within these components of the agent model, the domain model has been integrated which by itself consists of two (dynamical) models, as described in Section 3: a model for the functional state of the human and a model for the effects of information presentation. By incorporating such domain models within an agent model, an integrative agent model is obtained that has an understanding of the processes of its surrounding environment, which is a solid basis for knowledgeable intelligent behaviour. Note that here the domain model that is integrated refers to one agent (the human considered), whereas the agent model in which it is integrated refers to a different agent (the ambient software agent).

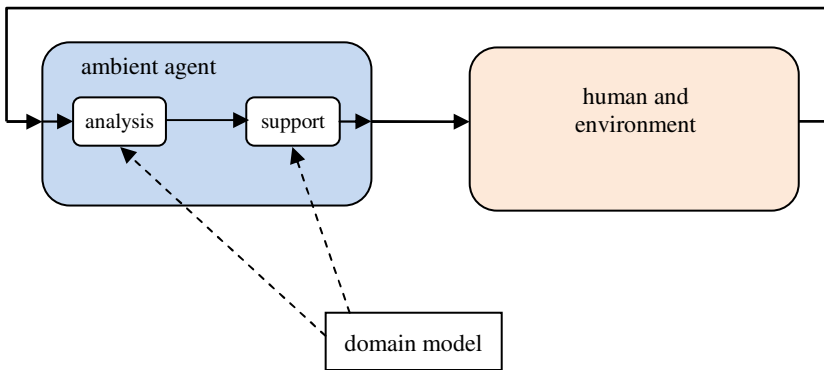


Fig. 3. Overall design of the ambient agent and the integration of the domain model. Here solid arrows indicate information exchange (data flow) and dotted arrows the integration of the domain model within the agent model.

Analysis Component

Within the analysis component, by model-based reasoning forward in time based on the domain model, predictions are made about future states of the human and the environment. The integration of the domain model relationships within such an analysis model for model-based reasoning forward in time is done in a systematic manner by replacing the atoms in a domain model relationship, for example

$$\text{has_value}(a, V_1) \ \& \ \text{has_value}(b, V_2) \ \rightarrow_D \ \text{has_value}(c, f(V_1, V_2))$$

with $f(V_1, V_2)$ a function of V_1 and V_2 by predictions of the ambient agent about them:

$$\begin{aligned} &\text{predicted_value_for}(a, V_1, t) \ \& \ \text{predicted_value_for}(b, V_2, t) \\ &\rightarrow \text{predicted_value_for}(c, f(V_1, V_2), t+D) \end{aligned}$$

An example of a function $f(V_1, V_2)$ is a weighted sum function with weights w_1 and w_2 : $f(V_1, V_2) = w_1 * V_1 + w_2 * V_2$. A more detailed description of the analysis component is given in Appendix B.

Support Component

Within the support component model-based reasoning based on the domain model takes place in a goal-directed manner, backward in time starting from desired (adjusted) future states. Within the support component this model-based reasoning can be done in a qualitative manner or in a quantitative manner. The former case is shown in Appendix C, where based on the causal graph as depicted in Figure 2, desires to increase or decrease values are derived (from right to left, against the direction of the arrows), in a heuristic manner without specifying numerically how much the increases or decreases should be. Below it is shown how a quantitative approach can be used, based on the more precise numerical relations of the information presentation model. In this case the integration of a domain model relationship within a support model for model-based reasoning backward in time can be done in a systematic manner by embedding some atoms in a domain model relationship in adjustment desires and some in beliefs and reversing the order, for example,

$$\text{has_value}(a, V_1) \ \& \ \text{has_value}(b, V_2) \ \rightarrow_D \ \text{has_value}(c, f(V_1, V_2))$$

for the case that the attribute b is kept fixed (not adjusted) is transformed into:

$$\text{desire_for}(c, V_3, t+D) \ \& \ \text{belief_for}(b, V_2, t) \ \rightarrow \ \text{desire_for}(a, g(V_2, V_3), t)$$

where $g(V_2, V_3)$ is a function of V_2 and V_3 that inverts the function $f(V_1, V_2)$ with respect to its first argument: $f(g(V_2, V_3), V_2) = V_3$ and $g(V_2, f(V_1, V_2)) = V_1$. For the example of a function $f(V_1, V_2)$ as a weighted sum with weights w_1 and w_2 the inverse function is found as follows:

$$\begin{aligned} f(V_1, V_2) = w_1 * V_1 + w_2 * V_2 \Leftrightarrow V_3 = w_1 * V_1 + w_2 * V_2 \Leftrightarrow w_1 * V_1 = V_3 - w_2 * V_2 \Leftrightarrow V_1 = (V_3 - w_2 * V_2) / w_1 \\ \Leftrightarrow g(V_2, V_3) = (V_3 - w_2 * V_2) / w_1. \end{aligned}$$

It is also possible to distribute a desire for adjustment over adjustment desires for multiple attributes. Suppose as a point of departure an adjustment Δv_l is desired, and that v_l depends on two variables v_{l1} and v_{l2} that are adjustable (the non-adjustable variables can be left out of consideration). Then by elementary calculus the following linear approximation can be obtained:

$$\Delta v_l = \frac{\partial v_l}{\partial v_{l1}} \Delta v_{l1} + \frac{\partial v_l}{\partial v_{l2}} \Delta v_{l2}$$

This is used to determine the desired adjustments Δv_{l1} and Δv_{l2} from Δv_l , where by weight factors μ_{l1} and μ_{l2} the proportion can be indicated in which the variables should contribute to the adjustment: $\Delta v_{l1} / \Delta v_{l2} = \mu_{l1} / \mu_{l2}$. Since

$$\Delta v_l = \frac{\partial v_l}{\partial v_{l1}} \Delta v_{l2} \mu_{l1} / \mu_{l2} + \frac{\partial v_l}{\partial v_{l2}} \Delta v_{l2} = \left(\frac{\partial v_l}{\partial v_{l1}} \mu_{l1} / \mu_{l2} + \frac{\partial v_l}{\partial v_{l2}} \right) \Delta v_{l2}$$

then the adjustments can be made as follows:

$$\Delta v_{l2} = \frac{\Delta v_l}{\frac{\partial v_l}{\partial v_{l1}} \mu_{l1} / \mu_{l2} + \frac{\partial v_l}{\partial v_{l2}}} \quad \Delta v_{l1} = \frac{\Delta v_l}{\frac{\partial v_l}{\partial v_{l1}} + \frac{\partial v_l}{\partial v_{l2}} \mu_{l2} / \mu_{l1}}$$

Special cases are $\mu_{11} = \mu_{12} = 1$ (*absolute equal contribution*) or $\mu_{11} = v_{11}$ and $\mu_{12} = v_{12}$ (*relative equal contribution*: in proportion with their absolute values). As an example, consider again a variable that is the weighted sum of two other variables: $v_1 = w_{11}v_{11} + w_{12}v_{12}$. For this case, the partial derivatives are w_{11} respectively w_{12} ; therefore

$$\Delta v_{11} = \frac{\Delta v_1}{w_{11} + w_{12} \mu_{12}/\mu_{11}} \quad \Delta v_{12} = \frac{\Delta v_1}{w_{11} \mu_{11}/\mu_{12} + w_{12}}$$

When $\mu_{11} = \mu_{12} = 1$ this results in $\Delta v_{11} = \Delta v_{12} = \Delta v_1 / (w_{11} + w_{12})$, and when in addition the weights are assumed normalized, i.e., $w_{11} + w_{12} = 1$, then it holds $\Delta v_{11} = \Delta v_{12} = \Delta v_1$. Another setting is to take $\mu_{11} = v_{11}$ and $\mu_{12} = v_{12}$. In this case the adjustments are assigned proportionally; for example, when v_1 has to be adjusted by 5%, also the other two variables on which it depends need to contribute an adjustment of 5%. Thus the relative adjustment remains the same through the backward desire propagations:

$$\frac{\Delta v_{11}}{v_{11}} = \frac{\Delta v_1}{w_{11} + w_{12} v_{12}/v_{11}} / v_{11} = \frac{\Delta v_1}{v_1}$$

This shows a general approach on how desired adjustments can be propagated in a backward manner using a domain model. For a detailed description of the support component see Appendix C.

5 Simulation Results

In order to analyse the behaviour of the integrative agent model, a number of simulations have been performed using the LEADSTO software environment; cf. [4]. The model exhibits behaviour as expected: after the assessment of alertness and/or processing fit as inadequate, the agent performs the relevant manipulations of information presentation aspects. As a result of the manipulations, both alertness and processing fits that are constituents of functioning fit are improved. As a consequence of alertness and processing fit improvement, functioning fit that represents general task performance becomes also better. For example, in the simulation depicted in Figure 4, it can be seen that after the manipulations of the ambient agent functioning fit, alertness fit and processing fit have improved. Time flow is represented on the horizontal axis and the values of alertness fit, processing fit and functioning fit are represented on the vertical axis. Dark bars in the figure represent the time intervals when a certain statement is true.

6 Discussion

Adaptive information presentation involves presenting information in types and forms that are strongly depending on circumstances, which may comprise a number of aspects (e.g., [7, 19]). Some of these aspects are considered constant over longer time periods (e.g., personality characteristics or preferences), and often can be estimated in an accurate manner progressively over time, using some type of (machine) learning method. Other aspects may be more dynamic: they may change all the time. Such a moving target is not easy to estimate in an accurate manner at each point in time. One

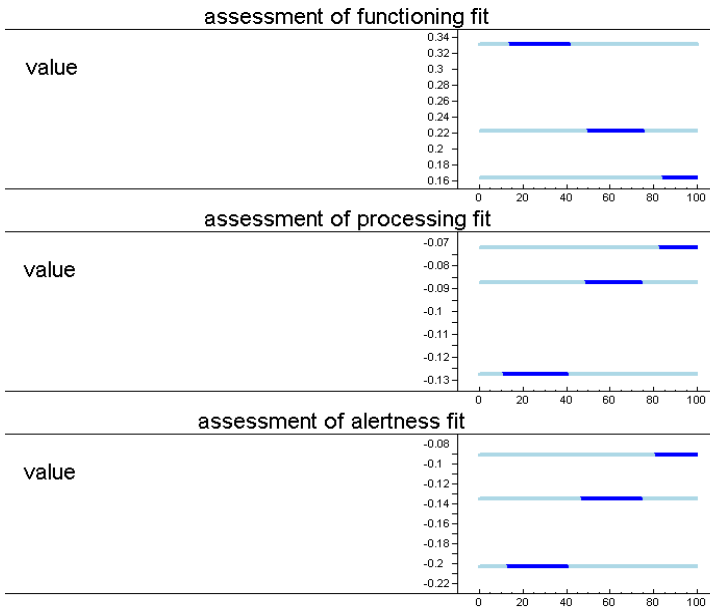


Fig. 4. Simulation trace: alertness fit assessment ‘demand dominance’, processing fit assessment ‘demand dominance’

way that is sometimes exploited assumes that there are attributes (e.g., by sensors) observable at each point in time that directly relate (in a non-temporal manner) to the aspect to be estimated. For example, in [10] the human’s anxiety state is determined in a non-temporal knowledge-based manner from monitor information. However, such attributes are not always available. A more general case is that there are relevant observable attributes, but they do not directly relate to the aspect to be estimated in a non-temporal manner, but instead, temporal, dynamic relations are available. This is the case addressed in the current paper. Model-based reasoning methods have been exploited by applying them to a dynamic model relating a human’s functional state to information presentation aspects and task performance.

Other approaches to adaptive information presentation often address the human’s characteristics and preferences; e.g., [17, 20, 22]. Such approaches usually do not address the human’s cognitive, affective or functional state, which within one session may show much variation over time. For use within educational systems the learner’s actions and progress can be monitored to get an estimation of the learner’s cognitive load (e.g., [13]). Especially for humans in demanding tasks monitoring the human’s cognitive, affective or functional state, and adapting information presentation based on this monitoring information may be crucial. As already mentioned, in [10] the human’s anxiety state is determined in a non-temporal knowledge-based manner from monitor information. In contrast to such approaches, the approach presented in the

current paper makes use of causal, dynamical domain models for the human's functional state and the information presentation aspects, and generic model-based reasoning methods applied to these models.

References

1. Baddeley, A.: Exploring the Central Executive. *Quarterly Journal of Experimental Psychology* 49, 5–28 (1996)
2. Bosse, T., Both, F., van Lambalgen, R., Treur, J.: An Agent Model for a Human's Functional State and Performance. In: Jain, L., Gini, M., Faltings, B.B., Terano, T., Zhang, C., Cercone, N., Cao, L. (eds.) *Proceedings of the 8th IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2008*, pp. 302–307. IEEE Computer Society Press, Los Alamitos (2008)
3. Bosse, T., Duell, R., Hoogendoorn, M., Klein, M.C.A., van Lambalgen, R., van der Mee, A., Oorburg, R., Sharpanskykh, A., Treur, J., de Vos, M.: An Adaptive Personal Assistant for Support in Demanding Tasks. In: Schmorow, D.D., et al. (eds.) *HCI 2009. LNCS*, vol. 5638, pp. 3–12. Springer, Heidelberg (2009)
4. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. *Intern. Journal of Artificial Intelligence Tools* 16, 435–464 (2007)
5. Brazier, F.M.T., Jonker, C.M., Treur, J.: Principles of Component-Based Design of Intelligent Agents. *Data and Knowledge Engineering* 41, 1–28 (2002)
6. Chung, S.T.L., Stephen Mansfield, J.S.J., Legge, G.E.: Psychophysics of Reading. XVIII. The Effect of Print Size on Reading Speed in Normal Peripheral Vision. *Vision Research* 38, 2949–2962 (1998)
7. De Carolis, B., Di Maggio, P., Pizzutilo, S.: Information Presentation Adapted to the User in Context. In: Esposito, F. (ed.) *AI*IA 2001. LNCS (LNAI)*, vol. 2175, pp. 314–319. Springer, Heidelberg (2001)
8. Fricke, N.: Effects of Adaptive Information Presentation. In: *Proc. of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pp. 292–298 (2007)
9. Hockey, G.R.J.: Compensatory Control in the Regulation of Human Performance under Stress and High Workload: a Cognitive-Energetical Framework. *Biological Psychology* 45, 73–93 (1997)
10. Hudlicka, E., McNeese, M.D.: Assessment of User Affective and Belief States for Interface Adaptation: Application to an Air Force Pilot Task. *User Modeling and User-Adapted Interaction* 12, 1–47 (2002)
11. Hull, J., Wright, K.P., Charles Jr., A., Czeisler Jr., C.A.: The Influence of Subjective Alertness and Motivation on Human Performance Independent of Circadian and Homeostatic Regulation. *Journal of Biological Rhythms* 18(4), 329–338 (2003)
12. Johnson, A., Proctor, R.W.: *Attention: Theory and Practice*. Sage Publications, CA (2003)
13. Kashihara, A., Kinshuk, Oppermann, R., Rashev, R., Simm, H.: A Cognitive Load Reduction Approach to Exploratory Learning and Its Application to an Interactive Simulation-Based Learning System. *Journal of Educational Multimedia and Hypermedia* 9, 253–276 (2000)

14. Krupinski, E.A., Roehrig, H., Furukawa, T.: Influence of Film and Monitor Display Luminance on Observer Performance and Visual Search. *Academic Radiology* 6(7), 411–418 (1999)
15. Macdonald, W.: The impact of Job Demands and Workload on Stress and Fatigue. *Australian Psychologist* 38(2), 102–117 (2003)
16. Morizio, N., Thomas, M., Tremoulet, P.: Performance Augmentation through Cognitive Enhancement (PACE). In: *Proceedings of the International Conference on Human Computer Interaction, Las Vegas* (2005)
17. Oppermann, R., Specht, M.: A Context-sensitive Nomadic Information System as an Exhibition Guide. In: Thomas, P., Gellersen, H.-W. (eds.) *HUC 2000. LNCS*, vol. 1927, pp. 127–142. Springer, Heidelberg (2000)
18. Penn, P.E., Bootzin, R.R.: Behavioural Techniques for Enhancing Alertness and Performance in Shift Work. *Work & Stress* 4, 213–226 (1990)
19. Sarter, N.: Coping with Complexity Through Adaptive Interface Design. In: Jacko, J.A. (ed.) *HCI 2007. LNCS*, vol. 4552, pp. 493–498. Springer, Heidelberg (2007)
20. Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., Kuflik, T., Not, E., Rocchi, C.: Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Modeling and User-Adapted Interaction* 17, 257–304 (2007)
21. Stone, N., English, A.: Task Type, Posters, and Workspace Colour on Mood, Satisfaction, and Performance. *Journal of Environmental Psychology* 18, 175–185 (1998)
22. Tarpin-Bernard, F., Habieb-Mammar, H.: Modeling Elementary Cognitive Abilities for Adaptive Hypermedia Presentation. *User Modeling and User-Adapted Interaction* 15, 459–495 (2005)
23. Thiffault, P., Bergeron, J.: Fatigue and Individual Differences in Monotonous Simulated Driving. *Personality and Individual Difference* 34, 159–176 (2003)
24. Treur, J.: A Virtual Human Agent Model with Behaviour Based on Feeling Exhaustion. In: Chien, B.C., Ali, M., Chen, S.M., Hong, T.P. (eds.) *IEA-AIE 2009. LNCS*, vol. 5579, pp. 11–23. Springer, Heidelberg (2009)
25. Wickens, C.D., Lee, J.D., Liu, Y., Gordon-Becker, S.E.: *An Introduction to Human Factors Engineering*, pp. 184–217. Prentice Hall, Upper Saddle River (2004)

Appendix A: Detailed Specification of the Domain Model

If alertness utilization has value V_1 and alertness demand value V_2 ,
 then alertness fit has value V_1-V_2
 $\text{has_value}(\text{alertness_utilization}, V_1) \ \& \ \text{has_value}(\text{alertness_demand}, V_2)$
 $\rightarrow \text{has_value}(\text{alertness_fit}, V_1-V_2)$

If available effort has value V_1 and processing demand value V_2 ,
 then processing fit has value V_1-V_2
 $\text{has_value}(\text{available_effort}, V_1) \ \& \ \text{has_value}(\text{processing_demand}, V_2)$
 $\rightarrow \text{has_value}(\text{processing_fit}, V_1-V_2)$

If alertness fit has value V_1 and processing fit has value V_2 ,
 then functioning fit has value $|V_1|+|V_2|$
 $\text{has_value}(\text{alertness_fit}, V_1) \ \& \ \text{has_value}(\text{processing_fit}, V_2)$
 $\rightarrow \text{has_value}(\text{functioning_fit}, |V_1|+|V_2|)$

If the basic task demand has value V_4 , luminance V_6 , visual demand V_1 , phonological demand V_2 , objects grouping V_8 , objects size V_9 , objects quantity V_{12} and object distinctness value V_3 ,

then processing demand has value

$$\alpha_9 * V_4 + \alpha_{10} * (I - V_6) + \alpha_{17} * V_1 + \alpha_{18} * V_2 + \alpha_{12} * (I - V_8) + \alpha_{13} * (I - V_9) + \alpha_{14} * V_{12} + \alpha_{16} * (I - V_3)$$

has_value(basic_task_demand, V_4) & has_value(luminance, V_6) &
 has_value(visual_demand, V_1) & has_value(phonological_demand, V_2) &
 has_value(objects_grouping, V_8) & has_value(objects_size, V_9) &
 has_value(objects_quantity, V_{12}) & has_value(objects_distinctness, V_3)
 → has_value(processing_demand,

$$\alpha_9 * V_4 + \alpha_{10} * (1 - V_6) + \alpha_{17} * V_1 + \alpha_{18} * V_2 + \alpha_{12} * (1 - V_8) + \alpha_{13} * (1 - V_9) + \alpha_{14} * V_{12} + \alpha_{16} * (1 - V_3)$$
)

If provided effort has value V_{I5} and experienced pressure value V_3 ,
 then available effort has value $\alpha_1 * V_I + \alpha_2 * (I - V_2) + \alpha_{19} * V_3 + \alpha_3 * V_5 + \alpha_4 * V_6 + \alpha_5 * V_{I0}$
 has_value(provided_effort, V_{I5}) & has_value(experienced_pressure, V_3)
 → has_value(available_effort, $\alpha_{20} * V_{I5} + \alpha_{21} * (1 - V_3)$)

If the effort motivation has value V_I , exhaustion V_2 , experienced pressure V_3 , background
 colour V_5 , luminance V_6 , and daytime value V_{I0} ,
 then alertness utilization has value $\alpha_1 * V_I + \alpha_2 * (I - V_2) + \alpha_{19} * V_3 + \alpha_3 * V_5 + \alpha_4 * V_6 + \alpha_5 * V_{I0}$
 has_value(effort_motivation, V_I) & has_value(exhaustion, V_2) &
 has_value(experienced_pressure, V_3) & has_value(background_colour, V_5) &
 has_value(luminance, V_6) & has_value(daytime, V_{I0})
 → has_value(alertness_utilization, $\alpha_1 * V_I + \alpha_2 * (1 - V_2) + \alpha_{19} * V_3 + \alpha_3 * V_5 + \alpha_4 * V_6 + \alpha_5 * V_{I0}$)

If time criticality has value V ,
 then alertness demand has value V
 has_value(time_criticality, V) → has_value(alertness_demand, V)

Appendix B: Detailed Specification of the Analysis Model

If agent A predicts that at T the alertness utilization has value V_I
 and agent A predicts that at T the alertness demand has value V_2
 then agent A will assess that at T the alertness fit has value $V_I - V_2$
 prediction(agentA, has_value_for(alertness_utilization, V_I , T) &
 prediction(agentA, has_value_for(alertness_demand, V_2 , T)
 → assessment(agentA, fit_value_for(alertness, $V_I - V_2$, T))

If agent A predicts that at T the available effort has value V_I
 and agent A predicts that at T the processing demand has value V_2
 then agent A will assess that at T the processing fit has value $V_I - V_2$
 prediction(agentA, has_value_for(available_effort, V_I , T) &
 prediction(agentA, has_value_for(processing_demand, V_2 , T)
 → assessment(agentA, fit_value_for(processing, $V_I - V_2$, T))

If agent A assesses that at T the alertness fit has value V_I
 and agent A assesses that at T the processing fit has value V_2
 then agent A will assess that at T the functioning fit has value $|V_I| - |V_2|$
 assessment(agentA, fit_value_for(alertness_fit, V_I , T)) &
 assessment(agentA, fit_value_for(processing_fit, V_2 , T))
 → assessment(agentA, fit_value_for(functioning, $|V_I| - |V_2|$, T))

If agent A assesses that at T the fit for F has value 0
 then agent A will assess the fit for F at T as perfect
 assessment(agentA, fit_value_for(F , 0 , T))
 → assessment(agentA, fit_for(F , T , perfect))

If agent A assesses that at T the fit for F has value V
 and $0 < V$ and $V \leq 0.1$
 then agent A will assess the fit for F at T as good
 assessment(agentA, fit_value_for(F , V , T)) & $0 < V$ & $V \leq 0.1$
 → assessment(agentA, fit_for(F , T , good))

If agent A assesses that at T the fit for F has value V
 and $-0.1 \leq V$ and $V < 0$
 then agent A will assess the fit for F at T as good
 assessment(agentA, fit_value_for(F, V, T)) & $-0.1 \leq V$ & $V < 0$
 → assessment(agentA, fit_for(F, T, good))

If agent A assesses that at T the fit for F has value V
 and $-1 \leq V$ and $V < -0.1$
 then agent A will assess the fit for F at T as demand dominance
 assessment(agentA, fit_value_for(F, V, T)) & $-1 \leq V$ & $V < -0.1$
 → assessment(agentA, fit_for($F, T, \text{demand_dominance}$))

If agent A assesses that at T the fit for F has value V
 and $0.1 < V$ and $V \leq 1$
 then agent A will assess the fit for F at T as effort dominance
 assessment(agentA, fit_value_for(F, V, T)) & $0.1 < V$ & $V \leq 1$
 → assessment(agentA, fit_for($F, T, \text{effort_dominance}$))

If agent A assesses the fit for F at T as demand dominance
 then agent A will assess the functioning fit at T as poor
 assessment(agentA, fit_for($F, T, \text{demand_dominance}$))
 → assessment_of(agentA, fit_for(functioning, T, poor))

If agent A assesses the fit for F at T as effort dominance
 then agent A will assess the functioning fit at T as poor
 assessment(agentA, fit_for($F, T, \text{effort_dominance}$))
 → assessment_of(agentA, fit_for(functioning, T, poor))

Appendix C: Detailed Specification of the Support Model

If agent A desires that functioning has an adequate fit
 and agent A assesses the functioning fit at T as poor
 and agent A assesses the alertness fit at T as demand dominance
 then agent A will desire an increased alertness fit
 desire(agentA, adequate_functioning_fit) & assessment(agentA, fit_for(functioning, T, poor)) &
 assessment(agentA, fit_for(alertness, $T, \text{demand_dominance}$))
 → desire(agentA, increased(alertness_fit))

If agent A desires that functioning has an adequate fit
 and agent A assesses the functioning fit at T as poor
 and agent A assesses the alertness fit at T as effort dominance
 then agent A will desire a decreased alertness fit
 desire(agentA, adequate_functioning_fit) & assessment(agentA, fit_for(functioning, T, poor)) &
 assessment(agentA, fit_for(alertness_fit, $T, \text{effort_dominance}$))
 → desire(agentA decreased(alertness_fit))

If agent A desires that functioning has an adequate fit
 and agent A assesses the functioning fit at T as poor
 and agent A assesses the processing fit at T as demand dominance
 then agent A will desire an increased processing fit
 desire(agentA, adequate_functioning_fit) & assessment(agentA, fit_for(functioning, T, poor)) &
 assessment(agentA, fit_for(processing, $T, \text{demand_dominance}$))
 → desire(agentA, increased(processing_fit))

If agent A desires that functioning has an adequate fit
 and agent A assesses the functioning fit at T as poor
 and agent A assesses the processing fit at T as effort dominance

```

then agent A will desire an decreased processing fit
  desire(agentA, adequate_functioning_fit) & assessment(agentA, fit_for(functioning, T, poor)) &
  assessment(agentA, fit_for(processing, T, effort_dominance))
  → desire(agentA, decreased(processing_fit))
If agent A desires an increased alertness fit
then agent A will desire an increased alertness utilization
  desires(agentA, increased(alertness_fit) ) → desires(agentA, increased(alertness_utilization) )
If agent A desires a decreased alertness fit
then agent A will desire a decreased alertness utilization
  desires(agentA, decreased(alertness_fit) ) → desires(agentA, decreased(alertness_utilization))
If agent A desires an increased processing fit
then agent A will desire a decreased processing demand
  desires(agentA, increased(processing_fit) )
  → desires(agentA, decreased(processing_demand) )
If agent A desires a decreased processing fit
then agent A will desire an increased processing demand
  desires(agentA, decreased(processing_fit) )
  → desires(agentA, increased(processing_demand) )

```

Consumer Decision Making in Knowledge-Based Recommendation

Monika Mandl, Alexander Felfernig, and Monika Schubert

Applied Software Engineering, Graz University of Technology
Inffeldgasse 16b, A-8010 Graz, Austria

{monika.mandl,alexander.felfernig,monika.schubert}@ist.tugraz.at

Abstract. In contrast to customers of bricks and mortar stores, users of online selling environments are not supported by human sales experts. In such situations recommender applications help to identify the products and/or services that fit the user's wishes and needs. In order to successfully apply recommendation technologies we have to develop an in-depth understanding of decision strategies of users. These decision strategies are explained in different models of human decision making. In this paper we provide an overview of selected models and discuss their importance for recommender system development. Furthermore, we provide an outlook on future research issues.

Keywords: Knowledge-based Recommendation, Interactive Selling, Consumer Buying Behavior, Consumer Decision Making.

1 Introduction

Traditional approaches to recommendation (collaborative filtering [1], content-based filtering [2], and different hybrid variants thereof) are well applicable for recommending quality & taste products such as movies, groceries, music, or news. Especially in the context of high-involvement products such as computers, cars, apartments, or financial services, those approaches are less applicable. For example, cars are not bought very frequently – consequently the corresponding items will not receive a critical mass of ratings needed for making reasonable predictions; [3] propose to use the 100 nearest neighbors in their collaborative filtering recommendation approach. Furthermore, a low frequency of ratings would require to take into consideration a rather long time period of ratings – this would make it infeasible for a content-based filtering algorithm to derive meaningful predictions.

Especially in domains where traditional recommendation approaches are not the first choice, knowledge-based recommendation technologies come into play [4,5]. Knowledge-based recommender applications are exploiting explicitly defined requirements of the user and additionally dispose of deep knowledge about the underlying product assortment. Thus, knowledge-based recommender applications exploit knowledge sources that are typically not available in collaborative and content-based filtering scenarios. A direct consequence of the availability of deep knowledge about the product assortment and explicitly defined customer requirements is that no ramp-up problems occur with knowledge-based recommenders [4,5]. The other side of the

coin is that – due to the explicit representation of recommendation knowledge in a recommender knowledge base – knowledge-based recommenders cause so-called knowledge acquisition bottlenecks: knowledge engineers and domain experts have to invest considerable time efforts in order to develop and keep those knowledge bases up-to-date.

In this paper we focus on the discussion of selected models of consumer decision making and their importance for the development of knowledge-based recommender applications. The remainder of this paper is organized as follows. In Section 2 we introduce the basic functionalities supported by knowledge-based recommender applications. We provide an overview of a selected set of models of consumer decision making in Section 3. Section 4 includes a discussion of different types of decoy effects that can have a major impact on the item selection behavior of users. In Section 5 we discuss further theories of consumer decision making. With Section 6 we provide an outlook of future research. The paper is concluded with Section 7.

2 Knowledge-Based Recommendation

The major difference between filtering-based recommendation approaches and knowledge-based recommendation [4,5] is that explicit knowledge about customers, the product assortment, and the dependencies between customer preferences and product properties is stored in a corresponding recommender knowledge base. The rules for the identification of a solution are explicitly defined and thus allow the derivation of intelligent and deep explanations as to why certain products have been recommended to a customer. Since advisory knowledge is represented in the form of variables and constraints we are able to automatically determine diagnoses and repair actions in situations where no solution can be found for the given set of customer requirements [6]. Knowledge-based recommendation problems can be defined on the basis of simple conjunctive database queries as well as on the basis of so-called constraint satisfaction problems (CSPs) [7]. Figure 1 presents an example of a knowledge-based recommender application that has been developed for one of the largest financial service providers in Austria. Such an application guides a user (repeatedly) through the following phases:

1. *Requirements specification* (Phase I.): in the first phase users are interacting with the recommender application in order to identify and specify their requirements. Examples for such requirements in the financial services domain are *the investment period should be below four years*, *the profit per year should be more than 5 percent*, or *the recommended items should not contain shares*.
2. *Repair of inconsistent requirements* (Phase II.): in the case that the recommender application is not able to identify a solution, it proposes a set of repair actions (change proposals for requirements) that (if accepted by the user) can guarantee the identification of a recommendation. An example for such an infeasibility in the financial services domain is a *low willingness to take risks* combined with *high return rates* of the investment. Another example for an infeasibility is the combination of *high return rates* and *short investment periods*.
3. *Result presentation* (Phase III.): if the defined requirements can be fulfilled, the recommender application presents a set of product alternatives. Those alternatives

are typically ranked on the basis of a utility function (for a detailed example see [5]) and are either presented in the form of an ordered list or on a product comparison page.

4. *Explanations* (Phase IV.): For each of the identified and presented product alternatives the customer can activate a corresponding explanation as to why this product has been recommended. Each explanation consists of a set of argumentations that relate specified requirements with the corresponding product properties. An example for an explanation is *we recommend this product since it supports the specified investment period and additionally provides reasonable return rates with low risks.*

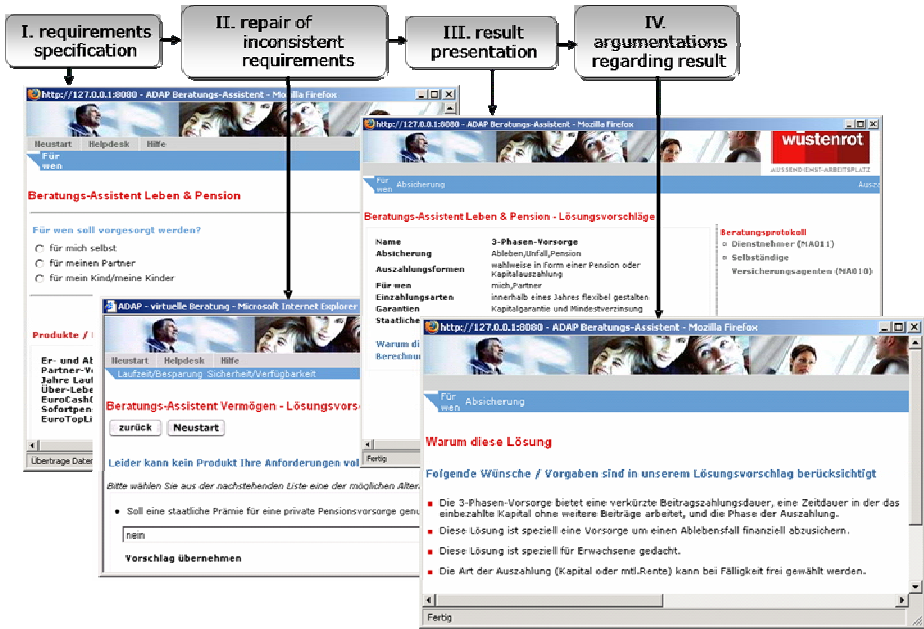


Fig. 1. Example knowledge-based recommender application. A typical recommendation process consists of the phases requirements specification (I.), repair of inconsistent requirements (II.), presentation of recommended products (III.), and explanation of each selected product (IV.).

3 Models of Consumer Decision Making

Traditional Economic Models. Those models are assuming that all users are able to take decisions that are optimal and that have been derived on the basis of rational and formal processes. An assumption of economic models in this context is that preferences remain stable, i.e., are not adapted within the scope of a decision process. However, it is a fact that preferences can be extremely unstable, for example, a customer who buys a car first sets the upper limit for the overall price to 20.000€. This does not mean that the upper limit is strict since the customer could change his mind and set the upper limit for the price to 25.000€ simply because he detected additional technical features for which he is willing to pay the higher price, for example, high-quality

headlights, park-distance control, satellite navigation, and rain-sensor for the wind-screen wipers. Solely on the basis of this simple example we immediately see that preferences could change over time, i.e., are not stable within the scope of a recommendation session. This insight led to the development of new decision models [8] that take into account this constructive nature of preferences. The most important ones will be discussed in the following.

Effort Accuracy Framework. Following this model, users are taking into account cost-benefit aspects. A decision process is now characterized by a trade-off between the effort to take a decision and the expected quality of the decision. The effort-accuracy framework is based on the fact that users (customers) show an adaptive decision behavior and select from an available set of different decision heuristics depending on the current situation. Criteria for the selection of a certain heuristic are on the one hand the needed decision quality and on the other hand the cognitive efforts needed for successfully completing the decision task. Per definition this framework clearly differs from the above mentioned economic model of decision making. In those models, optimality plays a dominant role and the efforts related to successfully completing a decision task are neglected. However, especially the effort for completing a decision task has to be taken into account as an important factor that determines whether the user is willing to apply the recommender application or chooses a different provider.

Construction of Preferences. The concept of preference construction in human choice has been developed by [9]. The basic idea of preference construction is that users tend to identify their preferences within the scope of a recommendation session but only in rare cases are able to state their preferences before the beginning of the decision process. Thus decision processes are more focused on *constructing* a consistent set of preferences than *eliciting* preferences from the user which is still the predominantly supported type of decision process in many existing knowledge-based recommender applications. Since user preferences are constructed within the scope of a recommendation session, the design of the user interface can have a major impact on the final outcome of the decision process.

Table 1. Selected theories of decision psychology

theory	explanation
decoy effects	inferior products added to a result set can significantly change the outcome of the decision process.
primacy/recency	information units at the beginning of a list and at the end of a list are analyzed and remembered significantly more often than information units in the middle of a list.
framing	the way in which we describe a certain decision alternative can have a significant impact on the final decision.
defaults	pre-selected decision alternatives have the potential to significantly change the outcome of a decision process.

In order to improve the applicability of recommender applications we must integrate recommendation technologies with deep knowledge about human decision making. Such an integration can help to improve the perceived quality of the recommender application for the user as well as the predictability of decision outcomes. In the remainder of this paper we review selected theories from decision psychology w.r.t. their potential impact on preference construction processes. An overview of those theories is provided in Table 1 – they will be discussed in the following subsections.

4 Decoy Effects




Decoy products are items that are inferior to other items in a given set of recommended products.¹ In this context, the inferiority respectively superiority of items is measured by simply comparing the underlying properties of items with regard to their distance to the optimal value, for example, *robot X* dominates *robot Y* in the dimensions *price* and *reliability* if it has both a lower price and a higher reliability. The inclusion of such decoy products can significantly influence the outcome of the decision process and therefore has to be taken into account when implementing recommender applications. The phenomenon that users change their selection behavior in the presence of additional inferior items is denoted as *decoy effect*. Decoy effects have been intensively investigated in different application contexts, see, for example [10,11,12].

In the following subsections we will discuss different types of decoy effects and explain how those effects can influence the outcome of decision processes. Note that the existence of decoy effects provides strong evidence against the validity of traditional economic models of choice that suppose rational and optimal strategies in human decision making.

4.1 Compromise Effects

Compromise effects denote one specific archetype of decoy effects which is shown in Table 2. It is possible to increase the attractiveness of robot X compared to robot Y by adding robot D to the set of alternatives. Robot D increases the attractiveness of robot X since, compared to robot D, X has a significantly lower price and only a marginally lower reliability (this effect is denoted as *tradeoff-contrast*). This way, X is established as a compromise between the alternatives Y and D. By the insertion of decoy robot D the comparison focus of the user is set to XD since D is more similar to X than to Y (*similarity effect*). Note that the compromise of choosing X can as well be explained by the aspect of *extremeness aversion*.

Table 2. Compromise effect

product (robot)	X 	Y 	D 
price [0..10.000€]	3.000	1.500	5.000
reliability [0..10]	9	4.5	10




¹ Note that we use the *robot product domain* in the following examples.

More formally, we can explain decoy effects as follows. Under the assumption that the probability of selection for item X out of the item set {X,Y} is equal to the probability of selection of Y out of {X,Y}, i.e., $P(X,\{X,Y\}) = P(Y,\{X,Y\})$, the addition of D causes a preference shift to X, i.e., $P(Y,\{X,Y,D\}) < P(X,\{X,Y,D\})$.

4.2 Asymmetric Dominance Effects

The second archetype of decoy effect is called asymmetric dominance (depicted in Table 3). In this scenario, robot X dominates robot D in both attributes (price and reliability) whereas robot Y dominates robot D in only one dimension (the price). The addition of robot D to the set of {X,Y} can help to increase the share of X. In this context the comparison focus is set to XD (D is more *similar* to X than Y) which makes X the clear winner in the competition, i.e., $P(Y,\{X,Y,D\}) < P(X,\{X,Y,D\})$.




Table 3. Asymmetric dominance effect

product (robot)	X 	Y 	D 
price [0..10.000€]	3.000	1.000	3.500
reliability [0..10]	9	5	8

4.3 Attraction Effects

The third archetype of decoy effects is called attraction effect. In this context, X appears to be only a little bit more expensive and simultaneously has a significantly higher reliability compared to robot D (*tradeoff-contrast* – see Table 4). In this scenario the inclusion of D can trigger an increased probability of selection for robot X since X appears to be more attractive than D, i.e., $P(Y,\{X,Y,D\}) < P(X,\{X,Y,D\})$. The attraction effect moves the comparison focus to the combination of items XD since D is more similar to X than to Y (*similarity effect*). Note that both compromise effects and attraction effects are based on the ideas of tradeoff-contrast and similarity. The difference lies in the positioning of decoy items. In the case of the compromise effect decoy products are representing extreme solutions (see Table 2) whereas in the case of the attraction effect decoy products are positioned between the target and the competitor product (see Table 4).

Table 4. Attraction effect

product (robot)	X 	Y 	D 
price [0..10.000€]	5.000	2.000	4.900
reliability [0..10]	7	3	5

4.4 Application of Decoy Effects in Recommendation Scenarios

If decoy items are added to a result set, this can change the selection probability for items that were included in the original result set. Decoy effects have been shown in a

number of empirical studies in application domains such as financial services, e-tourism, and even software agents (see, for example, [10,11,14]). The major possibilities of exploiting decoy effects in knowledge-based recommendation scenarios are the following:

- *Increased selection probability for target items:* as already mentioned, adding additional inferior items to a result set can cause an increased share of target items (in our example denoted as item X). This scenario definitely has ethical aspects to be dealt with since companies can potentially try to apply decoy effects for selling products that are maybe suboptimal for the customer.
- *Increased decision confidence:* beside an increase of the share of the target product, decoy effects can be exploited for increasing the decision confidence of a user. In this context, decoy effects can be exploited for resolving cognitive dilemmas which occur when a user is unsure about which alternative to choose from a given set of nearly equivalent alternatives.
- *Increased willingness to buy:* from empirical studies we know that a user's level of trust (confidence) in recommendations is directly correlated with the willingness to buy, i.e., increasing the level of trust directly means that the purchase probability can be increased as well [13].

The important question to be answered now is how to predict decoy effects within the scope of a recommendation scenario. Predicting the selection of products contained in the set of possible product alternatives (the consideration set CSet) requires the calculation of dominance relationships between the items contained in a result set. Exactly for this calculation different models have been developed [14,15] – the outcome of each of these models are dominance relationships between the items in CSet. The calculation of such dominance relationships can be based on Formula 1 which is a simplified version of the approach introduced in [14]. This Formula allows the calculation of dominance relationships between different products in a consideration set, i.e., $d(u, CSet)$ denotes the dominance of product u compared to all other items in CSet.

$$d(u, CSet) = \sum_{v \in CSet - \{u\}} \sum_{a \in properties} \sqrt{\frac{diff(u_a, v_a)}{a_{max} - a_{min}}} * sign(u_a, v_a) .$$

Formula 1. Calculating dominance value d for u in $CSet$: $diff(u_a, v_a) = u_a - v_a$ if $a=reliability$, otherwise $diff(u_a, v_a) = v_a - u_a$, $sign(u_a, v_a)=1$ if $u_a \geq v_a$, -1 otherwise.

Applying Formula 1 to the product set $\{X, Y, D\}$ depicted in Table 2 results in the dominance values that are depicted in Table 5. For example, product v_1 (Y) has a better price than product u (X; the target item) – the corresponding dominance value is -0.65, i.e., product u is inferior regarding the attribute price. The sum of the attribute-wise calculated dominance values, i.e., $d(u, CSet)$, provides an estimation of how dominant item u appears to be in the set of candidate items CSet. The values in Table 3 clearly show a dominance of item X over the items Y and D.

Table 5. Dominance values for $A \in \text{CSet}$ for Table 2

	u	v₁	v₂	Sum	d(u,CSet)
	X	Y	D		d(X,{X,Y,D})
price		-0.65	0.75	0.10	
reliability		0.90	-0.42	0.48	
					0.58
	Y	X	D		
price		0.65	1.0	1.65	
reliability		-0.90	-1.0	-1.9	
					-0,25
	D	X	Y		
price		-0.75	-1.0	-1.75	
reliability		0.42	1.0	1.42	
					-0.33

The dominance relationships between items in a result set can be directly used by a corresponding configuration algorithm [14]. If the recommendation algorithm determines, for example, 20 possible products (the consideration set) and the company wants to increase the sales of specific items in this set, a configuration process can determine the optimal subset of items that should be presented to the user such that purchase probability is maximized.

5 Further Effects

5.1 Primacy/Recency

Primacy/recency effects occur in situations where products are presented in the form of a list. Products presented at the beginning and at the end of the list are evaluated more often than those positioned somewhere in the middle of the list. Significant changes in the product selection behavior that have been triggered by changed product orderings are discussed in [16]. Similar experiences are reported in the context of web search where web links at the beginning and the end of a list are activated significantly more often than those in the middle of the list [17]. Users do not want to evaluate large lists of decision alternatives but rather find those alternatives that fit their wishes and needs as soon as possible. As a consequence, recommender applications must be able to calculate item rankings that reduce the cognitive overheads of users as much as possible. An approach to take into account primacy/recency effects in the presentation of result sets has been introduced in [18].

5.2 Framing

Framing effects occur when one and the same decision alternative is presented in different variants [19]. An example is *price framing* where – depending on the granularity

of the presented price information (price information presented in one attribute or distributed over several attributes) – users have different heuristics to evaluate the remaining product attributes [20]. If the price information is provided for different subparts of a product, users tend to focus on evaluating those subparts with a corresponding price information. If the product price on the contrary is represented by one attribute, users focus on evaluating those parts that are important (e.g., the resolution or zoom factor of a digital camera). Another occurrence of framing is *attribute framing* that describes the phenomenon that different but equivalent descriptions of a decision task can trigger completely different outcomes. For example, a fund with a 95% probability of no loss is interpreted as a better solution compared to the same product described with a 5% probability of loss.

5.3 Defaults

Defaults can be used to support users in the specification of their requirements, especially if users are non-experts in the product domain and therefore are not sure about which alternative to select [21]. For example, if the user is interested in *high return rates*, the runtime of the financial service should be *longterm* (the default). Thus defaults are a means to help the user to identify meaningful alternatives that are compatible with their current preferences. Obviously, defaults could as well be abused for manipulation purposes, i.e., to mislead users to purchase items that are unnecessary for their requirements [22]. Especially for knowledge-based recommender applications defaults play a very important role since users tend to accept preset values (defaults are representing the status quo) compared to other alternatives [23,24]. Maintaining the status quo and refusing to change the status quo is denoted as the status-quo bias in the literature [24]. This can be explained by the fact that users typically see new alternatives in the light of potential losses. Since users are typically loss-averse, they are very often reluctant to accept changes to the status quo. Typical attributes for which default values are specified are those with an associated risk, for example, warranties of a product or safety equipment.

6 Future Work

In the previous sections we focused on the discussion of selected decision-psychological aspects relevant for the development of knowledge-based recommender applications. In the remainder of this paper we are discussing relevant research topics especially related to decoy effects.

Decoy effects in repair actions. Repair actions help users to get out of the so-called “no solution could be found” dilemma (see Section 2). If a given set of requirements does not allow the calculation of a recommendation there exist potentially many different alternative combinations of repair actions (exponential in the number of requirements [25]) that resolve the current conflict. As a consequence, it is not possible to present the complete set of possible repair actions and we have to select a subset that best fits with the requirements of the user. An approach to personalize the selection of repair actions has been introduced in [26]. Our major goal for future work is to extend the approach of [26] by additionally taking into account different types of decoy effects that potentially occur in the repair selection process.

Decoy effects on result pages. Similar to the selection of repair alternatives we are also interested in general properties of decoy effects in the context of product lists. In this context we are interested in answering questions regarding the upper bound for the number of products such that decoy effects still occur. Furthermore, we are interested in the relevance of item distances for the existence of decoy effects. The question is whether we have to organize target, competitor, and decoy items in a cluster or do decoy effects still occur if the distance between those items is increased.

Decoy effects in compound critiques. Critiquing-based recommender applications [27,28] often support the concept of compound critiques. Critiques are a natural way to support users in item selection processes without forcing them to explicitly specify values for certain item properties (query-based approach). Users are more engaged in a navigation process where they are articulating requirements on a more abstract level such as *lower price* or *higher resolution*. In order to fasten the interaction with a critique-based recommender application, prototype systems have been developed that support the articulation of so-called compound critiques, i.e., critiques that include two or more change requests regarding basic product properties. An example for such a compound critique is *lower price and higher resolution*. A typical critiquing-based recommender presents a list of alternative compound critiques to the user. In this context, we are interested in answering the question whether decoy effects occur in the selection of compound critiques.

7 Conclusions

We have presented a selected set of decision-psychological phenomena that have a major impact on the development of recommender applications. A number of related empirical studies clearly show the importance of taking into account such theories when implementing a knowledge-based recommender application. We see our contribution as a first one on the way towards more intelligent recommender user interfaces that know more about the user and also know how to exploit this knowledge for improving the quality of applications in different dimensions such as prediction accuracy or satisfaction with the presented recommendations.

Acknowledgements. The work presented in this paper has been developed within the scope of the research projects XPLAIN-IT (funded by the Privatstiftung Kärnter Sparkasse) and Softnet Austria that is funded by the Austrian Federal Ministry of Economics (bm:wa), the province of Styria, the Steirische Wirtschaftsförderungsgesellschaft mbH (SFG), and the city of Vienna in terms of the center for innovation and technology (ZIT).

References

1. Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., Riedl, J.: GroupLens: applying collaborative filtering to Usenet news Full text. *Communications of the ACM* 40(3), 77–87 (1997)
2. Pazzani, M., Billsus, D.: Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning* 27, 313–331 (1997)

3. Bell, R., Koren, Y.: Improved Neighborhood-based Collaborative Filtering. In: Proceedings of the 1st KDDCup 2007, San Jose, California (2007)
4. Burke, R.: Knowledge-based Recommender Systems. *Encyclopedia of Library and Information Systems* 69(32), 180–200 (2000)
5. Felfernig, A., Burke, R.: Constraint-based Recommender Systems: Technologies and Research Issues. In: ACM International Conference on Electronic Commerce (ICEC 2008), Innsbruck, Austria, August 19–22 (2008)
6. Felfernig, A., Friedrich, G., Jannach, D., Stumptner, M.: Consistency-based Diagnosis of Configuration Knowledge Bases. *Artificial Intelligence* 2(152), 213–234 (2004)
7. Tsang, E.: *Foundations of Constraint Satisfaction*. Academic Press, London (1993)
8. Payne, J., Bettman, J., Johnson, E.: *The adaptive decision maker*. Cambridge University Press, Cambridge (1993)
9. Bettman, J., Luce, M., Payne, J.: Constructive Consumer Choice Processes. *Journal of Consumer Research* 25(3), 187–217 (1998)
10. Huber, J., Payne, W., Puto, C.: Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research* 9, 90–98 (1982)
11. Simonson, I., Tversky, A.: Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research* 29, 281–295 (1992)
12. Yoon, S., Simonson, I.: Choice set configuration as a determinant of preference attribution and strength. *Journal of Consumer Research* 35, 324–336 (2008)
13. Chen, L., Pu, P.: Trust Building in Recommender Agents. In: 1st International Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces (WPRSUI 2005), Reading, UK, pp. 135–145 (2005)
14. Teppan, E., Felfernig, A.: The asymmetric dominance effect and its role in e-tourism recommender applications. In: Proceedings of Wirtschaftsinformatik 2009, Vienna, Austria, pp. 791–800 (2009)
15. Roe, R., Busemeyer, J., Townsend, T.: Multialternative decision field theory: a dynamic connectionist model of decision making. *Psychological Review* 1, 7–59 (2001)
16. Felfernig, A., Friedrich, G., Gula, B., Hitz, M., Kruggel, T., Leitner, G., Melcher, R., Riepan, D., Strauss, S., Teppan, E., Vitouch, O.: Persuasive Recommendation: Exploring Serial Position Effects in Knowledge-based Recommender Systems. In: de Kort, Y.A.W., IJsselsteijn, W.A., Midden, C., Eggen, B., Fogg, B.J. (eds.) *PERSUASIVE 2007*. LNCS, vol. 4744, pp. 283–294. Springer, Heidelberg (2007)
17. Murphy, J., Hofacker, C., Mizerski, R.: Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication* 11(2), 522–535 (2006)
18. Felfernig, A., Gula, B., Leitner, G., Maier, M., Melcher, R., Teppan, E.: Persuasion in Knowledge-based Recommendation. In: Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerst ahl, K., Øhrstr om, P. (eds.) *PERSUASIVE 2008*. LNCS, vol. 5033, pp. 71–82. Springer, Heidelberg (2008)
19. Tversky, A., Kahnemann, D.: Rational choice and the framing of decisions. *Journal of Business* 59(4), 251–278 (1986)
20. Levin, I., Schneider, S., Gaeth, G.: All Frames are not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Processes* 76, 90–98 (1998)
21. Huffman, C., Kahn, B.: Variety for sale: Mass Customization or Mass Confusion. *Journal of Retailing* 74(4), 491–513 (1998)
22. Herrmann, A., Heitmann, M., Polak, B.: The Power of Defaults. *Absatzwirtschaft* 6, 46–47 (2007)

23. Ritov, I., Baron, J.: Status-quo and omission biases. *Journal of Risk and Uncertainty* 5(2), 49–61 (1992)
24. Samuelson, W., Zeckhauser, R.: Status quo bias in decision making. *Journal of Risk and Uncertainty* 108(2), 370–392 (1988)
25. O’Sullivan, B., Papadopoulos, A., Faltings, B., Pu, P.: Representative Explanations for Over-Constrained Problems. In: *AAAI 2007*, pp. s323–328 (2007)
26. Felfernig, A., Friedrich, G., Schubert, M., Mandl, M., Mairitsch, M., Teppan, E.: Plausible Repairs for Inconsistent Requirements. In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, Pasadena, California, USA (2009) (to appear)
27. Reilly, J., Zhang, J., McGinty, L., Pu, P., Smyth, B.: A comparison of two compound critiquing systems. In: *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI 2007)*, Honolulu, Hawaii, pp. 3127–3320 (2007)
28. Chen, L., Pu, P.: The Evaluation of a Hybrid Critiquing System with Preference-based Recommendations Organization. In: *Proceedings of the ACM Conference on Recommender Systems*, Minneapolis, Minnesota, USA, pp. 169–172 (2007)

Incremental Learning of Triadic PLSA for Collaborative Filtering

Hu Wu and Yongji Wang

Institute of Software, Chinese Academy of Sciences,
Beijing, China
{wuhu, ywang}@itechs.iscas.ac.cn

Abstract. PLSA which was originally introduced in text analysis area, has been extended to predict user ratings in the collaborative filtering context, known as Triadic PLSA (TPLSA). It is a promising recommender technique but the computational cost is a bottleneck for huge data set. We design an incremental learning scheme for TPLSA for collaborative filtering task that could make forced prediction and free prediction as well. Our incremental implementation is the first of its kind in the probabilistic model based collaborative filtering area, to our best knowledge. Its effectiveness is validated by experiments designed for both rating-based and ranking-based collaborative filtering.

1 Introduction

Probabilistic Latent Semantic Analysis (PLSA) is an effective unsupervised learning model for a wide range of applications such as text analysis [1], image analysis [2] and collaborative filtering (CF) [7]. It is one of the state-of-art model-based collaborative filtering algorithms and has many applications in real-world recommender systems [10]. However, it is necessary to distinguish collaborative filtering from other applications for PLSA because the user rating value can not be mapped directly to the co-occurrence value which is the case in text or image analysis. We need to model the relationship between user and item in a triplet $\langle user, item, rating \rangle$ where rating value indicates whether the relationship is positive or negative and how strong it is. [7] extended the classical PLSA [1] to provide an effective while computational intensive method for collaborative filtering, sometimes referred to as Triadic PLSA.

Due to the tremendous growth and changing of users and products, incremental learning ability is of growing importance for real-world recommender systems, especially for PLSA algorithm [11]. Triadic PLSA training is even more computational consumptive comparing with classical PLSA and other memory-based recommender methods. In this paper, we take one step further on the basis of TPLSA model [7]. By making its learning process incremental, the proposed model is much easier to learn while only with a little prediction precision. We believe our work is valuable in itself as well as in its implications for other potential incremental extensions on model based collaborative filtering methods.

The remainder of the paper is structured as follows: Section 2 reviews the prior works in PLSA model for collaborative filtering together with some incremental improvements for dyadic PLSA model; our incremental speedup algorithm for triadic version

of PLSA model is given in Section 3 then we perform a thorough comparison with some state-of-art algorithms in Section 4. Finally, Section 5 concludes the paper and points the possible future directions.

2 Related Works

Before we continue, some notations used in this paper are listed below.

Table 1. Notations

Notations	Meanings	Notations	Meanings
u^l	user	d	document
y	item	w	word
v	rating	z	latent class
K	number of classes	μ	average rating
σ	variance of rating	N	total rating triplets number
n_{iter}	EM iteration number	$N_{training}, N_{testing}$	training and testing rating triplets number

2.1 PLSA for Collaborative Filtering

We start from a brief introduction of collaborative filtering problem setting. Users choose to rate a subset of items in the system, so the training data are depicted by many triplets $\langle u, y, v \rangle$, indicating the user preference of an item. The goal of collaborative filtering is boiled down to predicting the most likely rating value of a given user (known as the active user) on an item. There are two different but related prediction scenarios been called forced prediction and free prediction. In the forced prediction setting, for the active user and all the items, an estimated rating value is predicted, the learning result is therefore a mapping: $\mathcal{G} : U \times Y \rightarrow V$; on the other hand, the goal of the free prediction task is to choose some items that the user might be interested in and to predict the possible ratings of these items, i.e., the learning result is a function: $\mathcal{F} : U \rightarrow Y \times V$.

PLSA is believed to be one of the state-of-art collaborative filtering algorithms. It is previously introduced in the text analysis context where the learning data are co-occurrence of the documents and words. PLSA tries to construct the latent semantic topic behind the words for each document in a probabilistic fashion. Noticing the similarity between co-occurrence matrix and user rating matrix, [7] extended the classical PLSA model which was designed to learn latent label from dyadic data (i.e., $\langle d, w \rangle$ concurrence data) to triadic data $\langle u, y, v \rangle$. Specifically, Hofmann proposed two graphic models for the forced prediction and free prediction, respectively (Fig. 1 (b) and (c)). The two models and their corresponding Expectation Maximization (EM) learning algorithms are given formally below:

Forced prediction case

I. Model definition:

$$P(v|u, y) = \sum_z P(z|u)P(v|y, z), \quad (1)$$

¹ We use the lower case for a single user or item and upper case for corresponding set of these objects in the system.

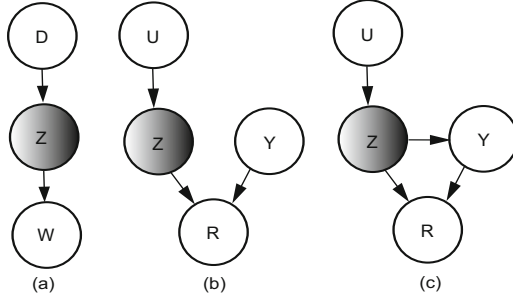


Fig. 1. The graphic model of PLSA for collaborative filtering: (a) is the classical d-w model, (b) and (c) are forced prediction and free prediction extensions of PLSA on collaborative filtering, respectively

– The **multinomial** case:

$$P(v|y, z) = \frac{\sum_{\langle u, v', y' \rangle : v' = v, y' = y} P(z|u, v, y, \hat{\theta})}{\sum_{\langle u, v', y' \rangle : y' = y} P(z|u, v, y, \hat{\theta})}, \quad (2)$$

– The **Gaussian** case:

$$P(v|y, z) = P(v|\mu_{y,z}, \sigma_{y,z}) = \frac{1}{\sqrt{2\pi}\sigma_{y,z}} \exp\left[-\frac{(v - \mu_{y,z})^2}{2\sigma_{y,z}^2}\right]. \quad (3)$$

II. Model fitting: The EM algorithm [7] consists of two steps that are performed alternatively:

– **E-Step:**

$$P(z = k|u, v, y, \hat{\theta}) = \frac{\hat{P}(z = k|u)\hat{P}(v|y, z)}{\sum_{z'} \hat{P}(z'|u)\hat{P}(v|y, z')}, \quad (4)$$

– **M-Step:**

$$P(z|u) = \frac{\sum_y P(z|u, v, y, \hat{\theta})}{\sum_z \sum_y P(z|u, v, y, \hat{\theta})}. \quad (5)$$

• The **multinomial** case:

$$P(v|y_{\bar{u}}, z) = \frac{\sum_{\langle u, v, y \rangle : y = y_{\bar{u}}} P(z = k|u, v, y, \hat{\theta})}{\sum_{\langle u, v', y \rangle : y = y_{\bar{u}}} P(z = k|u, v', y, \hat{\theta})}, \quad (6)$$

• The **Gaussian** case:

$$\mu_{y_{\bar{u}}, z} = \frac{\sum_{\langle u, v, y \rangle : y = y_{\bar{u}}} v P(z|u, v, y, \hat{\theta})}{\sum_{\langle u, v, y \rangle : y = y_{\bar{u}}} P(z|u, v, y, \hat{\theta})}, \quad (7)$$

$$\sigma_{y_{\bar{u}}, z}^2 = \frac{\sum_{\langle u, v, y \rangle : y = y_{\bar{u}}} (v - \mu_{y,z})^2 P(z|u, v, y, \hat{\theta})}{\sum_{\langle u, v, y \rangle : y = y_{\bar{u}}} P(z|u, v, y, \hat{\theta})}. \quad (8)$$

III. Recommendation:

$$\mathbf{E}[v|u, y] = \int_v v P(v|u, y) dv = \begin{cases} \sum_z P(z|u) \int_v v P(v|y, z) dv & (\text{multinomial case}), \\ \sum_z P(z|u) \mu_{y,z} & (\text{Gaussian case}). \end{cases} \quad (9)$$

Free prediction case

I. Model definition:

$$P(v, y|u) = \sum_z P(z|u)P(y|z)P(v|y, z), \quad (10)$$

II. Model fitting:

– **E-Step**:

$$P(z = k|u, v, y, \hat{\theta}) = \frac{\hat{P}(z|u)\hat{P}(v|y, z)\hat{P}(y|z)}{\sum_{z'} \hat{P}(z'|u)\hat{P}(v|y, z')\hat{P}(y|z')}, \quad (11)$$

– **M-Step**²:

$$P(y|z) = \frac{\sum_{\langle u, y', v, z' \rangle : y' = y, z' = z} P(z'|u, y', v, \hat{\theta})}{\sum_{\langle u, y, v, z' \rangle : z' = z} P(z'|u, y, v, \hat{\theta})}. \quad (12)$$

2.2 Incremental Learning of PLSA

Typical EM treatment for PLSA is processed in the batched fashion. When new data arrive, the whole model needs to be retrained with new data and old data together. This is obviously problematic. The focus of this paper is to deal with incremental learning of the Traidic PLSA.

There are three different scenarios for incremental learning should be considered.

1. New users register to the system and rate a few items, in this situation, the $P(z|u)$ corresponding to these users should be calculated and recorded;
2. New items are added to the system and rated by a few users, $P(v|y, z)$ should be calculated for the new items and updated for the old items with respect to all the latent classes;
3. An existing user rates an existing item, both $P(z|u)$ and $P(v|y, z)$ should be updated for all the users and items.

We note that much work has been done on the incremental implementation of classical diadic PLSA model. [4] provided a clustering based method for collaborative filtering. [6] presented a complete Bayes solution for incremental PLSA learning.

The advantage of the algorithm proposed by [6] is that they consider the first two scenarios aforementioned, i.e., they can handle the situation that new users evaluate new items or old items. The updating takes 4 steps as follows:

1. Discard old documents and terms $P(w|z) = \frac{P_0(w|z)}{\sum_{w' \in w_0} P_0(w'|z)}$, $P(d|z) = \frac{P_0(d|z)}{\sum_{d' \in D_0} P_0(d'|z)}$.
2. Fold in new documents:

$$P(z|w, d_{new}) = \frac{P(w|z)P(z|d_{new})}{\sum_{z' \in Z} P(w|z')P(z'|d_{new})}, P(z|d_{new}) = \frac{\sum_{w \in d_{new}} f(w, d_{new})P(z|w, d_{new})}{\sum_{z' \in Z} \sum_{w \in d_{new}} f(w, d_{new})P(z'|w, d_{new})}. \quad (13)$$

3. Fold in new words:

$$P(z|w_{new}, d_{new}) = \frac{P(d_{new}|z)P(z|w_{new})}{\sum_{z' \in Z} P(d_{new}|z')P(z'|w_{new})}, \quad (14)$$

² $P(z|u)$ and $P(v|y, z)$ is calculated following Equ.5 and 6, respectively.

$$P(z|w_{new}) = \frac{\sum_{d \in D_{new}} f(w_{new}, d)P(z|w_{new}, d)}{\sum_{d' \in D_{new}} f(w_{new}, d')}. \quad (15)$$

4. Update PLSA parameters, all $P(w|z)$ are normalized using the following:

$$P(w|z) = \frac{\sum_{d \in D \cup D_{new}} f(w, d)P(z|w, d)}{\sum_{d' \in D \cup D_{new}} \sum_{w' \in d'} f(w', d')P(z|w', d')}. \quad (16)$$

The complexity analysis: the algorithm needs $O(n_{iter} \times (n_{nd} + n_{od}) \times (n_{nw} + n_{ow}) \times k)$ operations to converge whenever there are new documents added, where n_{nd} is the number of new documents, n_{od} is the number of old documents, n_{nw} is the number of new words and n_{ow} is the number of old words.

[9] presented another method for incremental learning PLSA parameters.

$$\hat{P}_{MAP}(w_j|z_k) = \frac{\sum_{i=1}^N n(d_i, w_j)P(z_k|d_i, w_j) + (\alpha_{j,k} - 1)}{\sum_{m=1}^M [\sum_{i=1}^N n(d_i, w_m)P(z_k|d_i, w_m) + (\alpha_{j,m} - 1)]}, \quad (17)$$

$$\hat{P}_{MAP}(z_k|d_i) = \frac{\sum_{j=1}^M n(d_i, w_j)P(z_k|d_i, w_j) + (\beta_{k,i} - 1)}{n(d_i) + \sum_{l=1}^K (\beta_{l,i} - 1)}, \quad (18)$$

$$\alpha_{j,k}^{(n)} = \sum_{i=1}^{N_n} n(d_i^{(n)}, w_j^{(n)})P^{(n)}(z_k|d_i^{(n)}, w_j^{(n)}) + \alpha_{j,k}^{(n-1)}, \quad (19)$$

$$\beta_{k,i}^{(n)} = \sum_{j=1}^M n(d_i^{(n)}, w_j^{(n)})P^{(n)}(z_k|d_i^{(n)}, w_j^{(n)}) + \beta_{k,i}^{(n-1)}. \quad (20)$$

Besides these works, [11] and [8] modified original PLSA model and provided some experimental works on how to get a balance between efficiency and precision.

2.3 Incremental Learning for Collaborative Filtering: Fold-in and Other Naive Solutions

[7] gave a simple solution for the incremental learning for collaborative filtering named Fold-in. They focus on the first and the third incremental scenarios. When a new rating is added to a new or old user \tilde{u} profile, this method only updates the corresponding $P(z|\tilde{u})$ and $P(z|\tilde{u}, y, v)$, leaving $P(v|y, z)$ unchanged. Thus, the offline complexity is $O(n_{iter} \times K \times N)$, while it has an online complexity of $O(k)$ for old users and $O(n_{iter} \times |U| \times K)$ for new user who has rated n_u items.

[3] presented a user-based incremental learning algorithm (IUCF) for collaborative filtering. It updates the user similarities according to the new ratings. One thing needs to be noticed that it only reduced the learning cost, while the recommendation cost is still high. We will compare it with our method in the later section.

Recently, there is also some progress on the incremental update for other model-based CF algorithms such as SVD [4] and NNMF [5], which, in our opinion, are also extensible for collaborative filtering task.

3 Incremental Learning for Collaborative Filtering

In this section, we provide two algorithms named ITPLSA-Forced and ITPLSA-Free for the Incremental learning of TPLSA (ITPLSA) in the forced prediction and free

prediction scenarios, respectively. The intuition behind these algorithms is that for a new rating $\langle \tilde{u}, \tilde{y}, \tilde{v} \rangle$, not only the corresponding $P(z|\tilde{u})$ needs to be updated duly, but also the corresponding $P(v|\tilde{y}, z)$ and $P(z|\tilde{y})$ (for the free prediction mode) need incremental update.

3.1 Incremental Learning Algorithm for Forced Prediction

When a user \tilde{u} gives an item \tilde{y} the rating \tilde{v} , we get a new triplet $\langle \tilde{u}, \tilde{y}, \tilde{v} \rangle$ to update the learnt model. In the forced prediction mode, the corresponding $P(z|u)$ should be updated for this user and the users who also rated \tilde{y} before considering all the items that they rated. In the mean time, $P(v|y, z)$ should be updated for this item together with all the items that the user \tilde{u} has rated before. Finally, $P(z|u, y, v)$ should be updated in the E-Step for all the triplets $\langle u, y, v \rangle$ that $u = \tilde{u}$ or $y = \tilde{y}$. The algorithm for the forced prediction task listed in Algorithm 1 in the Appendix.

3.2 Incremental Learning Algorithm for Free Prediction

For the free prediction mode, it is harder to make the learning process incremental because the updating complexity of the denominator part of Equation (12) is $O(N)$. This makes it infeasible for the model to update online. Luckily, we could rewrite the denominator into four parts as follows:

$$\begin{aligned} \alpha_{z'}^{N+1} = & \sum_{\langle u, y, v, z \rangle : z=z'} P(z|u, y, v) = \alpha_{z'}^N + \\ & \sum_{\langle u, y, v, z \rangle : u=u', z=z'} (P^{N+1}(z|u, y, v) - P^N(z|u, y, v)) + \\ & \sum_{\langle u, y, v, z \rangle : y=y', z=z'} (P^{N+1}(z|u, y, v) - P^N(z|u, y, v)). \end{aligned} \quad (21)$$

Where $\alpha_{z'}^N$ is the old $\sum_{\langle u, y, v, z \rangle : z=z'} P(z|u, y, v)$ before update, and $\alpha_{z'}^{N+1}$ is the updated parameter that will be used in the next incremental update. The incremental part could be calculated efficiently. Therefore we have the update algorithm for the free prediction task as listed in Algorithm 2 in the Appendix.

3.3 Complexity Analysis

The time complexity of Algorithm 1 is $O(n_{iter} \times K \times (n_{\tilde{u}} \times n_{item_rating} + n_{\tilde{y}} \times n_{user_rating}))$, and Algorithm 2's time complexity is $O(n_{iter} \times K \times (n_{\tilde{u}} \times n_{item_rating} + n_{\tilde{y}} \times n_{user_rating} + |Y|))$, a little more computational intensive than Algorithm 1. Where $n_{\tilde{u}}$ is the number of items that the user \tilde{u} has rated before, n_{item_rating} is the average number of rating that these items get, $n_{\tilde{y}}$ is the number of users who has rated item \tilde{y} before and n_{user_rating} is their average number of ratings.

4 Experiments

We proposed incremental speedups for both forced prediction and free prediction in the last section. These two different prediction methods are used in two different recommendation scenarios: rating prediction and item ranking for recommendation. In this section, we experiment our algorithms on these two different tasks separately.

4.1 Forced Prediction

First, we test our forced prediction algorithm on a rating prediction task.

Data Set. Our test data set is MovieLens (<http://www.movielens.org>), a standard data set widely used in the CF literature. It contains 1000209 ratings given by 6040 users to 3883 movies. The ratings range from 1 to 5. The sparsity of the original data is 95.8%. Of these ratings, 80% are training data, and the rest are held as test set. For all the incremental methods, half of the training data, i.e., 40% of all the rating data are learnt incrementally.

The Protocol and Evaluation Metrics. Empirical risks for forced prediction scenario can be computed as the mean absolute error (MAE) and the rooted mean square error (RMSE) [12]. The time cost is the time cost for every single rating update. The speedup ratio is that how many times that the incremental algorithm is faster than it’s non-incremental version.

$$MAE = \frac{1}{N_{testing}} \sum_{r \in Testing} |p_{u,y} - v_{u,y}|,$$

$$RMSE = \sqrt{\frac{1}{N_{testing}} \sum_{(u_i, v_j) \in Testing} (p_{u,y} - v_{u,y})^2}.$$

Comparison Results. We compare our proposed methods with three other methods: IUCF [3], Fold-in and batched TPLSA [7]. For all the PLSA related methods, the topic number is 40. The results are shown in Table 2.

To further investigate our incremental algorithm’s performance for different incremental training dataset percentages, we increase the percentage of incremental training set, the results are shown in Fig. 2. The left Y-Axis is the MAE while the right Y-Axis is the RMSE. We can see the MAE and RMSE increase slightly while the percentage of incremental training part increases from 5% to 50% (the batched training part decreases from 75% to 30%).

Table 2. Comparison between different algorithms for forced prediction

Algorithm		MAE	RMSE	Time Cost (s)	Speedup
IUCF		0.864 ± 0.02	1.125 ± 0.03	0.1 ³	1157.13
Fold-in	Multinomial	0.932 ± 0.06	1.293 ± 0.10	3.5	2244.06
	Gaussian	1.090 ± 0.08	1.458 ± 0.10	4.1	2260.44
TPLSA	Multinomial	0.801 ± 0.04	1.030 ± 0.05	7854.2	0
	Gaussian	0.776 ± 0.03	0.990 ± 0.04	9267.8	0
ITPLSA	Multinomial	0.887 ± 0.05	1.141 ± 0.06	7.6	1033.45
	Gaussian	0.847 ± 0.04	1.112 ± 0.03	8.5	1090.33

³ This only includes the average update time for model learning. The predicting time is not included, which may be very high.

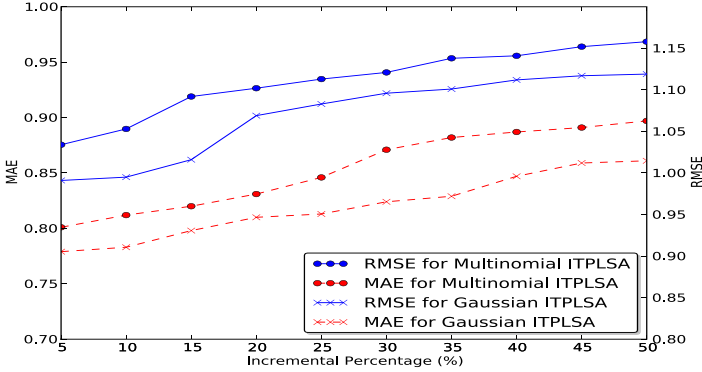


Fig. 2. RMSE and MAE for different incremental training percentages (solid line for RMSE and dashed line for MAE)

4.2 Free Prediction

In order to evaluate the performance of our method on ranking-based recommendation scenario, we experiment our free prediction algorithm on two real-world recommender systems: book recommendation and music recommendation.

Data Set. The book recommendation data are crawled from the largest Chinese book recommendation website: Douban (<http://www.douban.com>). The music recommendation data are crawled from the largest online music recommender system: Last.FM (<http://www.last.fm>). There are no explicit ratings available; instead, we acquire 0/1 implicit ratings from users' visiting history, i.e., a user reads a book or listens to a music. For experiment, we use part of the users and items as our training/testing dataset. For Douban dataset, there are 124 users, 14842 books and 24862 reading records; for Last.FM, there are 675 users, 8010 albums and 14007 listening records. Our training data include 90% of these visiting records (for incremental training algorithms, 40% of the total visiting records are trained incrementally), the rest are held as test set.

The Protocol and Evaluation Metrics. As for free prediction scenario, instead of predicting possible ratings for every item, we generate a ranked list of items that are most probably liked by the user. If the item in the recommendation list is actually visited by the user, we add this item in the *Hitting* set. Then we use two metrics: Precision [13] and Ranking Accumulation (RA) [14] to measure the quality of the generated recommendation list. Apparently, $Precision \in [0, 1]$, the higher the better; $RA \in [\frac{|List|+1}{2}, |List|+1]$ (here $|List|$ is the length of our recommendation list), and smaller value is better. Note that the *Hitting* set only contains the items that both are interesting to the user and already in our test set. So the measured precision and ranking accumulation both underestimates the real performance.

$$Precision = \frac{\# Hitting}{|List|},$$

$$RA = \sum_{item \in Hitting} \frac{rank(item)}{|List|} + \sum_{item \notin Hitting} \frac{|List| + 1}{|List|}.$$

Comparison Results. We also compare our algorithms with Fold-in, batched TPLSA, the results are shown in Table 3 and 4 for Douban and Last.FM dataset, respectively.

Table 3. Comparison between different algorithms for ranking performance on Douban

Algorithm		Precision	RA	Time Cost (s)	Speedup
Fold-in	Multinomial	0.010	20.854	2.1	629.14
	Gaussian	0.012	20.841	2.7	601.59
TPLSA	Multinomial	0.034	20.568	1321.2	0
	Gaussian	0.035	20.562	1624.3	0
ITPLSA	Multinomial	0.030	20.581	4.7	281.10
	Gaussian	0.032	20.570	5.2	312.37

Table 4. Comparison between different algorithms for ranking performance on Last.FM

Algorithm		Precision	RA	Time Cost (s)	Speedup
Fold-in	Multinomial	0.049	20.388	1.5	1118.47
	Gaussian	0.050	20.386	1.6	1139.89
TPLSA	Multinomial	0.081	19.915	1677.7	0
	Gaussian	0.083	19.899	1823.8	0
ITPLSA	Multinomial	0.079	20.093	3.7	453.44
	Gaussian	0.080	20.051	3.9	467.65

4.3 Discussions

The experiments demonstrated that our incremental TPLSA algorithms significantly reduced the update cost for collaborative filtering learning which is very common and important for recommender systems. Although in terms time complexity, our algorithms are not better than Fold-in method or incremental version of user-based CF algorithms. But for both rating-based and ranking-based tasks, our algorithms, especially the Gaussian version, outperform the other two. Moreover, user-based and item-based algorithms are effective in terms of offline computations but more costly in online recommendation phase. We can also see that when incremental training dataset percentage increases, our algorithms' performance does not deteriorate obviously, which means the whole learning phase can be trained incrementally, if needed.

5 Conclusions and Future Work

On the one hand, Internet scale recommender systems are expecting incremental ability of their learning algorithms for their huge and fast changing data set; on the other

hand, as a cutting-edged recommendation algorithm TPLSA lacks the incremental ability which limits its application for collaborative filtering. We proposed incremental treatment of TPLSA for two different while related recommendation problems in this paper. Our methods successfully reduced the update computational complexity while their recommendation results is almost the same as the results given by batched TPLSA algorithm. Further, their effectiveness is proved on a rating based scenario and two ranking based scenarios.

Sometimes incremental improvement is not only about the reduction of their computational cost, both in time and memory, but also makes them more suitable for large, on-line recommender systems. Although this paper focuses on the incremental implementation of TPLSA, other probabilistic models (such as Latent Dirichlet Allocation (LDA), User Rating Profile (URP) models) also suffer from high computational cost problem. This is our future work.

References

1. Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Maching Learning Journal* 42(1-2), 177–196 (2001)
2. Monay, F., Gatica-Perez, D.: PLSA-based Image Auto-Annotation: Constraining the Latent Space. In: *Proceeding of ACM International Conference on Multimedia*, pp. 348–351 (2004)
3. Papagelis, M., Rousidis, I., Plexousakis, D., et al.: Incremental Collaborative Filtering for Highly-Scalable Recommendation Algorithms. In: *Hacid, M.-S., Murray, N.V., Raś, Z.W., Tsumoto, S. (eds.) ISMIS 2005. LNCS (LNAI), vol. 3488*, pp. 553–561. Springer, Heidelberg (2005)
4. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Incremental singular value decomposition algorithms for highly scalable recommender systems. In: *Proc. Fifth International Conference on Computer and Information Technology*, pp. 399–404 (2002)
5. Bucak, S.S., Gungel, B., Gursoy, O.: Incremental Non-Negative Matrix Factorization for Dynamic Background Modelling. In: *Proceedings of PRIS*, pp. 107–116 (2007)
6. Chou, T.C., Chen, M.C.: Using Incremental PLSA for Threshold Resilient Online Event Anlysis. *IEEE TKDE* 20(3), 289–299 (2008)
7. Hofmann, T.: Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.* 22(1), 89–115 (2004)
8. Zhang, L., Li, C., et al.: An Efficient Solution to Factor Drifting Problem in the PLSA Model. In: *Proceedings of the The Fifth International Conference on Computer and Information Technology*, pp. 175–181 (2005)
9. Chien, J.T., Wu, M.S.: Adaptive Bayesian Latent Semantic Analysis. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 198–207 (2008)
10. Marlin, B.: Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto (2004)
11. Das, A., Datar, M., Garg, A., Rajaram, S.: Google News Personalization: Scalable Online Collaborative Filtering. In: *Proc. of the 16th Int. Conf. on World Wide Web*, pp. 271–280 (2007)
12. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Information Systems* 22(1), 5–53 (2004)
13. Wang, J., Robertson, S., De Vries, A.P., Reinders, M.J.T.: Probabilistic relevance ranking for collaborative filtering. *Information Retrieval* 11(6), 477–497 (2008)
14. Zhou, T., Ren, J., Medo, M., Zhang, Y.: Bipartite network projection and personal recommendation. *Physical Review E* 76(4) (2007)

Appendix: Algorithms

Input: New rating triplet: $\langle \tilde{u}, \tilde{y}, \tilde{v} \rangle$

Output: Updated $P(z|u_{\tilde{y}})$ for all the users $u_{\tilde{y}}$ who rated \tilde{y} before and $P(v|y_{\tilde{u}}, z)$ for all the items $y_{\tilde{u}}$ the user \tilde{u} has rated

```

1 if user  $\tilde{u}$  is new then
2   for all the  $z$  do
3      $P(z|\tilde{u}) = 1/K$ ;
4   end
5 end
6 if item  $\tilde{y}$  is new then
7   for all the  $z$  do
8      $P(v|\tilde{y}, z) = \frac{1}{|V|}$ ;
9   end
10 end
11 while not convergent do
12   for all the  $z$  do
13     for all the  $\langle u, y, v \rangle$  triplets where  $u = \tilde{u}$  or  $y = \tilde{y}$  do
14        $P(z = k|u, y, v) = \frac{P(z=k|u)P(v|y,z)}{\sum_{z'} P(z'|u)P(v|y,z')}$ ; // E-Step
15     end
16   end
17   for all the user  $u_{\tilde{y}}$  who rated  $\tilde{y}$  before do
18     for all the  $z$  do
19        $P(z = k|u_{\tilde{y}}) = \frac{\sum_{y'} P(z=k|u_{\tilde{y}}, y', v)}{\sum_{z', y'} P(z'|u_{\tilde{y}}, y', v)}$ ; // M-Step
20     end
21   end
22   for all the  $y_{\tilde{u}}$  that user  $\tilde{u}$  has rated do
23     for all the  $z$  do
24       if Multinomial case then
25          $P(v|y_{\tilde{u}}, z) = \frac{\sum_{u'} P(z|u', y_{\tilde{u}}, v)}{\sum_{u', y'} P(z|u', y_{\tilde{u}}, y')}$ ; //M-Step
26       end
27       else if Gaussian case then
28          $\mu(y_{\tilde{u}}, z) = \frac{\sum_{\langle u, y, v, z \rangle: y=y_{\tilde{u}}} v \times P(z|u, y, v)}{\sum_{\langle u, y, v, z \rangle: y=y_{\tilde{u}}} P(z|u, y, v)}$ ;
29          $\sigma^2(y_{\tilde{u}}, z) = \frac{\sum_{\langle u, y, v, z \rangle: y=y_{\tilde{u}}} (v - \mu_{y, z})^2 \times P(z|u, y, v)}{\sum_{\langle u, y, v, z \rangle: y=y_{\tilde{u}}} P(z|u, y, v)}$ ;
30          $P(v|y_{\tilde{u}}, z) = \frac{1}{\sqrt{2\pi\sigma_{y_{\tilde{u}}, z}^2}} \exp\left[-\frac{(v - \mu_{y_{\tilde{u}}, z})^2}{2\sigma_{y_{\tilde{u}}, z}^2}\right]$ ; // M-Step
31       end
32     end
33   end
34 end

```

Algorithm 1. Incremental learning algorithm for forced prediction

Input: New rating triplet: $\langle \tilde{u}, \tilde{y}, \tilde{v} \rangle$; α_z for all the topics;

Output: Updated $P(z|u_{\tilde{y}})$ for all the users $u_{\tilde{y}}$ who rated \tilde{y} before, $P(y|z)$ for all the users and $P(v|y_{\tilde{u}}, z)$ for all the items $y_{\tilde{u}}$ the user \tilde{u} has rated, new α_z for all the topics

```

1  if user  $\tilde{u}$  is new then
2  |   for all the  $z$  do
3  |   |    $P(z|\tilde{u}) = 1/K$ ;
4  |   end
5  end
6  if item  $\tilde{y}$  is new then
7  |   for all the  $z$  do
8  |   |    $P(v|y, z) = \frac{1}{|V|}$ ;
9  |   end
10 end
11 while not convergent do
12 |   for all the  $z$  do
13 |   |   for all the  $\langle u, y, v \rangle$  triplets where  $u = \tilde{u}$  or  $y = \tilde{y}$  do
14 |   |   |    $Diff\_P(z = k|u, y, v) = \frac{P(z=k|u)P(v|y,z)}{\sum_{z'} P(z'|u)P(v|y,z')} - P(z = k|u, y, v)$ ;
15 |   |   |    $P(z = k|u, y, v) = Diff\_P(z = k|u, y, v) + P(z = k|u, y, v)$ ; // E-Step
16 |   |   end
17 |   end
18 |   for all the user  $u_{\tilde{y}}$  who rated  $\tilde{y}$  before do
19 |   |   for all the  $z$  do
20 |   |   |    $P(z = k|u_{\tilde{y}}) = \frac{\sum_{y'} P(z=k|u_{\tilde{y}}, y', v)}{\sum_{z', y'} P(z'|u_{\tilde{y}}, y', v)}$ ; // M-Step
21 |   |   end
22 |   end
23 |   for all the  $z'$  do
24 |   |    $\alpha_{z'} = \alpha_{z'} + \sum_{\langle u, y, v, z \rangle: y=\tilde{y}, z=z'} Diff\_P(z|u, y, v) + \sum_{\langle u, y, v, z \rangle: u=\tilde{u}, z=z'} Diff\_P(z|u, y, v)$ ;
25 |   |   for all the  $y'$  do
26 |   |   |   // M-Step, the rest part is the same as Alg. 1
27 |   |   |    $P(y'|z') = \frac{\sum_{\langle u, y, v, z \rangle: y=y', z=z'} P(z|u, y, v)}{\alpha_{z'}}$ ;
28 |   |   end
29 |   end
30 |   for all the  $y_{\tilde{u}}$  that user  $\tilde{u}$  has rated do
31 |   |   for all the  $z$  do
32 |   |   |   // M-Step, the rest part is the same as Alg. 1
33 |   |   |    $P(v'|y_{\tilde{u}}, z') = \frac{\sum_{\langle u, y, v, z \rangle: y=y_{\tilde{u}}, z=z'} P(z|u, y, v)}{\sum_{\langle u, y, v, z \rangle: y=y_{\tilde{u}}, z=z'} P(z|u, y, v)}$ ;
34 |   |   end
35 |   end
36 end

```

Algorithm 2. Incremental learning algorithm for free prediction

Interactive Storyboard: Animated Story Creation on Touch Interfaces

Kun Yu¹, Hao Wang¹, Chang Liu², and Jianwei Niu²

¹ Mobile Social Networking, Nokia Research Center Beijing,
No. 5 Donghuan Zhonglu, BDA, 100176, P.R. China

² Department of Computer Sciences, Beihang University,
Haidian, Beijing, 100083, P.R. China

{kun.l.yu, hao.ui.wang}@nokia.com,
{liuchang, niujianwei}@buaa.edu.cn

Abstract. Storytelling, or the telling of narrative, plays a central role in human experience. However, prevalent tools for animated story creation are so complex that expertise is required to grasp them. To help novices such as children create a wide range of visual stories quickly and easily, in this paper we present the design and evaluation of the Interactive Storyboard, an application that enables novices to create vivid animations on touch interfaces, and further to generate live digital stories in a cooperative way. User interviews help us reach the basic assumptions of use cases and design principles for storytelling via Interactive Storyboard. The prototyping application is evaluated by potential users including children and parents, after which improvement on usability and extension of functionalities is conducted. Further usability tests indicate that the Storyboard application is easy to use, more efficient and attractive than formal tools in concrete scenarios, and effectively enhances novices' creativity.

Keywords: Digital storytelling, animation, pen-based, user interface (UI).

1 Introduction

Storytelling “goes back as far as time allows us to remember” [1]. Stories that pass on human experience are now told through many media because of fast development of computers and personal multimedia devices such as mobile phones. There exist significant opportunities that new technologies offer in reshaping the way in which narrative is conceived and presented, especially for the evolution of children's narrative for education and entertainment.

The art of storytelling has always been under the impaction of new technologies [2]. Interactivity, traditionally used in story books to engage, is greatly enhanced in digital environment. Balabanovic et al. [3] presented a StoryTrack device that demonstrates storytelling with digital photos which is similar to some kinds of story sharing that people enjoy with print photos. The form of story creation with StoryTrack can be simply summarized as to select proper photos onto the working track and record audio clips as annotation of the story. The authors also depicted their findings that there appeared to be two different styles of storytelling: *Photo-driven*, for which the

subject explains every photo in turn, and the story prompted by the existing sequence of pictures; and *Story-driven* in which the subject has a particular story in mind, then gathers the appropriate photos and recounts the story.

Salovaara [4] reported a case study on the use of Comeks, a mobile tool for creating and sharing comic strips as MMS messages. Somewhat similar to this, Jokela et al. [5] presented the design and evaluation of the Mobile Multimedia Presentation Editor, an application that makes it possible to author sophisticated multimedia presentations that integrate several different media types on mobile devices. The focus of this work is to enable creation of more continuous media, giving the user more fine-grained control over the temporal structure of the presentation, and also enabling the use of audio in the created presentations.

Landry and Guzdial [6] presented an examination on the support mechanisms used by the Center for Digital Storytelling (CDS) to help everyday people unlock the stories captured in their images and video through the practice of digital storytelling. They also recorded the observed challenges such as story development, content preparation, movie production, and process management, because composing digital narratives requires a more involved creation process.

All the above work utilized rich media such as text, image, video, audio and their combinations. It is a big challenge for novices to learn how to prepare content in different media forms and compose them with a complex process in a short period. More specifically, the most expressive form of narratives for children might not be photos of real life, but rather to be cartoon-like drawings in animations, as appeared in easy books. Another factor, as revealed by Millard et al. [7], children are more concerned about levels of fun rather than levels of efficiency, and they often do not stay focused on a task for very long. This implies that a storytelling tool for novices (esp. for children) should be simple and attractive enough.

Inspired by a drawing program which is specifically designed for children [8], we have developed a pen-based visual storytelling system called Interactive Storyboard. One similar design on touch interfaces is K-Sketch proposed by Davis et al [9], a general-purpose, 2D animation sketching system which is simple and effective for novices to create a wide range of works. While the user interface of K-Sketch is complex with buttons and temporal control, we aim to enable creation of elegant and colorful characters with scene construction, and simplify traditional temporal structure to enable novices to create animations and tell stories more naturally. Puppet Master is an animation creation design proposed by Young et al [10], in which paired behavior is learned from the motion training of designers, and then the system synthesizes the motion of the reacting character based on the control of user on the main character. It is a funny design, but if the user wants accurate control on the reacting character, it is not possible. Instead of the use-after-training metaphor, Interactive Storyboard endows the characters with inherent motion features, which are intuitive, vivid and easy to use. Moreover, in our proposed design, it also supports the collaborative creation from multiple users, and each user does not need to spend a long time working on the whole animation composition.

With this application, users can create elegant visual elements, and use these 'living' elements to make vivid animations to tell narrative. Our goal has not been set to design novel interaction technologies but rather to focus on high-level selections of tool features, and optimal combination of these features to maximize user experience

in the content organization of a story, and try to release the creativity of users which is technically baffled when they are using other tools.

The rest of this paper is structured as follows. Our interview of users is first introduced, in which we try to gain understanding of user's expectations on the application. Based on this, the basic assumptions of the use cases and design principles for Interactive Storyboard are introduced. Then we describe our approach, and report the evaluations of the developed prototype. After this we present the usability issues raised during the tests and discussions before concluding this paper.

2 User Interview

For better understanding of user needs, we began our study with interviewing fourteen people including five children and nine adults. The questions were focused on the key features they expected, and the time they would like to spend on creating an animated story, as shown in Table 1. Thirteen out of the fourteen interviewees have limited usage with pen-based animation tools. Although their background varies, commonalities in their answers did emerge in some sense. All participants, especially children showed great interest in creating digital stories, but many blamed that the tools they used were either too weak like MS Paint that could not meet their needs, or too complex like Photoshop or Flash, which required too much time to learn.

Table 1. Summary of interviews

Occupation	Domain	Expected features and composing time
1. Graduate student	Engineering	Arbitrary drawing with easy operations; 2-10 min.
2. Designer	Art	What you see is what you get; 2-3 hours
3. Secretary	Business	Easy to learn, drawing patterns available, colorful; 10 min.
4. Researcher	Science	Copy/paste, cartoon; 20 min.
5. Engineer	EE	Easy to learn, intuitive; 5 min.
6. Researcher	Science	Convenient to learn, shapes and colorful models; 1 hour
7. Assistant	Sociology	Funny, simple, professional; 15 min.
8-12. Children	N/A	Easy, colorful, diverse brushes, easy shapes; 5-30 min.
13. Postgraduate	Math	Consistent with common sense; no requirement on time
14. Researcher	Linguistics	Easy to learn and use; 50 min.

During the interviews, we also described possible designs for a storyboard tool, and suggested ways that users might be able to create visual stories, after asking for their expectations on the tool. As shown in Table 1, most people expressed their

expected time for drawing a picture and making animation to be about ten minutes or at the same level. All the children wanted the new tool to be simple while providing colorful and funny effects. Almost all participants expected intuitive user interface and operations, and would not spend much time on learning to use the tool.

The interviews convinced us that there is a need for an easy-to-grasp tool that requires little time to learn while giving more flexibility for enjoyment.

3 Design Principles

With the information collected during the interviews, we based our design on the expectations of the interviewed users.

3.1 Design Assumptions

Two basic assumptions were derived from the user needs explored during the interviews. Based on the observation of children who can not concentrate upon a task if it is not fun and unnecessarily complex, our primary assumption is that most non-professional users can afford an average of less than 20 minutes to compose an animation, including the training time. The second assumption is that most users will focus on the process rather than the results. This assumption is derived from the nature of storytelling since most people enjoy the process of telling a story itself, but do not care too much about the final result it produced (most interviewees do not expect to be real artists just with the help of a tool, but they want to be engaged in the process of creation). This means that functionalities of the tool will be acceptable if they are interesting enough to attract people.

3.2 Targets of Design

Based on the above assumptions, three design targets were derived as follows: consistency, intuitiveness, and attraction.

The principle of consistency was mainly learnt from the fact that most users had the experience of using computer-based drawing tools. The build-in habits of using familiar functions will accelerate the adaptation of users to a new tool. Users feel comfortable to follow a process according to their existing knowledge, which means that the design of the Interactive Storyboard can be inspired by some known UI features.

The second principle, intuitiveness, is tightly related to the consistency principle. While previous knowledge can help users get familiar with some common features, understanding new features introduced by a new tool will need users' imagination and intuition. This is essential when the users want to learn to use the tool quickly and naturally by themselves.

The third principle is attraction, which is based on the second assumption that participants pay particular attention to the creation process rather than the final result. Most users are best-effort triers who do not have the whole picture in mind when starting to create a digital visual story; therefore the outcome of the story typically changes during the composition process. If any intermediate step lacks expressiveness or attraction, the users might lose interest to go forward.

4 Storyboard Application

The Interactive Storyboard is intended for pen-based mobile devices, e.g. Nokia N800 (Figure 1a), an Internet tablet, and Nokia N5800 device (Figure 1b). For benchmarking purposes, we have also implemented this application on a Dell Tablet PC (Figure 1c).



Fig. 1. Currently Interactive Storyboard runs on two portal devices. (a) Nokia N800 device and (b) Nokia N5800 mobile phone (c) Dell Tablet PC.

The process of our development and evaluation is shown in Figure 2. Based on the assumptions derived from user interview, paper prototype was created at first to demonstrate the basic functionalities and interactions. Afterwards prototyping application was built on tablet PC and first round usage test with twelve participants (ranging in age from 6 to 40) was conducted. Functionality improvement and extension was carried out based on the first round feedbacks. Then the second round usability test was conducted to validate the improvement. Finally the design and implementation was tailored for mobile devices. Due to the space limitation, here we only present the form of application after the first round usage test.

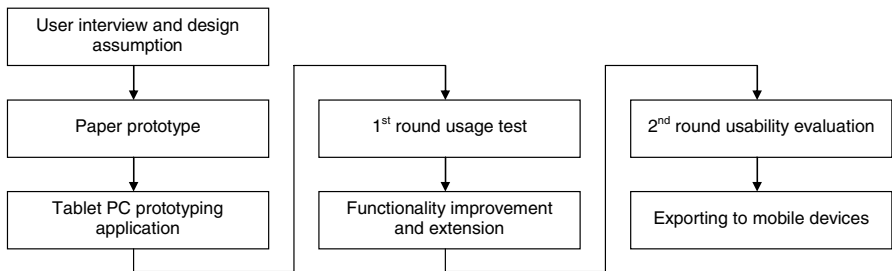


Fig. 2. Development and evaluation process

4.1 User Interface

The user interface of the Interactive Storyboard is illustrated in Figure 3. It is visually divided into three parts: the primary tool bar on the left, the drawing canvas in the center, and the secondary panel on the right. Buttons in the primary tool bar are raw categories of the operations, including drawing, object selection, loading scenes, etc. Buttons in the secondary panel further specify the operations, including pen/brush selection, element selection and specific animation setting and tuning functionalities.

The creation of a digital visual story starts with the selection of scenes and characters, and here we refer to the characters as elements. Thus our design mainly includes two main modules. The scene and element composition module allows users to create and edit scenes and elements, while the animation creation and storytelling module enables users to make animation based on scenarios, and share the story with others. Although each module corresponds to different stages of storytelling, users can work at any stage any time, giving them enough flexibility to operate according to their current idea without worrying much about the sequence of the operations.



Fig. 3. User interface of Interactive Storyboard

4.2 Scene and Element Composition

In our application, users have several methods to apply scenes into their works. Besides drawing a scene from scratch, the user could also choose from a set of typical background pictures that are available in the system. Users can combine all the scenes and elements provided in their desired way to create new scenes. Commonly used elements such as trees, houses and animals can be loaded onto the existing scenes as separated layers and dragged freely to any position. Their front-back relationship can be adjusted by clicks, where the element clicked last will be topmost.

4.3 Animation Creation

Animation creation used to be challenging for users, especially when they are struggling with the frame-level animation composition and tuning. This boring but seemingly necessary step has jeopardized most fun during creation, and meanwhile kept most novices away from their animated imaginations. In the proposed method, we try to facilitate the operation of users with the most natural and intuitive touch/pen gestures, to provide users with rich contextual support while maintaining the flexibility of the creation.

4.3.1 Animation of an Element

The underlying concept of the animation creation is based on pen stroke. Any graphical object on the drawing canvas could be endowed with an animated feature if a stroke is drawn from it, and the motion path is marked by the stroke. The animation is

not confined to positional change only, and it also includes the varying poses of the selected graphical object. To support this feature, an embedded graphical database is used to store various gestures and postures of the respective graphical object. At each point along the specified stroke (motion path), the tangent direction is computed by the system, and mapped to a specific gesture of the character as shown with the butterfly in Figure 4. In this way the users can easily make all desired objects moving along any tracks, and the application itself takes care of the vivid animation effects.

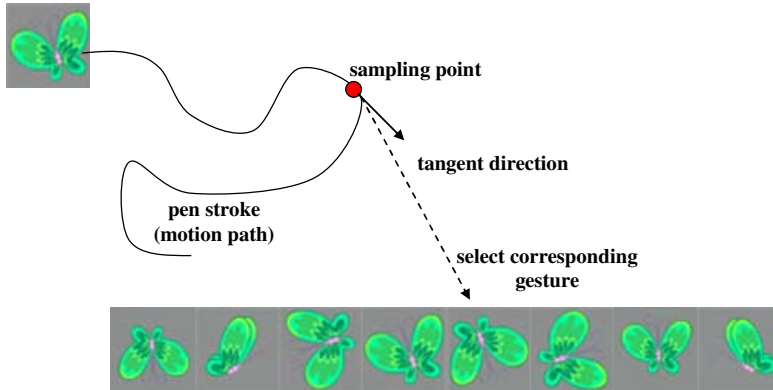


Fig. 4. Gesture of the object along the tangent of the motion track

A speed-control method is also supported while drawing the stroke. Speed options (e.g., very slow, slow, moderate, fast, very fast) are widely adopted in existing tools such as PowerPoint. But this solution has several shortages. First, users have to explicitly specify the speed option after the motion path is drawn, which is not convenient. Secondly, it is not easy for the users to get clear understanding of the quantitative meaning of each option. An alternative is to let users input the accurate duration of the motion (e.g. in seconds), which might be even more accurate but boring. Thirdly, an object can only move uniformly along the defined motion path in this case. If users want an object to move faster and faster, they have to split the motion path into several small segments and specify the speed option for each segment respectively. Our design solves this problem in an intuitive way: the speed of an object is determined by the rate of the pen stroke: the faster the stroke is drawn, the faster the object will move, and vice versa. Thus the users can specify a complex motion path with non-uniform rate by just drawing one stroke.

4.3.2 Multi-element Animation

When more than one element is involved in an animation, our application should be able to support both sequential and synchronous motions. The synchronous motion setting is something challenging. It is not easy for novice users to organize such a motion process, for instance, in commercial software, e.g. MS PowerPoint and Macromedia Flash. In our design, we intend to keep the control of animation mode simple and intuitive, so an extra button is introduced to control the animation mode, as shown in the center of Figure 5. Every time this button is pressed, the animation is switched between sequential movement mode and synchronous movement mode.

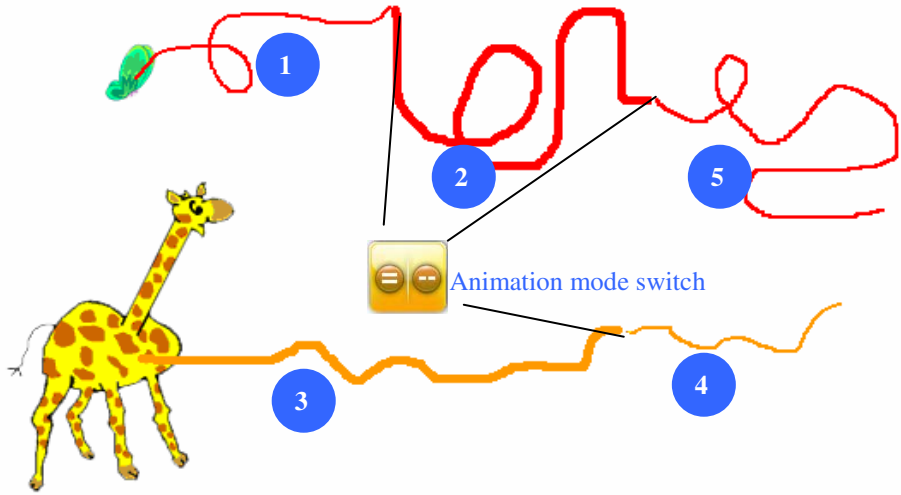


Fig. 5. Animation mode control. Original state is sequential, one switch changes to synchronous state, another switch changes back to sequential state. The animation is like such: the butterfly flies along path 1, then the butterfly and the giraffe move along path 2 and 3 at the same time, then the giraffe moves along path 4, before the butterfly flies along path 5.

4.3.3 Multi-user Cooperative Animation Creation

The current platform also supports animation creation with multiple users. To be specific, several devices could be connected together via wireless network to form a creation group, and thus the operations of one user could be synchronized with all the other peer users. A sketch, a selected graphical object, or the animation setting could be shared with others synchronously. One evident benefit of this creation method is that users could focus on the parts which they are most adept at. The communication between users becomes increasingly important in this type of teamwork, as everyone is expected to contribute at the correct time.

5 Usability Test and Discussions

Six right-handed participants, including 3 male and 3 female ranging in age from 6 to 32, volunteered for the test. Among the participants there were two children (between 6 and 9 years old), and for the four adult participants, three of them had experience in using MS PowerPoint. Most of the adults have children.

All the testers were asked to compose an animation shown in Figure 6 as fast as possible on the interface of a Dell Tablet PC. This task requires the participants to compose a short episode of animation (approximately 40 seconds in length). The script of the animation story was shown to all participants together with a piece of video example, which is about a “forest sports game” and the protagonists include an elephant, a giraffe, and two butterflies as well as some scene elements such as houses and trees. The motion paths of these objects are complex with specific sequence,

which are illustrated in Figure 6. Seven different objects are selected, and their respective motion is set at different time segments. Specifically, between t_0 and t_1 , only A will move. After that, B will join A in the movement between t_1 and t_2 . Then, after t_2 both of them stop moving, while the movement of C starts. There is no strict requirement on the accuracy of the motion traces. Test sessions lasted between 1.1 and 2.75 hours. We minimized our interaction with participants during the animation task. The three volunteers who had PowerPoint experience also performed the same task with PowerPoint. They needed help 2-5 times during the tasks when working with PowerPoint. When using our tool, only one adult needed help once to finish the task, and the children got occasional help from us.

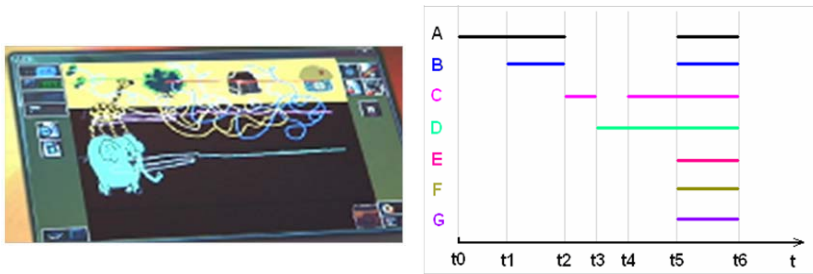


Fig. 6. Motion sequence setting for the testing task

The task time was recorded as a direct reflection of the efficiency of the design, as illustrated in Table 2. Testers spent much less time for a specific animation creation, compared with PowerPoint. After spending more than one hour with PowerPoint, some testers even gave up composing the animation with it. Besides the time to complete the animation, we also recorded subjective feedback on Interactive Storyboard and PowerPoint from the participants, which included how satisfied they felt with

Table 2. Time to complete the animation task with individual records, means and deviations in minutes. (PT: practice time, TT: task time, EXP: experience time, SD.: standard deviation).

Tester	Storyboard		PowerPoint	
	<i>PT</i>	<i>TT</i>	<i>EXP/PT</i>	<i>TT</i>
1	7.0	3.2	30	55
2	1.3	3.8	>150	28
3	6.5	10.3	>120	29
4	4.0	5.2	/	
5	13.8	6.3		
6	5.6	8.5		
Mean	6.37	6.22	>100	37.3
SD.	4.18	2.75	/	

their animations created, and how easy it was to work with the tools. The four adults were also asked to give an acceptance score of how acceptable it was for them to use our tool to create animations and show to others (e.g., their children). All the subjective feedbacks were measured on a Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), and the result is shown in Table 3.

Considering the design principles of consistency and intuitiveness, the buttons did not frustrate the participants, and they felt the tool was easy to use and significantly less effort was required to grasp it compared with PowerPoint. It was exciting for them to verify a new feature by easy trials, e.g. animation setting, which was identical to their intuition. Regarding the principle of attraction, the participants were highly engaged and never lost their interests of practice, which is also proved with the high average acceptance score of 6.25 and satisfaction score of 6.

We have not carried out focused tests on the cooperative creation of animation, because no similar tools have been found for benchmarking purposes. However, in the practical but informal trials with children, they agree that they enjoy this type of recreation. One interesting finding is that the people involved in the co-creation may not be cooperative, and some of them may intentionally make troubles with others during the creation. However, as we claimed before, the users care much more on the process than the result. And the integration between artists and trouble-makers really entertains all of them.

Table 3. Subjective feedbacks on the animation task

Evaluation Criteria	Storyboard		PowerPoint
	<i>Mean</i>	<i>SD.</i>	<i>Mean</i>
Satisfaction	6	1.26	3.67
Easiness	5.33	1.37	3.33
Acceptance	6.25	0.96	

6 Conclusion

We have presented the Interactive Storyboard, a pen-based application that enables novices to create elegant visual components and make animations to generate vivid digital stories. Based on the investigation of user needs, our design is adaptive to users' intuition and consistent experience. A novel animation control method is introduced and implemented, which is much faster than a more formal tool (MS PowerPoint) in our experiment. The experiments proved that our design is intuitive and attractive, which retains users' interests and simplifies the process of animation creation.

In the near future, we plan to push the research forward, with emphasis on the different cooperative modes of multiple users, or user teams. Some exemplary use cases may be observation and creation, team competition, and also online instructions. In addition we also plan to conduct a field study of the Interactive Storyboard with the participation of a larger population of users.

Acknowledgments

The authors would like to thank all the people who have supported us with constructive advice and sincere help during the design and development of Interactive Storyboard. Our appreciations also go to those who have helped us in testing the application, whose contribution is also indispensable to improve its overall performance and fine-tune our idea.

References

1. Landry, B., Guzdial, M.: Learning from Human Support: Informing the Design of Personal Digital Story-Authoring Tools. In: Proc. of CODE 2006 (2006), http://www.units.muohio.edu/codeconference/proceedings/conference_papers4.htm
2. Madej, K.: Towards Digital Narrative for Children: From Education to Entertainment: A Historical Perspective. *ACM Computers in Entertainment* 1 1, 1–17 (2003)
3. Balabanovic, M., Chu, L., Wolff, G.: Storytelling with Digital Photographs. In: Proc. CHI 2000, pp. 564–571. ACM Press, New York (2000)
4. Salovaara, A.: Appropriation of MMS-Based Comic Creator: From System Functionalities to Resources for Action. In: Proc. CHI 2007, pp. 1117–1126. ACM Press, New York (2007)
5. Jokela, T., Lehtikoinen, J.Y., Korhonen, H.: Mobile Multimedia Presentation Editor: Enabling Creation of Audio-Visual Stories on Mobile Devices. In: Proc. CHI 2008, pp. 63–72. ACM Press, New York (2008)
6. Landry, B., Guzdial, M.: Learning from Human Support: Informing the Design of Personal Digital Story-Authoring Tools. In: Proc. of CODE 2006 (2006), http://www.units.muohio.edu/codeconference/proceedings/conference_papers4.htm
7. Millard, N., Lynch, P., Tracey, K.: Child's Play: Using Techniques Developed to Elicit Requirements from Children with Adults. In: Proc. Third Int. Conf. on Requirements Engineering 1998, pp. 66–73 (1998)
8. Tux Paint, <http://www.tuxpaint.org/>
9. Davis, R., Colwell, B., Landay, J.: K-Sketch: A Kinetic Sketch Pad for Novice Animators. In: Proc. CHI 2008, pp. 413–422. ACM Press, New York (2008)
10. James, E.Y., Takeo, I., Ehd, S.: Puppet Master: Designing Reactive Character Behavior by Demonstration. In: Proc. ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Dublin, Ireland, pp. 183–191 (2008)

Comparative Evaluation of Reliabilities on Semantic Search Functions: Auto-complete and Entity-Centric Unified Search

Hanmin Jung¹, Mi-Kyoung Lee¹, Beom-Jong You¹, and Do-Wan Kim²

¹ Dept. of Information Technology Research, KISTI

52-11 Eueon-dong, Yuseong-gu, Daejeon, Korea 305-806

² Dept. of information and communications, Paichai University

14 Yeon-Ja 1 Gil Seo-gu, Daejeon, Korea 302-735

jhm@kisti.re.kr

Abstract. Although users require increased reliabilities on the function and information that Web information services offer, there are few studies in the Semantic Web applications on reliability and usability evaluation on semantic service functions owing to the lack of successful use cases and technological immaturity. This research is originated from a previous lesson that reliabilities needs to be assured in ways of improving service functions as well as retrieved information, particularly, in the services giving new user experiences such as semantic search services. We comparatively evaluated reliabilities of ‘auto-complete’ and ‘entity-centric unified search’ functions in semantic services with the following two criteria: the one is ‘precision and satisfaction scores’ on the functions and retrieved information that the user rates, and the other is ‘precision in task performance’ as a quantitative analysis and reliability of expectation-result as a qualitative analysis that an observer assesses. The experimental results indicate that precision is closely related with satisfaction in the view point of the user, and further reliabilities can not be taken off various factors for evaluating software quality.

Keywords: Reliability, Usability Test, Semantic Search, Auto-complete, Unified Search, Entity-centric Search, Semantic Web, Ontology, Named Entity.

1 Introduction

The Web, as a core infrastructure in information society, is providing more and more services including information search, community service, and shopping. Particularly, semantic services using the Semantic Web technologies such as ontology and reasoning are emerging as a new information service paradigm. However, there are few studies on reliability and usability evaluation on the services because of the lack of successful use cases and technological immaturity although users require increased reliabilities on the function and information that Web information service offer. Only after 2008, a workshop in ASWC (Asian Semantic Web Conference) was held to identify and bring to the attention of the research community the human factors in the

Semantic Web technology at large¹. This is the reason why we study on reliabilities of semantic services with the background of the experience about usability test.

ISO 9126-1 defines reliability, one of software qualitative characteristics, as “a set of attributes that bear on the capability of software to maintain its level of performance under stated conditions for a stated period of time” [1]. ISO 15489 defines reliability as an information feature of contents management² whereas ISO 9126-1 treats as a functional feature of software as the above mentioned [2]. When regarding the Web as software and further as information management service system, reliabilities on the Web information services including semantic services can be divided into the followings.

- Reliabilities on the functions that information services provide
- Reliabilities on information retrieved as a result of function execution

In conclusion, reliabilities on information services need to be assessed with the above criteria by users as well as with the criteria of compliance, fault tolerance, recoverability, and maturity.

2 Semantic Search Services

OntoFrame, has been developed since 2005, is a semantic service platform to provide semantic search on IT/BT information [3]. It gathers full-text documents with metadata, and then transfers them into semantic knowledge in the form of RDF (Resource Description Framework) triples while referring to pre-defined ontology schema. A rule-based reasoner expands the knowledge by interpreting user-defined reasoning rules.

New services on the platform are yielded every year. OntoFrame 2007 and OntoFrame 2008 which are test beds of this research also are opened to the public³. They include hundreds of thousands international journal papers, and provides tens of service functions such as ‘auto-complete’ and ‘entity-centric unified search’. Particularly, the services give analytic information about research topic keywords. For example, when a user select a research topic keyword, they show ‘topic trends’, ‘significant researchers’, ‘significant institutions’, ‘researcher network’ and so on. All information in the services are controlled and identified by URI (Uniform Resource Identifier) scheme.

We have performed usability tests for each service [4]. A result observed from the usability test of OntoFrame 2007 is that there is a high variation in task performance among the users although all service functions are enough to accomplish given tasks. We found the cause of this instability is originated from user’s reliability on service functions. The users with experiences suffered from functional defects during search and navigation tend not to rely on the information service provides any more, but to try to use their own priori knowledge for task performance.

¹ <http://sites.google.com/site/humanfactorsandsemanticweb/1st-workshop-human-factors-and-the-semantic-web>

² “A reliable record is one whose contents can be trusted as a full and accurate representation of the transactions, activities or facts to which they attest and can be depended upon in the course of subsequent transactions of activities”.

³ <http://ontoframe.kr/2008/>

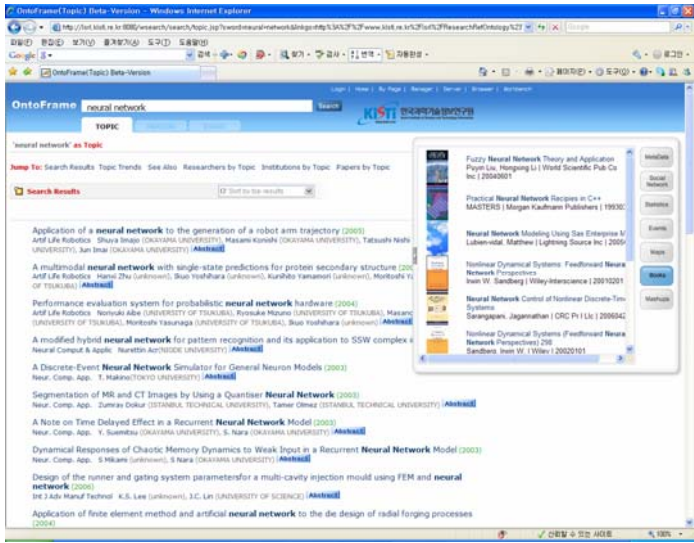


Fig. 1. A screen shot of OntoFrame 2007 (search keyword: “neural network”)

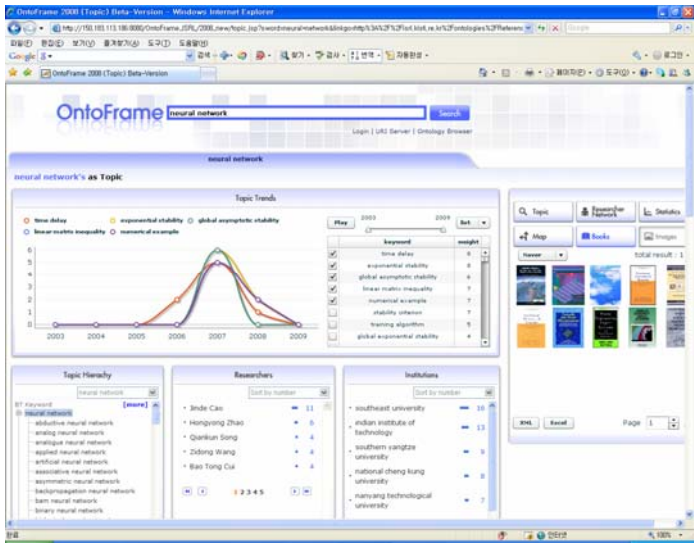


Fig. 2. A screen shot of OntoFrame 2008 (search keyword: “neural network”)

This empirical result gives us a lesson that reliabilities needs to be assured in ways of improving service functions as well as search results, particularly, in the services giving new user experiences such as semantic search services. Thus, some search functions including the followings were improved in OntoFrame 2008.

- Auto-complete: This function displays entities reserved in service system (partially) matched with user's input keyword. The entities are grouped by their types such as 'TOPIC' and 'PERSON'. They also stand guarantee for successful search results retrieval, that is, for existence of coincided documents.
- Entity-centric unified search: This function generates dynamically search results for entity types corresponding to user's search query. It is improved to cover combination of multiple entities such as 'TOPIC + TOPIC' and 'TOPIC + PERSON' as well as single entity.

3 Comparative Evaluation of Reliabilities

Comparative evaluation of reliabilities on semantic search functions is performed by both the users participating in this experiment and an observer who is an HCI expertise. The users rate 'precision and satisfaction scores' on the functions and retrieved information. The observer assesses 'precision in task performance' as a quantitative analysis and reliability of expectation-result as a qualitative analysis.

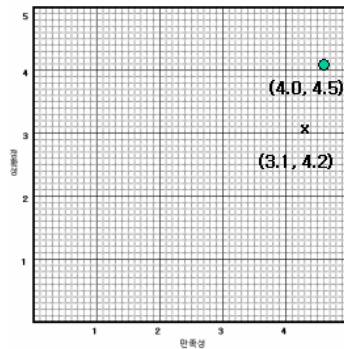


Fig. 3. Example of section paper marked for precision and satisfaction factors

Five undergraduate students participated in the experiment. All of them are acquainted with internet search and majoring in information and telecommunications. The observer monitors the users using a camcorder and separate monitors. The section paper shown in figure 3 is used for marking 'precision and satisfaction scores' by the users. The range of each score is 0 ~ 5.

3.1 Auto-complete

'Auto-complete' is a feature predicting a word or phrase that the user wants to type in without the user actually typing it in completely⁴. OntoFrame 2008 improved this function so that only keywords ensuring successful search results can appear in its own auto-complete list in ways of taking a new approach using auto-complete table management [5]. The previous version in OntoFrame 2007 does not guarantee

⁴ <http://en.wikipedia.org/wiki/Autocomplete>

successful search results because it contains static keyword list extracted from full-text papers once in advance before document indexing. Many enterprise search services have also the same problem similar with that of OntoFrame 2007 as their auto-completes entirely depend on user's popularity in search keywords or search keyword dictionary.

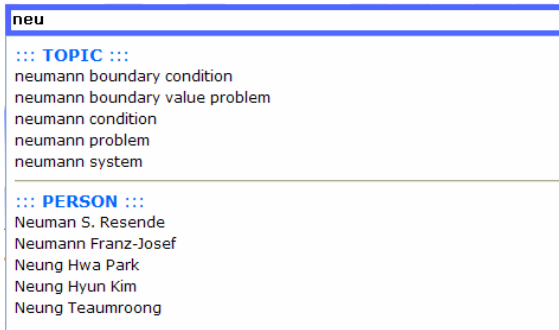


Fig. 4. Example of auto-complete list generated from topic keyword “neu” in OntoFrame 2008. Keywords in the list are divided into two entity types of ‘TOPIC’ and ‘PERSON’.

The following task is given to the users for evaluating reliabilities of the two services.

“Find the number of papers published by the most significant researcher and the most significant institution in each of search results retrieved from top-5 ranked topic keywords in auto-complete list, using a topic keyword “neural”. (1)

Table 1 shows an example of result performed and recorded by a user after the task performance. ‘Observed result’ and ‘precision in task performance’ are assessed by the observer. ‘User expectation’, ‘user’s opinion’ and ‘user’s conclusion’ are recorded by thinking-aloud protocol. ‘Precisions in task performance’ of all the users and their average scores for the two versions are shown in table 2. Table 3 indicates comparative results in the viewpoint of the users; the upper part is for precision and satisfaction scores in ‘auto-complete’ function itself, the lower part is for precision and satisfaction scores in retrieved information. All the scores are acquired from the users. We normalize average scores into the range of 0 ~ 1 to directly compare with ‘precisions in task performance’.

When analyzing based on reliability of expectation-result inspired from GOMS (Goals, Operators, Methods, and Selection rules) [6], both versions are satisfactory for ‘fault tolerance’ and ‘recoverability’. However, ‘auto-complete’ function of OntoFrame 2007 shows lack of ‘maturity’ in that its auto-complete list does not filter inappropriate search keywords such as “functional magnetic resonance imaging study” and “neural activity associated with stress-induced cocaine craving”, which bring out search failures. It also needs to build ‘compliance’. The users should type ‘space’ key to operate, and have to use only mouse instead of upper/lower direction keys to select a keyword in auto-complete list.

Table 1. Example of result performed and recorded by a user after the performance

	OntoFrame 2007	OntoFrame 2008
User expectation	After giving “neural” in search box, I will get a set of topic keywords in auto-complete list, and then search results including researchers and institutions.	After giving “neural” in search box, I will get a set of topic keywords in auto-complete list, and then search results including researchers and institutions.
User’s opinion	Finding topic keywords was easy, but search results retrieved from the keywords did not show researcher information.	Researchers and institutions were successfully found in the search results retrieved from the top 5 keywords in auto-complete list.
User’s conclusion	Incomplete auto-complete	Satisfactory auto-complete
Observed result	All of the top 5 keywords in the auto-complete list failed in the given task	All of the top 5 keywords in auto-complete list succeeded in the given task
Precision in task performance	0.0 (0/5)	1.0 (5/5)

Table 2. ‘Precisions in task performance’ of the users assessed by the observer

User	OntoFrame 2007	OntoFrame 2008
1	0.2	1.0
2	0.2	1.0
3	0.0	0.6
4	0.4	1.0
5	0.0	1.0
Average	0.16	0.92

Table 3. ‘Precision and satisfaction scores’ on ‘auto-complete’ function and retrieved information

User	OntoFrame 2007		OntoFrame 2008	
	<i>Function</i>			
	Precision	Satisfaction	Precision	Satisfaction
1	3.0	3.0	4.0	4.0
2	3.0	3.0	4.9	4.9
3	3.8	2.5	4.5	4.5
4	2.0	2.0	4.2	4.2
5	2.5	2.6	4.5	4.3
Average	2.86	2.62	4.42	4.38
Normalized Average	0.57	0.52	0.88	0.88
	<i>Information</i>			
	Precision	Satisfaction	Precision	Satisfaction
	3.2	3.1	4.5	3.6
	2.5	2.9	4.9	4.9
	2.0	3.0	3.0	4.0
	2.8	2.8	4.0	4.2
	3.0	3.0	4.0	4.0
Average	2.7	2.96	4.08	4.14
Normalized Average	0.54	0.59	0.82	0.83

The comparison between ‘precision in task performance’ assessed by the observer and ‘precision and satisfaction scores (*in normalized average*)’ rated by the users indicates that the function of OntoFrame 2008 is superior to that of OntoFrame 2007 in both criteria as we expect. ‘Precision and satisfaction scores’ of OntoFrame 2007 is much higher than ‘precision in task performance’ of OntoFrame 2007 unlike the case of OntoFrame 2008 because users tend to avoid giving excessive scores.

3.2 Entity-Centric Unified Search

‘Entity-centric unified search’ is to generate an appropriate search result according to entity types acquired from the user’s search query. This function of OntoFrame 2008 improved to cover combination of multiple entities with flexibility to cope with various search queries whereas that of OntoFrame 2007 limits to only single entity [3] [7]. For example, when search query “neural network exponential stability” is given, the latter generates a simple search result consisting of papers, but the former generates a mixed search result including information relevant with the two topic keywords (See figure 5).

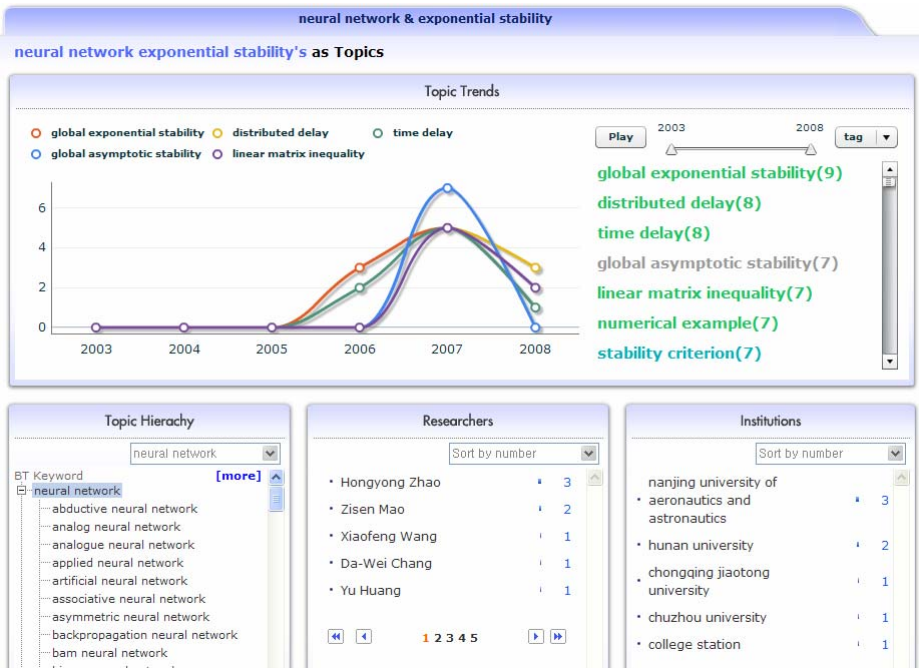


Fig. 5. Example of topic page generated from two topic keywords “neural network” and “exponential stability” in OntoFrame 2008

The following task is given to the users for evaluating reliabilities of the two versions.

“Find the number of papers published by the most significant researcher and the most significant institution, using topic keywords “neural network” and “exponential stability”.” (2)

Table 4 shows an example of result performed and recorded by a user after the task performance. ‘Precisions in task performance’ of all the users and their average scores for the two versions are shown in table 5. Table 6 indicates comparative results in the viewpoint of the users; the one is for precision and satisfaction scores in ‘entity-centric unified search’ function itself, the other is for precision and satisfaction scores in retrieved information.

Table 4. Example of result performed and recorded by a user during his task performance

	OntoFrame 2007	OntoFrame 2008
User expectation	After giving “neural network” and “exponential stability” in search box, I will get search results including researchers and institutions.	After giving “neural network” and “exponential stability” in search box, I will get search results including researchers and institutions.
User’s opinion	I found a set of papers retrieved from the two keywords, but the search results do not show significant researchers and institutions.	Researchers and institutions were successfully found in the search results retrieved from the two keywords.
User’s conclusion	Disappointed in results lower than expected	Satisfactory unified search
Observed result	All of the search keywords failed in the given task	All of the search keywords retrieved successful results
Precision in task performance	0.0 (0/5)	1.0 (5/5)

Table 5. ‘Precisions in task performance’ of the users assessed by the observer

User	OntoFrame 2007	OntoFrame 2008
1	0.0	1.0
2	0.0	1.0
3	0.0	1.0
4	0.0	1.0
5	0.0	1.0
Average	0.00	1.00

When analyzing based on expectation-result reliability, both services are satisfactory for ‘compliance’, ‘fault tolerance’, and ‘recoverability’. However, ‘entity-centric unified search’ function of OntoFrame 2007 shows lack of ‘maturity’ in that it does not successfully generate search results for search queries including multiple entities such as “neural network exponential stability⁵” and “neural network Jinde Cao⁶”.

The comparison between ‘precision in task performance’ assessed by the observer and ‘precision and satisfaction scores (*in normalized average*)’ rated by the users indicates that the function of OntoFrame 2008 is superior to that of OntoFrame 2007

⁵ TOPIC (“neural network”) + TOPIC (“exponential stability”).

⁶ TOPIC (“neural network”) + PERSON (“Jinde Cao”).

Table 6. ‘Precision and satisfaction scores’ on ‘entity-centric unified search’ function and retrieved information

User	OntoFrame 2007		OntoFrame 2008	
	<i>Function</i>			
	Precision	Satisfaction	Precision	Satisfaction
1	1.7	2.3	3.4	3.6
2	3.0	4.0	4.9	4.9
3	2.0	3.0	3.0	4.0
4	1.6	1.7	4.8	4.7
5	1.6	1.6	4.6	4.7
Average	1.98	2.52	4.14	4.38
Normalized Average	0.40	0.50	0.83	0.88
	<i>Information</i>			
	Precision	Satisfaction	Precision	Satisfaction
	2.2	2.3	3.6	4.5
	1.0	1.0	4.9	4.9
	2.0	1.0	3.0	4.0
	1.0	1.0	4.8	4.8
	2.3	1.3	3.5	4.0
Average	1.70	1.32	3.96	4.44
Normalized Average	0.34	0.26	0.79	0.89

in both criteria as we expect. ‘Precision and satisfaction scores’ on function is higher than ‘precision and satisfaction scores’ on retrieved information in OntoFrame 2007 because the users would not recognize the fact that the function needs to be improved, but are discontented with incomplete information.

‘Entity-centric unified search’ has higher gap in reliability than ‘auto-complete’ between the two versions. It would come from relative importance in search service because the former is one of crucial features affecting the quality of search service whereas the latter is a complementary feature for guiding user’s keywords to proper search keywords. This is the reason why service developers should give priority to relatively important features in service improvement.

4 Conclusions

This study comparatively evaluated reliabilities of two functions in semantic services for the first time. Both ‘auto-complete’ and ‘entity-centric unified search’ functions are operated semantically based on named entities under the URI (Uniform Resource Identifier) scheme. We divided the evaluation into two criteria: the one is ‘precision and satisfaction scores’ on the functions and retrieved information that the user rates, and the other is ‘precision in task performance’ as a quantitative analysis and reliability of expectation-result as a qualitative analysis that the observer assesses. Measured precision scores are very similar with satisfaction scores on both functions and retrieved information. It indicates that precision is closely related with satisfaction in the view point of the user, and further reliabilities can not be taken off various factors for evaluating software quality.

Future work will focus on measuring relative importance among search functions and giving investment priority to them. The relationship between reliability and other usability factors will be also studied.

References

1. ISO 9126-1, Software Engineering, Product Quality (2001)
2. ISO 15489-2, Information and Documentation – Records Management (2001)
3. Sung, W., Jung, H., Kim, P., Kang, I., Lee, S., Lee, M., Park, D., Hahn, S.: A Semantic Portal for Researchers Using OntoFrame. In: ASWC 2007 and ISWC 2007 (2007)
4. Jung, H., Lee, M., Sung, W.: Lessons Learned from Observing User Behavior through Repeated Usability Evaluations. *J. Computer Science and Engineering* 1(3) (2007)
5. Jung, H., Lee, M., Sung, W., You, B.: Auto-complete for Improving Reliability on Semantic Web Service Framework. In: 13th International Conference on Human-Computer Interaction (2009)
6. Kieras, D.: A Guide to GOMS Task Analysis. Technical report, Univ. of Michigan (1994)
7. Jung, H., Lee, M., Kim, P., Lee, S., Sung, W.: Implementation of a Semantic Service Framework with Three Features for Improving Reliability. In: 1st Workshop on Human Factors and the Semantic Web in ASWC 2008 (2009)

Integrated Recommender Systems Based on Ontology and Usage Mining

Liang Wei and Song Lei

School of Management, University of Jinan, Jinan 250022, China
{sm_liangw, sm_songl}@ujn.edu.cn

Abstract. Lots of researches show that ontology as background knowledge can improve document clustering quality with its concept hierarchy knowledge. Meanwhile Web Usage Mining plays an important role in recommender systems and web personalization. However, not many studies have been focused on how to combine the two methods for recommender systems. In this paper, we propose a hybrid recommender system based on ontology and Web Usage Mining. The first step of the approach is extracting features from web documents and constructing relevant concepts. Then build ontology for the web site use the concepts and significant terms extracted from documents. According to the semantic similarity of web documents to cluster them into different semantic themes, the different themes imply different preferences. The hybrid approach integrates semantic knowledge into Web Usage Mining and personalization processes. The experimental results show that the combination of the two approaches can improve the precision rate, coverage rate and matching rate effectively.

1 Introduction

Recommender systems suggest items to users and used on many web sites to help users find interesting items [18]. It measures the user interest in given items or products to provide personalized recommendations for items that will suit the user's taste. More broadly, recommender systems attempt to profile user preferences and model the interaction between users and products [1].

The growing popularity of e-commerce brings an increasing interest in recommender systems. While users browse a web site, well calculated recommendations expose them to interesting products or services that they may consume. The huge economic potential led some of the biggest e-commerce web, for example web merchant Amazon.com and the online movie rental company Netflix, and make the recommender system a salient part of their web sites. High quality personalized recommendations add another dimension to the user experience [2].

There are many different approaches to creating recommender systems [7], but the most common systems fall into three broad classes: collaborative-filtering systems [4], content-based systems [17] and Web Usage Mining approach. Content-based systems make recommendations based on an item similarity measure, based on item features. On the other hand, Collaborative-filtering systems use

patterns in user ratings to make recommendations. Both types of recommender systems require significant data resources [19]. Web Usage Mining approach consists of three phases, namely preprocessing, sequential pattern discovery and sequential pattern analysis. The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By analyzing sequential patterns we can predict future users' visit patterns and recommend web pages to users [21].

1.1 Approach

In this paper, we propose a hybrid recommender system based on ontology and Web Usage Mining. The first step of the approach is collect and process information from web site. The textual features and keywords of web pages are captured and semantic properties of objects of interest as might be available through these textual features and keywords we call them semantic cues. We make use of these semantic properties to clustering different web pages into different themes. The different themes mean different preferences.

Semantic cues are similar to collaborative cues in that they measure similarity of the user preference model from the active access pattern. We get the current visitor's active access pattern from the web server log files in real time and then measure similarity between it and the theme of the clustering result before. Next, we can get some similar web pages from the semantic clusters and recommend them to the current visitor.

The detail steps as follow:

1. Applying Web Usage Mining algorithm Predictor 1.2 [8] to analyze web log files and getting visitors' frequent patterns.
2. Selecting significant sentences from web documents and partitioning significant sentences into terms and feature vectors (Extracting features from web documents).
3. Clustering the features into concepts.
4. Constructing website's ontology based on feature vectors and concepts.
5. Clustering web documents based on ontology.
6. Getting the recommend set of web documents by integrating 5 and 1.

Fig. 1 shows a framework for a hybrid recommender system we discussed in this paper.

1.2 Organization

The next section provides a detailed analysis of how to cluster web documents based on ontology. Section 3 analyses the document representation. In section 4, we provide a detail analysis of how to construct a single website's ontology. Section 5 analyses the application of Web Usage Mining in recommender systems and we propose a recommender algorithm Predictor1.4 and provide a detailed

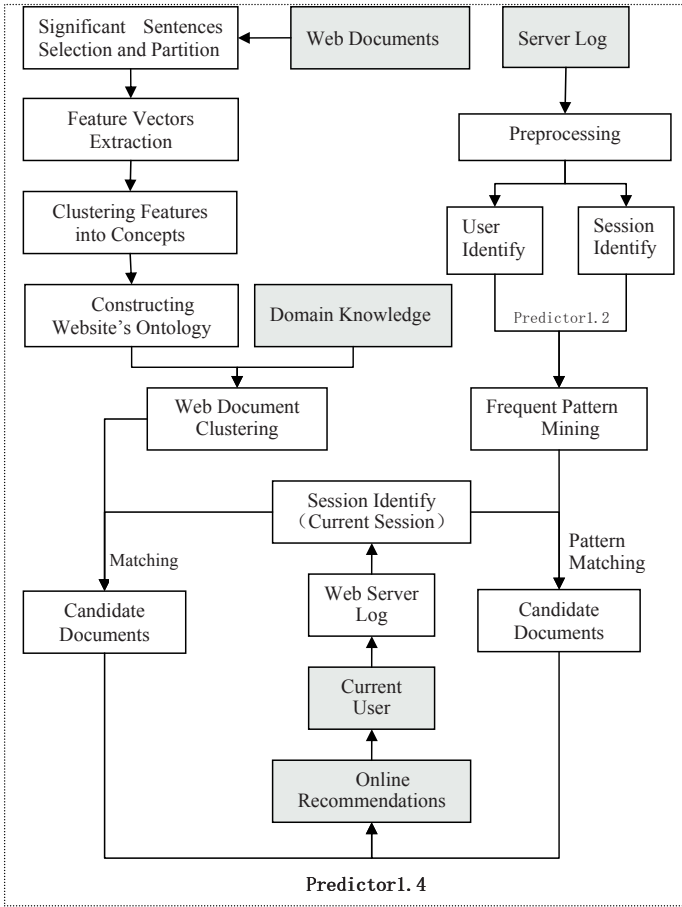


Fig. 1. A general framework for hybrid recommender system

analysis of how to combining ontology-based web clustering into Web Usage Mining. Section 6 describes our experimental work. It provides details of evaluation metrics, methodology and results of different experiments and discussion of the results. The final section provides some concluding remarks and directions for future research.

2 Web Documents Clustering

Ontology based similarity measure has some advantages over other measures. First, ontology is created by human being manually for a domain and thus more precise; second, compared to other methods such as latent semantic indexing, it's much more computationally efficient; Third, it helps integrate domain knowledge into the data mining process. Comparing two terms in a document using ontology

information usually exploit the fact that their corresponding concepts within ontology usually have properties in the form of attributes, level of generality or specificity, and their relationships with other concepts [10], [23]. It should be noted that there are many other term semantic similarity measures such as latent semantic indexing, but it's out of scope of our research, our focus here is on term semantic similarity measure using ontology information.

2.1 Feature Based Measure

Feature based measure assumes that each term is described by a set of terms indicating its properties or features. Then, the more common characteristics two terms have and the less non-common characteristics they have, the more similar the terms are [22]. In our experimental study, we take all the ancestor nodes of each compared concept as their feature sets. The following measure is defined according to [7], [9]:

$$S_{BasicFeature}(C_1, C_2) = \frac{|Ans(C_1) \cap Ans(C_2)|}{|Ans(C_1) \cup Ans(C_2)|} \quad (1)$$

where $AnsC_1$ and $AnsC_2$ correspond to description sets (the ancestor nodes) of terms C_1 and C_2 respectively, $C_1 \cap C_2$ is the join of two parent node sets and $C_1 \cup C_2$ is the union of two parent node sets. Knappe [5] defines a similarity measure as below using the information of generalization and specification of two compared concepts:

$$S_{Knappe}(C_1, C_2) = p \times \frac{|Ans(C_1) \cap Ans(C_2)|}{|Ans(C_1)|} + (1 - p) \times \frac{|Ans(C_1) \cap Ans(C_2)|}{|Ans(C_2)|} \quad (2)$$

where p 's range is $[0,1]$ that defines the relative importance of generalization vs. specialization. This measure scores between 1 (for similar concepts) and 0. In our experiment, p is set to 0.5.

In this paper, we adopt feature based measure to cluster web documents. Before clustering, we need to solve the problem of web documents' representation first.

3 Document Representation

For the clustering experiments described subsequently, we prepare different representations of web documents suitable for the clustering algorithms.

Let us first consider documents to be bags of terms [17]. Let $tf(d, t)$ be the absolute frequency of term $t \in T$ in document $d \in D$, where D is the set of documents and $T = \{t_1, t_2, \dots, t_m\}$ is the set all different terms occurring in D . We denote the term vectors $td = (tf(d, t_1), \dots, tf(d, t_m))$. Later on, we will need the notion of the centroid of a set X of term vectors. It is defined by:

$$t_X := \frac{1}{|X|} \sum_{t_d \in X} t_d \quad (3)$$

In the sequel, we will apply tf also on sets of terms: for $T' \subseteq T$, we let:

$$tf(d, T') := \sum_{t \in T'} tf(d, t) \quad (4)$$

As initial approach we have produced this standard representation of the texts by term vectors. The initial term vectors are further modified as follows. Stop-words are words which are considered as non-descriptive within a bag-of-words approach. Following common practice, we removed stopwords from T .

We have processed our web page using the Porter stemmer introduced in [16]. We used the stemmed terms to construct a vector representation td for each web page. Then, we have investigated how pruning rare terms affects results.

Depending on a pre-defined threshold δ , a term t is discarded from the representation (i. e., from the set T). We have used the values 0, 5 and 30 for δ . The rationale behind pruning is that infrequent terms do not help for identifying appropriate clusters, but may still add noise to the distance measures degrading overall performance [17].

$tfidf$ weighs the frequency of a term in a document with a factor that discounts its importance when it appears in almost all documents. The $tfidf$ of term t in document d is defined by:

$$tfidf(d, t) := \log(tf(d, t) + 1) * \log\left(\frac{|D|}{df(t)}\right) \quad (5)$$

Where $df(t)$ is the document frequency of term t that counts in how many documents term t appears. If $tfidf$ weighting is applied then we replace the term vectors $td = (tf(d, t_1), \dots, tf(d, t_m))$ by $td = (tfidf(d, t_1), \dots, tfidf(d, t_m))$.

Based on the initial web page representation, we have first applied stopword removal. Then we performed stemming, pruning and $tfidf$ weighting in all different combinations. This also holds for the initial document representation involving background knowledge described subsequently. When stemming and/or pruning and/or $tfidf$ weighting was performed, we have always performed them in the order in which they have been listed here.

4 The Integration of Semantic Knowledge and Web Document Clustering

The semantic knowledge we have exploited is given through a simple ontology. We first describe its structure and then provide the actual ontology and its integration into the initial web document representation.

4.1 Ontology

The background knowledge we will exploit further on is encoded in a core ontology. We here present those parts of our wider ontology definition [3] that we have exploited:

Definition: A core ontology is a tuple $O := (C, \leq c)$ consisting of a set C whose elements are called concept identifiers, and a partial order $\leq c$ on C , called concept hierarchy or taxonomy. Often we will call concept identifiers just concepts, for sake of simplicity.

Definition: If $c_1 < c c_2$, for $c_1, c_2 \in C$, then c_1 is a sub concept of c_2 , and c_2 is a super concept of c_1 . If $c_1 < c c_2$ and there is no $c_3 \in C$ with $c_1 < c c_3 < c c_2$, then c_1 is a direct sub concept of c_2 , and c_2 is a direct super concept of c_1 . We note this by $c_1 \prec c_2$.

4.2 Concepts Vector

Enriching the term vectors with concepts from the core ontology has two benefits. First it resolves synonyms; and second it introduces more general concepts which help identifying related topics. For instance, a document about 'Michael Bay' may not be related to a document about 'George Lucas' by the cluster algorithm if there are only 'Michael Bay' and 'George Lucas' in the term vector. But if the more general concept 'Sci-Fi & Fantasy Director' is added to both documents, their semantical relationship is revealed. Because they all have directed Sci-Fi & Fantasy movies.

So how to add or replace terms by concepts as follows:

When applying this strategy, we have extended each term vector td by new entries for Wordnet concepts c appearing in the web document sets. Thus, the vector td was replaced by the concatenation of td and cd , where $cd := (cf(d, c_1), \dots, cf(d, c_n))$ is the concept vector with $n = |C|$ and $cf(d, c)$ denotes the frequency that a concept $c \in C$ appears in a document d . Hence, a term that also appears in Wordnet as a synset would be accounted for at least twice in the new vector representation, i.e., once as part of the old td and at least once as part of cd . It could be accounted for also more often, because a term like "Director" has several corresponding concepts in WorldNet. Thus, terms that appear in WorldNet are only accounted at the concept level, but terms that do not appear in WorldNet are not discarded.

Finally we construct an experimental ontology for a movie website showed by Fig. 2.

4.3 Web Documents Clustering Based on Feature

The approach we used to cluster web documents has already introduced in section 2 and the web documents' presentation has also presented in section 3. In order to combine the semantic knowledge to document clustering we adopt the clustering approach that based on feature. To construct concepts, we extract features from each web document, and cluster them into concepts that can be used to form website's ontology. Then every web document can be denoted by several concepts.

Nakata et al. [14] introduced a notion of Concept Index, which aims to index important concepts described in a collection of documents belonging to a group,

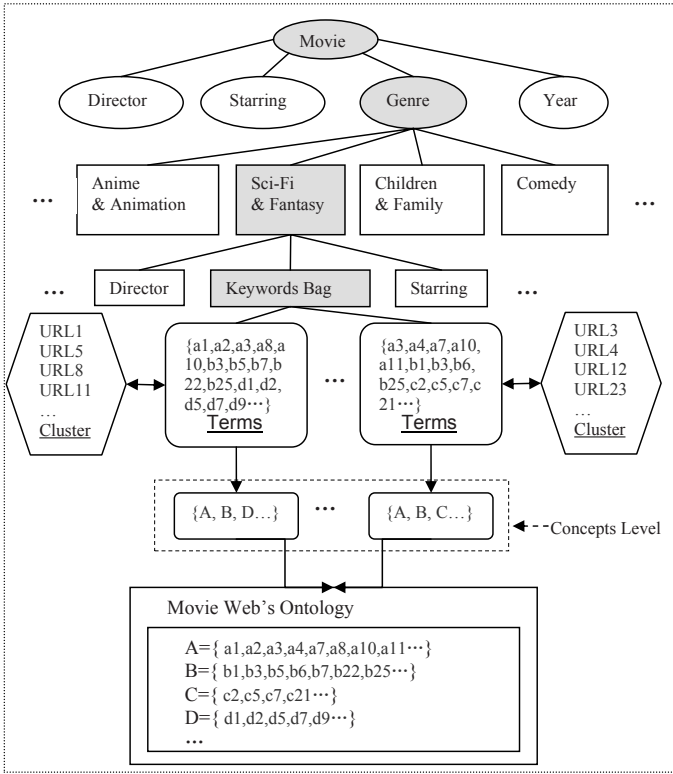


Fig. 2. The concept mapping from terms to ontology for a movie web site

and provide user-friendly cross-references among them to aid concept-oriented document space navigation. The Concept Index relied on users to identify important concepts by marking keywords and phrases that interest them. Nakata’s work addressed a group of individuals who shared the same interest or a task and would profit from making use of the knowledge possessed by the group. However this approach is different from ours since they use collaboration between the members of group for extracting concept. In our approach, we try to automatically construct concepts from a document space.

Suppose that we represent a sentence as a vector of terms in the sentence with their *tfidf* weight values. We can consider a set of vectors $S = \{s_1, s_2, \dots, s_m\}$, where m is the number of selected significant sentences for a document. To make feature vectors for the document, we partition S as follows. Let us consider each vector s_i as a subgraph such that the vertices of the subgraph are terms of the vector and they are connected. We then consider a feature vector as a connected component of the graph consisting of subgraphs $\{s_1, s_2, \dots, s_m\}$. Every significant sentence is represented as a set of terms and its *tfidf* weight value pairs.

Then we use the approach presented in section 2.3 to cluster documents. So far we accomplish the document clustering.

5 Web Usage Mining

Web Usage Mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context.

In Web usage mining, the primary Web resource that is being mined is a record of the requests made by visitors to a Web site, most often collected in a Web server log [20]. The content and structure of Web pages, and in particular those of one Web site, reflect the intentions of the authors and designers of the pages and the underlying information architecture. The actual behavior of the users of these resources may reveal additional structure.

Relationships may be induced by usage where a different relationship was intended. For example, sequence mining may show that many of the users who visited page C later went to page D , along paths that indicate a prolonged search (frequent visits to help and index pages, frequent backtracking, etc.) [21]. This can be interpreted to mean that visitors wish to reach D from C , but that this was not foreseen in the information architecture, hence that there is at present no hyperlink from C to D . This insight can be used for static site improvement for all users.

It is useful to combine Web usage mining with content and structure analysis in order to "make sense" of observed frequent paths and the pages on these paths. This can be done using a variety of methods.

Predictor1.4 is a hybrid algorithm based on semantic knowledge and Web Usage Mining. The algorithm as follows:

Input:

1. The pages of recent visited $[p_1, p_2, \dots, p_k]$, k is changeable, through experiment we take $k = 3$.
2. The output of Predictor1.2 [8], such as $P = [p_1, p_2, \dots, p_n]$.

Output: The final recommend results $P' = [p_1, p_2, \dots, p_m]$, $m \leq n$.

```

for (i=1; i<=k; i++)
{
    p=traverse(P);
    // traverse the P=[p1,p2,,pn];
    topic=search1(ontology,p);
    //search the subjects that p belongs to in the ontology;
    for (j=1; j<=a; j++)
        // a is the number of subjects that p belongs to;
        {
            P_candidate=search2(ontology, subject);
            //return the most N pages that similar to p from the concept;
            P_candidate_a= P_candidate_a+P_candidate;
        }
    P_candidate_all= P_candidate_all+ P_candidate_a;
}
P'=Cross1(P_candidate_all)+Cross2(P_candidate_all,P)

```

6 Evaluation Measures and Results

6.1 Evaluation Measures

There are two important evaluation measures: *Coverage* and *Precision* [12]. If we take RS as the recommend set of pages and take US as a set of pages that visitors accessed, then we can provide the definition for *Coverage* and *Precision*:

$$\text{Coverage} = \frac{|US \cap RS|}{|US|} \quad (6)$$

$$\text{Precision} = \frac{|US \cap RS|}{RS} \quad (7)$$

Meanwhile in order to get the best performance we need a measure to evaluate. Mobasher etc. provided a new evaluation measures M [11]. M is called matching rate.

$$M = \frac{2 \times \text{coverage} \times \text{precision}}{\text{coverage} + \text{precision}} \quad (8)$$

6.2 Results

Our experimental dataset comes from a single movie rental website. As of March 2008, the website contained more than 85,000 web pages and more than 5GB web server log files. There are some pages irrelative to our experiment. So after pre-processing there are still 12,530 web pages.

In order to evaluate the performance of Predictor 1.4, we compared the results that come from Predictor 1.4 with Predictor 1.2 and web documents clustering approach.

Table 1. Result from Predictor1.4 with different number of RS

RS	Precision	Coverage	M
6	0.831	0.521	0.64
9	0.805	0.582	0.676
12	0.782	0.658	0.715
15	0.769	0.735	0.752
Average	0.797	0.624	0.696

From Table 1 to Table 3 and Fig. 3 we can see that Predictor1.4 is better than Predictor1.2 and WDC in *Precision*, *Coverage* and M . The reasons are as the follow:

1. The Predictor1.2 is the traditional Web Usage Mining approach that is not considering the semantic knowledge between web pages.
2. The WDC approach based on documents' feature and content.

- Predictor1.4 is a hybrid recommender algorithm based on ontology and Web Usage Mining. This approach get the recommend set of web documents by integrating 1 with 2.

The experiment results show that the approach that integrates ontology with Web Usage Mining can improve the performance. It also show that the number of *RS* influence the *Precision*, *Coverage* and *M*. If we recommend 15 web pages to visitor we can get higher Coverage, but Precision is lower than the results that adopt other number of *RS*. So we must find a balance point. We recommend the number of *RS* is 9. If we recommend too many pages to visitor, it will confuse the visitor again. That is against the primary purpose of recommender system. We will introduce this in other paper.

Table 2. Result from Predictor1.2 with different number of RS

RS	Precision	Coverage	M
6	0.762	0.451	0.567
9	0.693	0.537	0.605
12	0.585	0.573	0.579
15	0.476	0.626	0.541
Average	0.629	0.547	0.573

Table 3. Result from WDC (web document clustering singly)

RS	Precision	Coverage	M
6	0.793	0.486	0.603
9	0.723	0.536	0.616
12	0.689	0.617	0.651
15	0.605	0.663	0.633
Average	0.703	0.576	0.625

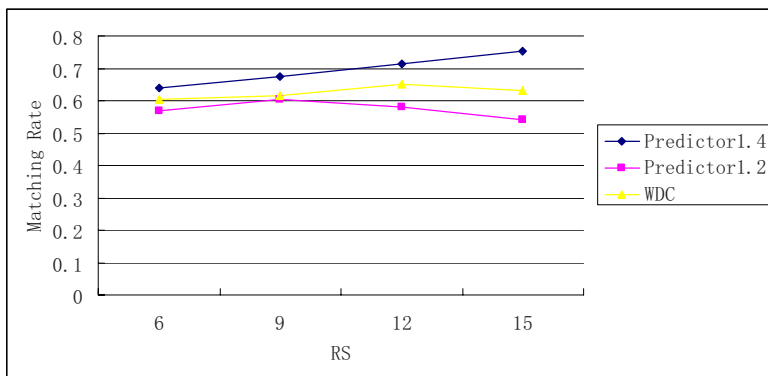


Fig. 3. Performance of three methods

7 Conclusions and Future Directions

Overall, our preliminary experiments are encouraging and suggest that integrate document clustering and semantic knowledge with Web Usage Mining can indeed be useful in recommender system. We believe that the successful integration of semantic knowledge with Web usage mining is likely to lead to the next generation of personalization tools which are more intelligent and more useful for Web users.

References

1. Adomavicius, G., Tuzhilin, A.: Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 634–749 (2005)
2. Bell, R., Koren, Y., Volinsky, C.: Modeling relationships at multiple scales to improve accuracy of large recommender systems. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 95–104 (2007)
3. Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S., Stojanovic, L., Stojanovic, N., Studer, R., Stumme, G., Sure, Y., Tane, J., Volz, R., Zacharias, V.: KAON - towards a large scale semantic web. In: *Bauknecht, K., Tjoa, A.M., Quirchmayr, G. (eds.) EC-Web 2002. LNCS*, vol. 2455, pp. 304–313. Springer, Heidelberg (2002)
4. Breese, J., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: *UAI 1998*, pp. 43–52 (1998)
5. Cooley, R.: *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, Faculty of the Graduate School (2000)
6. Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. In: *Proc. of the SIGIR 2003 Semantic Web Workshop* (2003)
7. Knappe, R., Bulskov, H., Andreasen, T.: Perspectives on Ontology-based Querying. *International Journal of Intelligent Systems* (2004)
8. Liang, W., Zhang, H.Y.: Study of Recommendation Models in E-Commerce Recommendation Systems. *Computer Engineering and Applications* 42(36), 183–186 (2006)
9. Lin, D.: Principle-Based Parsing Without Overgeneration. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 1993)*, Columbus, Ohio, pp. 112–120 (1993)
10. Mao, W., Chu, W.W.: Free text medical document retrieval via phrased-based vector space model. In: *Proc. of AMIA 2002*, San Antonio, TX (2002)
11. Mobasher, B., Dai, H.: Improving the effectiveness of collaborative filtering on anonymous web usage data. In: *Proc of the IJCAI2001 Workshop on Intelligent Techniques for Web personalization*, pp. 49–56. Springer, Berlin (2001)
12. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Discovery of aggregate usage profiles for Web personalization. In: *Proc. of the WebKDD 2000 Workshop at the ACM SIGKDD 2000*, Boston, pp. 142–151 (2000)
13. Montaner, M., Lopez, B., De La Rosa, J.L.: A taxonomy of recommender agents on theinternet. *Artificial Intelligence Review* 19(4), 285–330 (2003)

14. Nakata, K., Voss, A., Juhnke, M., Kreifelts: Collaborative concept extraction from documents. In: Proceedings of the 2nd Int. Conf. on Practical Aspects of Knowledge management (PAKM 1998), Basel, Switzerland, pp. 29–30 (1998)
15. Pazzani, M., Billsus, D.: Content-based recommendation systems. *The Adaptive Web* 5, 325–341 (2007)
16. Porter, M.F.: An algorithm for suffix stripping. *Program* 14(3), 130–137 (1980)
17. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
18. Schafer, J.B., Konstan, J.A., Riedl, J.: Recommender systems in e-commerce. In: ACM Conference on Electronic Commerce, pp. 158–166 (1999)
19. Shani, G., Chickering, M., Meek, C.: Mining recommendations from the web. In: Proceedings of the 2008 ACM conference on Recommender systems, Lausanne, Switzerland, pp. 35–42 (2008)
20. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: discovery and application of usage patterns from web data. *SIGKDD Explorations* 1(2), 12–23 (2000)
21. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations* 1(2), 12–23 (2000)
22. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G., Milios, E.E.: Semantic similarity methods in wordNet and their application to information retrieval on the web. In: WIDM 2005, pp. 10–16. ACM Press, New York (2005)
23. Zhang, X., Jing, L., Hu, H., et al.: A comparative study of ontology based term similarity measures on pubmed document clustering. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 115–126. Springer, Heidelberg (2007)

Knowledge-Based Concept Score Fusion for Multimedia Retrieval

Manolis Falelakis, Lazaros Karydas, and Anastasios Delopoulos

Multimedia Understanding Group
Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki, Greece
manf@mug.ee.auth.gr, kary@mug.ee.auth.gr, adelo@eng.auth.gr

Abstract. Automated detection of semantic concepts in multimedia documents has been attracting intensive research efforts over the last years. These efforts can be generally classified in two categories of methodologies: the ones that attempt to solve the problem using discriminative methods (classifiers) and those that build knowledge-based models, as driven by the W3C consortium. This paper proposes a methodology that tries to combine both approaches for multimedia retrieval. Our main contribution is the adoption of a formal model for defining concepts using logic and the incorporation of the output of concept classifiers to the computation of annotation scores. Our method does not require the computationally intensive training of new classifiers for the concepts defined. Instead, it employs a knowledge-based mechanism to combine the output score of existing classifiers and can be used for either detecting new concepts or enhancing the accuracy of existing detectors. Optimization procedures are employed to adapt the concept definitions to the multimedia corpus in hand, further improving the attained accuracy. Experiments using the TRECVID2005 video collection demonstrate promising results.

1 Introduction

The exponential growth of multimedia content during the last decade has made efficient indexing and retrieval a necessity. The appearance of evaluation frameworks such as TRECVID [1] and baseline frameworks such as Mediamill [2] and Columbia374 [3] reflects this trend.

Much of the research effort towards this direction is governed by the development of discriminative concept classifiers, often yielding satisfactory results. These can be further improved by using classifier score fusion [4] with concepts either selected manually [5] or determined automatically [6,7].

On the other hand lie methods relying on knowledge. These are mainly based on inference using expressive Description Logics [8]. Extensions of these methods that model certainty using Fuzzy Description Logics [9] have recently been successfully employed for multimedia retrieval [10,11]. These, however, have certain restrictions, imposed mainly by the computational cost of reasoning which can become prohibitive when dealing with large concept collections.

Present paper proposes a methodology that aims at combining virtues from both aforementioned approaches. The output scores of concept classifiers are fuzzified and used by formal fuzzy knowledge models to detect semantic concepts in multimedia. This way, the computationally intensive training of new classifiers is unnecessary. Furthermore, based on a notion that (existential or universal) quantifiers are of no use in this particular case, the knowledge models adopted are very simple, minimizing the complexity of reasoning.

The target of our methodology is two-fold, aiming at (i) providing an inexpensive yet reliable means of extending existing concept detector schemes and (ii) enhancing the detection accuracy of concepts for which specialized classifiers already exist. A virtue of this approach is that it does not dictate the type of features used. In fact, concept classifiers used may be trained using completely different feature vectors.

Moreover, our method is coupled with optimization schemes that help to adapt the fuzzy definitions to the dataset in-hand. To this end, we use a genetic algorithm which is accompanied with of k -fold cross validation [12] and RankBoost [13] re-sampling algorithms, in order to avoid over-fitting on the training sets and provide a more modest estimation on the behavior when applied to the test set.

We must point out that our method does not necessarily provide better results than concept classifiers in all cases. However, exploitation of the estimate of its performance can help us determine when its use is of potential benefit for classifier enhancement.

Experiments conducted using the LSCOM concept ontology on a TRECVID dataset demonstrate the effectiveness of the proposed approach on real data. Results show that new concepts can be efficiently defined, often attaining performance comparable to specifically trained concept classifiers (but with minimal computational effort) while it can provide significant improvement to detection of corresponding concepts for which classifiers exist.

The next section is devoted to describing the Fuzzy Models, while section 3 describes the fuzzy degrees adaptation procedure. Section 4 presents experiments conducted on real data and section 5 concludes the paper and includes some future directions.

2 Fuzzy Definition Models

The main idea of our approach is to rely on the result of reliable concept-classifiers to infer on other concepts.

Concept classifiers treat an image as a whole and provide information whether it (up to a certain degree) belongs to a certain class or not. However, they provide no evidence on the existence and the type of possible interrelations between the detected concepts. Due to this reason, it makes no sense to model these relations using object properties ('roles' in the Description Logics terminology).

Based on the previous notions we adopt a language which can contain statements based on conjunction, disjunction and negation operators, i.e., disregarding quantifiers and other expressivity tools provided by Description Logics. More formally, the expressions are constructed according to the following syntax rule:

C, D	\longrightarrow	A		(atomic concept)
		\top		(universal concept)
		\perp		(bottom concept)
		$\neg C$		(negation)
		$C \sqcup D$		(union)
		$C \sqcap D$		(intersection)

Furthermore, we allow subsumptions to hold up to a certain degree, i.e., we model uncertainty in a way similar to the one of [9]. In this direction, a concept S_i is subsumed by a concept C_i to the degree f_i , as displayed in equation 1

$$\langle S_i \sqsubseteq C_i, f_i \rangle \tag{1}$$

Let a hierarchy \mathcal{T} of such subsumptions, according to which, concept C subsumes concepts $S_1 \dots S_k$, i.e.,

$$\mathcal{T} = \left\{ \begin{array}{l} \langle S_1 \sqsubseteq C, f_1 \rangle, \\ \langle S_2 \sqsubseteq C, f_2 \rangle, \\ \dots, \\ \langle S_k \sqsubseteq C, f_k \rangle \end{array} \right. \tag{2}$$

Inference on the degree of the existence $\mu(C)$ of concept C , based on the existence of concepts S_i as given by the fuzzified classifier output $\mu_c(S_i)$, is made according to the *type 1* definition which is of the following form

$$\mu(C) = \mathcal{U}_i(\mathcal{I}(\mu_c(S_i), f_i)) \tag{3}$$

where the operators \mathcal{U} and \mathcal{I} denote fuzzy union and intersection operators respectively.

The existence of the other concepts of \mathcal{T} is computed with definitions of *type 2* that take the following form

$$\mu(S_i) = \mathcal{I}(\mu_c(C), \mathcal{I}_{j \neq i}(\mathcal{N}(\mathcal{I}(\mu_c(S_j), f_j)))) \tag{4}$$

where the operator \mathcal{N} denotes fuzzy complement (negation).

To illustrate these with an example, consider the hierarchy depicted in figure 1 that can be encoded as

$$\mathcal{T} = \left\{ \begin{array}{l} \langle Car \sqsubseteq Vehicle, f_{Car} \rangle, \\ \langle Bus \sqsubseteq Vehicle, f_{Bus} \rangle, \\ \langle Motorcycle \sqsubseteq Vehicle, f_{Motor} \rangle \end{array} \right. \tag{5}$$

Definitions of type 1, computed with equation 3 suggest that we compute the degree of existence of 'Vehicle' as the logical union of the degrees of existence of 'Car', 'Bus' and 'Motorcycle', meaning that a 'Vehicle' is a 'Car' or a 'Bus' or a 'Motorcycle'.

¹ The equation can be written this way (with union taking multiple inputs) due to the associativity and commutativity properties of fuzzy norms.

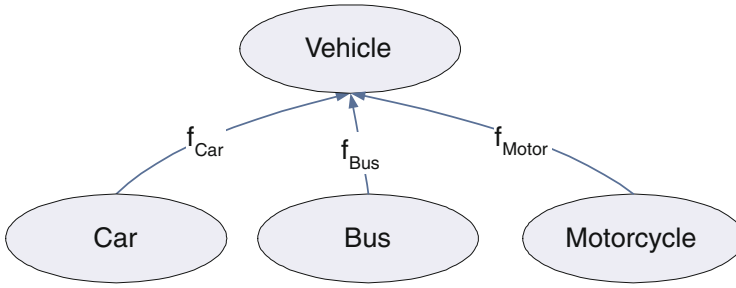


Fig. 1. An example of a simple hierarchy

On the other hand, with a type 2 definition, we can compute the existence of a 'Car', defining it as a 'Vehicle' and not a 'Bus' and not a 'Motorcycle', when scores for the latter concepts exist.

Note that by forming definitions of type 2, we make a disjointness assumption, i.e., all hierarchy siblings are assumed to be disjoint. This could not always be the case. For instance an image may contain two sibling concepts (such as a 'Car' and a 'Motorcycle' in our example) at the same time. However, this assumption leads us to easy definition extraction and proves to work in practice as shown in the experiments of section 4.

As stated before, our approach can also become useful when the classifier score for the concept under examination is available and the goal is to improve the retrieval performance. In this case, in order to compute $\mu(C)$ given the fuzzified output $\mu_c(C)$ of the corresponding classifier, we use $\mu_c(C)$ and eq. 3 in a disjunctive manner, and a type 1 definition takes the following form

$$\mu(C) = \mathcal{U}(\mathcal{I}(\mu_c(C), f_C), \mathcal{U}_i(\mathcal{I}(\mu_c(S_i), f_i))) \quad (6)$$

In a similar manner derives inference for concepts of type 2:

$$\mu(S_i) = \mathcal{U}(\mathcal{I}(\mu_c(S_i), f_i), \mathcal{I}(\mu_c(C), \mathcal{I}_{j \neq i}(\mathcal{N}(\mathcal{I}(\mu_c(S_j), f_j)))))) \quad (7)$$

Subsumptions in the form of Eq. 1 can be extracted from crisp domain ontologies such as LSCOM [14], and use optimization techniques to compute the degrees f_i , fuzzifying the hierarchy (i.e., making subsumptions hold up to a certain degree) and adapting to the dataset under examination.

3 Adaptation to the Corpus In-Hand

In order to fuzzify the knowledge source we have to compute the weights (i.e. the values for f_i) presented in the previous section. This is essentially an optimization problem.

The target is to compute f_i with respect to the current dataset with the goal of maximizing the average precision for each of the defined concepts. For the

parametric norms (we have experimented with Dubois-Prade and Yager class) the value of the norm parameter is also determined by training.

As the optimization surface proves to contain local minima we employ a genetic algorithm for this task. The fitness function to be minimized is the arithmetic negation of the average precision.

In order to improve the generalization properties of the computed weights we use two resampling methods; k -fold cross validation and a modified version of RankBoost [13].

3.1 k -Fold Cross Validation

In k -fold cross validation the original training samples are firstly partitioned into k subsets, called folds. Due to the nature of the video dataset (scarce positive samples tend to appear in bursts) the partitioning of the training set into folds is made by evenly distributing the positive samples in every fold. This assures that all k folds are representative of the whole set, with respect to the prior probability of positives.

Then the genetic algorithm is called to train the fuzzy weights using as training set all the samples in $k-1$ folds, leaving one fold out, which is used for validation. This procedure is repeated k times until each fold has been used for training $k-1$ times and for validation exactly once.

Finally, we end up with k sets of fuzzy weights which are averaged to obtain a single set. These are the values for f_i that are then incorporated to the inference function.

The results obtained by this method in the training set are used as a modest estimate of the expected average precision of a concept in the test set, therefore providing a way to determine the potential improvement of performance in classifier enhancement. Since the method is not so prone to over-fitting on the training set the selection of the potentially improved classifiers can be made by setting a performance threshold.

3.2 RankBoost

Boosting is a technique used in machine learning that tries to create a set of weak learners and combine them into a single strong learner.

The optimization scheme used here is a modified version of the RankBoost algorithm (see [13]). The algorithm runs in T rounds, randomly choosing all the positive and an approximately equal number of negative samples to form a new training set for the next round. A distribution function is initialized for the purpose of determining which samples are going to be used for training in each round. This distribution function is updated in each round, so as to promote the selection of samples that were misclassified in the previous rounds.

In each round, the genetic algorithm is called to train the weights with respect to the subset of the samples that have been chosen as training set. This means that the genetic algorithm training procedure will rerun from scratch exactly T times.

Finally, RankBoost computes T sets of fuzzy weights and inference is performed using a combination of them, which is a weighted average with the accuracy obtained in each round.

4 Experiments

For the purpose of our experiments we have used the Columbia374 [3] set of semantic concept detectors, which also includes the ground truth, the features, and the results of the detectors over the TRECVID datasets.

Our dataset consisted of the 47 videos of the TRECVID2005 development set that were not used for training the Columbia classifiers. These videos (corresponding to 20054 shots) were split to a training (23 videos) and an evaluation (24 videos) set, each one containing about 10000 shots.

In order to form the definitions automatically, we used a cut-down version of the LSCOM, that includes the 374 concepts of our dataset and exploited its hierarchy. We have conducted two experiments for evaluating our method. In the first experiment we demonstrate how new concepts can be defined, in the presence of no adequate classifier, while in the second one we perform a type of query expansion in order to improve the accuracy of concept detection.

4.1 Concept Scalability

This experiment simulates the case of extending the vocabulary of a multimedia retrieval system using knowledge. This is fully scalable and extensible as new concepts can be defined recursively, while training requires minimal effort compared to building a classifier model for every new concept.

The definitions used for this experiment are of the form displayed in equations [3] and [4].

We have chosen to define concepts, already existing in the ontology but without taking into account the actual classifier scores for them during inference. Instead, we use these scores as a baseline for comparison purposes.

Figure [2] displays the attained average precision for several concepts, using this kind of definitions. The concepts here were selected based on a certain threshold imposed on their performance on the training set when using the cross validation method. This gave us a hint of their performance on the evaluation set.

Commenting on figure [2], our methodology seems to yield very satisfactory results, often comparable to the ones of specifically trained classifiers. In some cases (see 'Road Overpass' for example), it outperforms the corresponding classifier. This is very important considering the computational cost of training the latter. Finally, in every case, the use of fuzzy weights, adapted to the set in-hand, significantly improves the performance.

4.2 Classifier Enhancement

In this experiment classifier scores are taken into account and the definitions formed correspond to the ones of the equations [6] and [7].

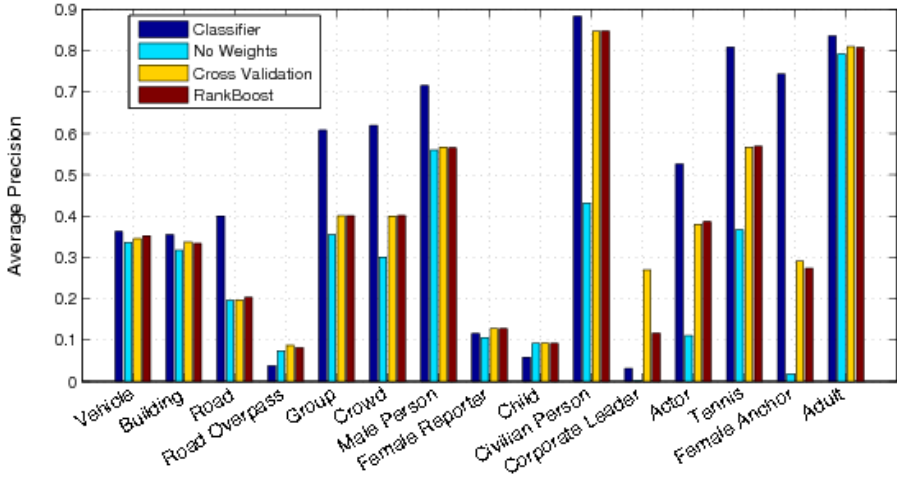


Fig. 2. Concept scalability experiment. Bars display the Average Precision attained by the Columbia classifier, our definition without using fuzzy weights, (i.e., with f_i 's set to 1), and our definition using weights calculated using the two proposed methods respectively.

The goal here is to enhance the retrieval performance of the classifiers, using a kind of knowledge-based query expansion.

The average precision attained in this case for the concepts of figure 2 is illustrated in figure 3.

As it can be seen, the results of our method provide improvement over the ones given by the Columbia classifiers. Once again, proper fuzzy weights seem to increase the performance. Finally, comparing these to the results of section 4.1 confirms our expectation that the use of classifiers, whenever available, is beneficial for our method.

4.3 Comparison of Fuzzy Norms

Finally, the same experiment was carried for multiple fuzzy T-norms coupled by their corresponding, dual in terms of fuzzy complement, T-conorms (see 15 for more on this subject). Table 1 displays the mean average precision in each case.

Some comments are worth to be made here: The pair algebraic product/ sum yields the best results in this dataset, while drastic product/ sum seem to be a completely inadequate choice. The standard (min/max) operators have decent, but far from optimal, performance.

Furthermore, contrary to one might expect, the parametric norms (Dubois-Prade and Yager class) have not performed very well. A potential reason might be that in this case optimization may have failed to train their extra parameter over the dataset.

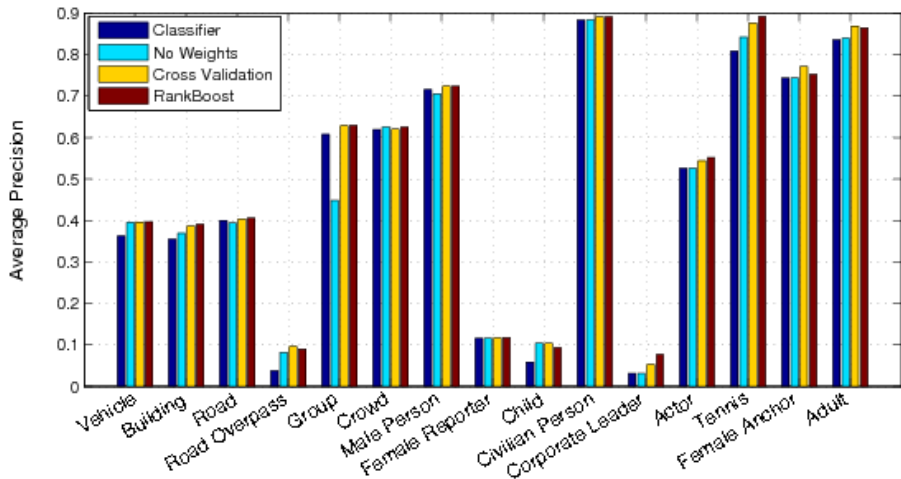


Fig. 3. Classifier enhancement experiment. Bars display the Average Precision attained by the Columbia classifier, our definition without using fuzzy weights, (i.e., with f_i 's set to 1), and our definition using weights calculated using the two proposed methods respectively.

Table 1. Mean Average Precision attained for various pairs of fuzzy norms

fuzzy T-norm	cross validation	RankBoost
Standard (min)	0.3065	0.3013
Algebraic product	0.3220	0.3206
Drastic product	0.0866	0.0871
Bounded difference	0.3132	0.3093
Dubois-Prade	0.3087	0.2889
Yager	0.2756	0.2970

5 Conclusions

We have presented a methodology for constructing fuzzy definitions and employing them for concept detection in multimedia, based on classifier results. The fuzzy weights are efficiently adapted to the corpus available. Our approach is useful both extending a concept collection and improving classifier scores in multimedia retrieval, all with a minimal computational effort. Experiments have shown that the method performs well on real data, often outperforming specifically trained discriminative classifiers.

Further improvement of the definition extraction methodology, fine tuning the optimization procedures as well as experimentation on other datasets are of potential interest for the future.

Acknowledgement

This paper is part of the 03ED853 research project, implemented within the framework of the “Reinforcement Programme of Human Research Manpower” (PENED) and co-financed by National and Community Funds (25% from the Greek Ministry of Development-General Secretariat of Research and Technology and 75% from E.U.-European Social Funding).

References

1. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR 2006: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, pp. 321–330. ACM, New York (2006)
2. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.-M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: MULTIMEDIA 2006: Proceedings of the 14th annual ACM international conference on Multimedia, pp. 421–430. ACM, New York (2006)
3. Yanagawa, A., Chang, S.-F., Kennedy, L., Hsu, W.: Columbia university’s baseline detectors for 374 lscm semantic visual concepts. Technical report, Columbia University ADVENT Technical Report #222-2006-8 (March 2007)
4. Hauptmann, A., Yan, R., Lin, W.-H., Christel, M., Wactlar, H.: Filling the semantic gap in video retrieval: An exploration. *Semantic Multimedia and Ontologies*, 253–278 (2008)
5. Christel, M., Hauptmann, A.: The use and utility of high-level semantic features in video retrieval. *Image and Video Retrieval*, 134–144 (2005)
6. Volkmer, T., Natsev, A.: Exploring automatic query refinement for text-based video retrieval, July 2006, pp. 765–768 (2006)
7. Neo, S.-Y., Zhao, J., Kan, M.-Y., Chua, T.-S.: Video retrieval using high level features: Exploiting query matching and confidence-based weighting, pp. 143–152 (2006)
8. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, Cambridge (2003)
9. Straccia, U.: Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research* (April 14, 2001)
10. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: A fuzzy description logic for multimedia knowledge representation. In: *Proc. of the International Workshop on Multimedia and the Semantic Web* (2005)
11. Athanasiadis, T., Simou, N., Papadopoulos, G., Benmokhtar, R., Chandramouli, K., Tzouvaras, V., Mezaris, V., Phiniketos, M., Avrithis, Y., Kompatsiaris, Y., Huet, B., Izquierdo, E.: Integrating image segmentation and classification for fuzzy knowledge-based multimedia indexing. In: Huet, B., Smeaton, A., Mayer-Patel, K., Avrithis, Y. (eds.) *MMM 2009*. LNCS, vol. 5371, pp. 263–274. Springer, Heidelberg (2009)
12. Mosteller, F.: A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics* 19(1), 58–65 (1948)

13. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.* 4, 933–969 (2003)
14. Naphade, M., Smith, J.R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. *IEEE MultiMedia* 13(3), 86–91 (2006)
15. Klir, G.J., Yuan, B.: *Fuzzy Sets and Fuzzy Logic; Theory and Applications*. Prentice-Hall, Englewood Cliffs (1995)

Example-Based Query Analysis Using Functional Conceptual Graphs^{*}

Hui Liu¹ and Yuquan Chen²

¹ School of Business Information Management, Shanghai Institute of Foreign Trade
liuh@shift.edu.cn

² Department of Computer Science and Engineering, Shanghai Jiao Tong University
yqchen@sjtu.edu.cn

Abstract. In this paper the authors suggest an example-based method to analyze user queries in questions into conceptual graphs. As a novel point, functional conceptual graphs (FCGs) are introduced as an abstract layer for example annotation to catch the transformation between an example sentence and its correspondent graph. Concepts and relations in graphs are denoted as functions with arguments. Therefore the main task of semantic analysis is decomposed into two parts: to construct an FCG using example-based machine translation methods; and to instantiate the FCG by solve the values of functions. The second part could be implemented by a lot of existing methods for retrieving relations between concepts. Moreover, this paper uses an active example selection approach to ease annotation work. Evaluation shows that this method is effective and can improve the labeling of relations.

1 Introduction

Query analysis forms the base of information retrieval systems. Most current enterprise systems view the users' queries, questions or phrases as bags of words. Such approaches distort the original requirements of users, which leads to further difficulties for IR. So, the semantic analysis of queries is essential for an IR system to better interpret users' intentions.

Some researchers take the availability of semantic representations of queries as granted. They assume that the users will write queries in a more structured language like RDQL^[1]. Consequently, there are not many discussions on the transformation from a user's natural language query into such a representation in some semantic retrieval systems^{[2][3]}. However we argue that it is not quite practical to rely on users only, because users in general will not bother to learn a new search language. Current search engines have not even persuaded the majority of users to use "advanced grammar" in searching like Boolean operators or quotation marks to identify a phrase.

^{*} This paper is supported in part by Chinese 863 project No. 2009AA01Z334 and the Shanghai Municipal Education Commission Foundation for Excellent Young University Teachers.

¹ www.w3.org/Submission/2004/SUBM-RDQL-20040109/

In this paper we focus on the semantic analysis of wh-questions in Chinese from real user queries. We use the semantic representation named *conceptual graph*(CG) [7], which describes concepts and their relations by a graph. There are a few papers on CG and query analysis. [8] generate CGs from questions using a unification based approach. Current researchers [9, 10] get CGs from queries by mapping a syntactic parsing tree onto a CG.

[11] regards a semantic parser as a translator, with conceptual graphs as target language and queries as source language. The idea is like the example-based machine translation (EBMT) suggested by Nagao [12]. In this paper, we suggest an improved framework of example-based query analysis enlightened by EBMT in which the examples are written in a novel representation called “functional conceptual graph”, which captures the transformation correspondence between the sentence and the graph. In this way we decompose the huge problem of semantic analysis by EBMT into two sub-problems: to analyzing the typology of the conceptual graph; and to resolve the labeling of concepts and relations. Consequently, we can effectively integrate other approaches for analyzing semantic relations into our framework, which will partially solve the data sparseness problem when we have only a few examples. We also suggest an active example selection algorithm to ease human work on annotation. Further evaluations in section 5 show that our method is effective and requires much less examples if an example selection algorithm is applied.

The paper is structured as following: in Section 2 we will introduce the original EBMT-based framework, its shortcomings and our proposed improvements in FCGs. Section 3 will discuss the definition of FCG and the analysis process depending on it. We will suggest an active example selection method in 4. In Section 5 we will show the evaluation results. The last section is the conclusion.

2 EBMT-Based Analysis of Queries

2.1 General Design

Conceptual graph is used as the semantic representation. Simply stated, a CG is a bipartite graph with two node sets: concept nodes and relation nodes. While annotating a group of user queries into conceptual graphs by hand, we use existing examples for guidance. If we have seen A and B labeled with relation R in previous examples, we may possibly link them up in similar sentences. If sentence or phrase patterns are limited, this approach could be very effective. The whole process is very similar to that of translation by examples. Enlightened by this similarity, we view the whole process as an example-based machine translation(EBMT) one. The main idea behind EBMT, just like our task, is translation by analogy.

Like the EBMT architecture, the prototype system has three components:

1. Matching. Constituents of the questions are matched with similar constituents of examples in the example base.

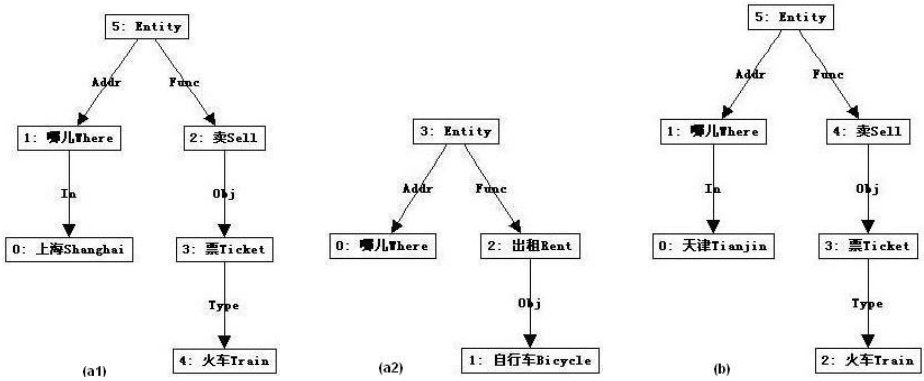


Fig. 1. Examples of concept graphs

2. Alignment. Picking up the correspondent tuples from the graph of the selected example of the matched constituents.
3. Recombination. The tuples are combined to form a graph, which is the smallest graph that covers all the concepts in the input sentence.

We will show the process through one simple example. Assume the input is a query “天津哪儿有火车票卖”(Where in Tianjin has train tickets to sell?). We will find two examples similar to it, one is “哪儿有自行车出租”(Where has bicycles to rent?)(Fig. 1 (a1)); the other is “上海哪儿能买到火车票”(Where can I buy train tickets in Shanghai?)(Fig. 1 (a2)). Parts of the alignments are shown in Table 1. Then the aligned segments are combined to form (b). The words in the graph are substituted in recombination.

Table 1. Word Alignment

Word/Word Pair	In Examples	Alignment
天津,哪儿	上海,哪儿	天津,哪儿,In
火车,票	火车,票	票,火车,Type
卖	出租	entity,卖,Func.
卖,火车票	出租,自行车	卖,火车票,Obj.

2.2 Problem and Approaches

In general, user queries are restricted in information needs and sentence patterns. Most users use similar phrases and sentences querying similar topics, such as movies, songs, etc. This similarity justifies our example-based approach, because it ensures that we do not need a very large sum of examples to reach satisfactory results.

But this approach still faces some problems in data sparseness. For example, when solving the relations between two concepts c_1 and c_2 , if c_1 and c_2 are frequently referred to in queries, such as “下载/download” and “电影/movies”, we can easily retrieve the relation as “对象/object” from examples. If not so, the situation is a little trickier. Optimistically, a large majority of related concepts in queries are popular ones that have appeared in the examples. But the queries we are studying are not totally restricted ones, so there will be lots of “free” concept pairs that do not exist in the examples. Therefore we should seek some methods to solve these relations. There are plenty of works on solving the relations between two concepts, like [13]. So our problem here is to adapt the EBMT framework to integrate these algorithms.

The same problem exists for the identification of concepts. The correspondence between a word and a concept node in the graph is not always explicit. In the example “上海哪儿能买到火车票”, the concept in the sentence is “买/buy” while the corresponding concept is “卖/sell” in the graph, which is the antonym of “买/buy”. Consequently, we have to note this correspondence in the graph by writing “卖” as “Ant(买)”.

To solve all these problems, we add an abstract layer to the conceptual graph in annotation. We explicitly notify the transformation from a sentence to a CG, i.e., all concept and relation nodes are viewed as the value of a function with its arguments. Therefore, we can first parse a query into a functional conceptual graph using the EBMT framework, and then solve all the function values using relation solving techniques.

3 Query Analysis Using Functional Conceptual Graphs

3.1 Functional Conceptual Graphs

Hereafter, we introduce the use of “functional conceptual graphs”, in which each concept and relation node is represented as a function with its arguments. Recall that a conceptual graph is a bipartite graph with a concept node set and a relation node set. Consequently, we will define the correspondent functional nodes.

Definition 1. A functional concept node $cu := \langle u, W' \rangle$, $W' \subseteq W$, $u : 2^{W'} \rightarrow C$. W is the set of all words, C is the set of concepts.

Moreover, we consider the situations that we do not know the function u , which is very common in annotations as in the “买” case. In such situations we only know the value of the function, e.g. “卖”. So we record it here and view the function as a hidden one to be clarified.

Definition 2. A hidden functional conceptual node $cc := \langle W', c \rangle$, $W' \subseteq W$, $v \in C$.

Following Definition 1 and 2, we have the following definitions:

Definition 3. A functional relation node $r_f := \langle f, p \rangle$, $f : C \times C \rightarrow R$, $p \in C \times C$. R is the set of relations.

Definition 4. A hidden functional relation node $rr := \langle p, r \rangle$, $r \in R$.

Summing these all up, we have the definition of functional annotation graphs:

Definition 5. A functional annotation graph $FAG := \langle C', R', L \rangle$, $C' \subseteq CU \cup CC$, $R' \subseteq RF \cup RR$. CU is the set of cus, CC the set of ccs, RF the set of rfs, and RR the set of rrs.

This is the graph model that we use in annotating questions. It includes hidden functional nodes because in annotation one may just not be able to recognize the function. For the sake of better illustration, we introduce a smaller model that excludes hidden nodes, i.e. all functions in the nodes are known.

Definition 6. A functional conceptual graph $FCG := \langle C', R', L \rangle$, $C' \subseteq CU$, $R' \subseteq RF$. CU is the set of cus, RF the set of rfs.

To get the FCGs from annotation graphs, we have to induce the hidden functions from the functions' independent and induced variables. That is to say, for a hidden function f , we have a set of examples $\langle x_i, f(x_i) \rangle_{0 < i < N}$ from annotations. We want to know the value of $f(x)$ for any $x \in D$, in which D is the field of definitions. In this implementation, we assume several kinds of known functions for hidden conceptual nodes. One is $I(x) = x$, which is the most common. The other is *Ant*, which maps a word to one of its antonyms. For example, $\text{Ant}(\text{买}) = \text{卖}$. The mapping of concepts from sentence is comparably easy so that this simple approach works well. It is a great challenge to learn a relation function from examples. There are a lot of researches focusing on this topic under different names, like [14] [13]. The adoption of functional nodes will make it possible to integrate such models into our work. However, in our current implementation, we use a simpler instance based approach. Assume that there is an instance set $S = \{ \langle c_1, c_2, r \rangle \}$. We construct a decision list D . Each rule in D is in the form of $\langle C_1, C_2 \rightarrow r; w \rangle$, in which C_1 and C_2 are two concept sets, r the relation label, and w the strength of the rule. The weight of each rule dr is decided as:

$$w = \frac{|S_{dr}|}{|S_{dr'}|} \quad (1)$$

In the above equation

$$S_{dr} := \{ \langle c_1, c_2, r_1 \rangle \mid c_1 \in C_1 \wedge c_2 \in C_2 \wedge r_1 = r \} \quad (2)$$

$$S_{dr'} := \{ \langle c_1, c_2, r_1 \rangle \mid c_1 \in C_1 \wedge c_2 \in C_2 \} \quad (3)$$

Given a concept pair $\langle c'_1, c'_2 \rangle$, we solve the relation between the two concepts by their similarity with a rule dr .

$$r' = \arg \max_r \text{SimDR}(\langle c'_1, c'_2 \rangle, dr) \quad (4)$$

In the above equation, the similarity function is further defined as:

$$SimDR(\langle c'_1, c'_2 \rangle, dr) = w \cdot \frac{1}{|C_1|} \sum_{c_1 \in C_1} SimSem(c_1, c'_1) \cdot \frac{1}{|C_2|} \sum_{c_2 \in C_2} SimSem(c_2, c'_2) \tag{5}$$

The similarity between two concepts is calculated as the similarity between two correspondent words. It is measured as a linear combination of two similarity models [15] [16]. [15] uses Hownet to calculate the similarity, while [16] bases the similarity on the similarity of dictionary definitions.

3.2 Matching

Matching is the first step of example-based query analysis. The goal of matching is to find a correspondent part in an example query for a word in the given query. Let W be a set of all words. Considering a given query $q \in 2^W$, an example base $E = \{e_i\}_{0 < i < N}$ in which $e_i \in 2^W$, we want to find a set M_{w_i} for each word w_i in q through matching.

$$M_{w_i} := \{ \arg \max_{w'_j} C_{e_k}(w_i, w'_j) \}_{0 < k < N, w'_j \in e_k} \tag{6}$$

Here C is a function that determines how well w_i and w'_j match. Basically, the word with the highest value in each example is selected. Though we match words in a query, the matching process is global instead of local. We consider not only the similarity between two words, but also the similarity between two sentences as well as the matching status of other words.

In a top-down manner, we define C as:

$$C(w_i)_e = Sim(q, e) \cdot ws_{w_i, w'_j} \tag{7}$$

In equation [7], Sim measures the similarity of two sentences, while ws_{w_i, w'_j} measures the similarity between two words. ws_{w_i, w'_j} considers not only the semantic similarity of two words as in the previous section, but also other aspects like dependency information, etc [11].

Each possible match between words in the given query and words in an example could be modeled as a *match path* which consists of a set of index pairs conforming to the criteria: $\forall (i, j), (k, l) \in p (i \neq k \wedge j \neq l)$. Thus we calculate the credit of each match as $CM(p) = \sum_{(i,j), (k,l) \in p} F_{i,j,k,l} \cdot ws_{i,j} * ws_{k,l}$, in which $F_{i,j,k,l}$ is a function concerning word order.

$$F_{i,j,k,l} = \begin{cases} 1 & \text{if } (i - j) * (k - l) > 0; \\ -1 & \text{else.} \end{cases} \tag{8}$$

The similarity measure of the two sentences is then decided by

$$Sim(s_1, s_2) = \max_{p \in P} CM(p) \tag{9}$$

in which P is the set of possible match paths of the two sentences.

3.3 Alignment and Recombination

We are to find a correspondent part in the example graph for the matched part in the given query. Alignment means two things here: to find the correspondent part in the FCGs; and to instantiate the graph to get a conceptual graph.

Let $\{ \langle w, w' \rangle \}$ be a set of matched words. w is a word in the given query, while w' is in the example. Let W'_m be the set of matched words in an example e . For an FCG g_e of example e , a concept node $c' = \langle u, W' \rangle$ is selected if and only if $W' \subseteq W'_m$. The aligned concept node is constructed as $c = \langle u, W_m \rangle$, in which W_m is the matched words in the given query. An instantiated node is $u(W_m)$.

Having the selected set C_s and the aligned set C'_s , we start to select relation nodes to be instantiated, which are relation nodes that are associated with the concept nodes. The selected nodes R'_s are:

$$\{r | r \in RU \wedge \exists c'_s \in C'_s ((c'_s, r) \in L \vee (r, c'_s) \in L)\} \tag{10}$$

Moreover, a new transition relation L_0 is constructed to reserve all the original transitions from concept nodes to relation nodes. An instantiated relation node is $f(c_1, c_2)$.

Some additional concept nodes are also selected, which may work as intermediate nodes in the final graphs.

$$C'_{sa} = \{c' | c' \in CU \wedge c' \notin C'_s \wedge \exists r (r \in R'_s)\} \tag{11}$$

We also construct aligned nodes which are just the same as the selected ones: $c = \langle u, W'_m \rangle$.

Basically, the essential parts of a conceptual graph are concepts and relations linking them. If we want to construct a graph, we'd better start from the tuples: $\langle c_1, c_2, r \rangle$, which represents a part of connected graph, not only a single concept or relation. Therefore, we take tuples as the basic constructing blocks for recombination. The tuple set can be easily constructed in the following way.

Definition 7. A functional tuple set $T_f := \{ \langle c_1, c_2, r \rangle | c_{1,2} \in C_s \cup C'_{sa} \wedge r \in R'_s \wedge (c_1, r) \in L_0 \wedge (r, c_2) \in L_0 \}$.

A tuple t is constructed as instantiations to the functional tuples. Moreover, the tuples are rated as:

$$C(t) = \begin{cases} \sqrt{\frac{C(c_1)^2 + C(c_2)^2}{2}} & \text{if } c_1 \in W_m \wedge c_2 \in W_m \\ \beta C(c_1) & \text{if } c_1 \in W_m \wedge c_2 \notin W_m \\ \beta C(c_2) & \text{if } c_2 \in W_m \wedge c_1 \notin W_m \end{cases} \tag{12}$$

In the above formula, β is a power coefficient between 0 and 1 which adjusts the credits for the situation that only one concept of the tuple is matched in the graph.

In recombination we use a forward/backward algorithm [11] to build the graph. In the forward stage we add tuples according to their credit values until all the

“core concepts” are covered and the graph is a connected one. Then, in the backward stage the tuples are removed one by one in the sequence reverse to that in the forward stage. If the removal affects the connectivity of the graph, the tuple is inserted back. In this manner we have a graph that is small but can cover all the essential concepts.

Finally, we give a certainty value for the final graph G_r which is a set of tuple, according to the rating of tuples $\{t_i\}_N$:

$$C(G_r) = \frac{1}{N} \sum_{0 < i \leq N} C(t_i) \quad (13)$$

4 Active Example Selection

For a data-driven task, one key factor of performance is the size of labeled data. However, manual labeling is time-consuming and error-prone. So the challenge here is how to achieve best result using only minimal annotations. In our system, we experiment a kind of example selection technique derived from active learning. Active learning [17] is a particular supervised learning method aiming at minimizing labeling effort. Unlike common learners that sample training examples randomly, in active learning the learner tries to select the most informative training examples, i.e. those contribute most to the performance of the classifier. In this way, active learning can minimize labeling work whilst retaining performance. We use an uncertainty based strategy here, which relies on the uncertainty given by our parser. If the parser is uncertain on an example using the current example base, then this example may contribute more to the system than other more certain examples.

Several un-annotated examples are first randomly selected and annotated to form an initial example base. We use 5 in the first iteration, though preliminary evaluation shows little difference from 3 initial examples or 5 initial ones. In each iteration, the system parses other un-annotated examples using current example base. We annotate manually N examples whose outputs have the lowest certainty values, and add them to the example base. In this way, we continue the building of example base until the required annotation number is reached. Following is the algorithm.

Algorithm

```

Build an initial annotated example base
While(the annotator is not tired)
  Use current annotated examples to parse un-annotated examples
  Select 5 examples with least certainty value
  Annotate the 5 examples
  Add them to the annotated example base

```

5 Experimental Results

All queries in this experiment are selected from Baidu Zhidao², which is a community-based Chinese question-answering Web site. In Zhidao, all queries are posted by users and answered by other users. Each query includes a title, mostly a question, and a short description. For our experiment, we select the titles only, since they are self-explanatory. For this experiment, we focus on three kinds of questions: 哪儿/Where, 哪个/Which and 什么是/What is.

Table 2. Examples of the three question types

Type	Data size (sentences)	Example
哪儿/Where	115	染发粉南京哪儿可以买到? Where to buy hair-dyeing powder in Nanjing?
哪个/Which	99	杭州哪个医院做手术好? Which hospital is good at surgery in Hangzhou?
什么是/What is	199	什么是笔记本上的小红帽? What is a TrackPoint on a laptop computer?

We select the questions randomly from Zhidao, though we delete some un-specific questions and some questions containing web slang, which are not the topic of this study. Totally, we have 314 questions in which there are 100 什么是 sentences, 115 哪儿 sentences and 99 哪个 sentences. Some examples along with their English translations are shown in Table 2. The questions are first segmented and tagged by an automatic tagger. A parser is also used to extract some useful syntactic information like dependency relations. Then the processed questions are edited and annotated by 3 graduate students in computational linguistics. The results are discussed in a research group. Each graph is represented by tuples in annotation.

Ten-fold cross-validation is used in evaluation. In each iteration, the data set is divided randomly into the test data and the potential examples according to a 1:3 ratio. The example base is a subset of the potential examples. Since there are ten different example/test set pairs, the experimental results are averaged to form the final result. Since the annotation graphs are in tuples, the precision, recall and f-measure of the results are measured based on tuples of graphs. Also, we consider graph-scale correctness which focuses on the correctness of a whole graph.

$$P = \frac{\text{Number of correct tuples}}{\text{Number of all tuples in the test set}} \quad (14)$$

$$R = \frac{\text{Number of correct tuples}}{\text{Number of all tuples in the gold set}} \quad (15)$$

$$F_{\beta=1} = \frac{2 \cdot P \cdot R}{P + R} \quad (16)$$

² zhidao.baidu.com

$$Correctness = \frac{\text{Number of correct graphs}}{\text{Number of graphs}} \quad (17)$$

Table 3 shows the evaluation results of our system using random sampling of examples. We consider two evaluation situations here: whether to take relation names into account or not. In some applications we only want to know the typology of the conceptual graph, not the labeling of relations. Therefore we include the no-relation situation in which the relation names are not evaluated. Furthermore, we compare two strategies in Table 3: one is the original example-based algorithm we proposed in [11], the other is the FCG-based approach. In the original approach, the relation names are just the same as that in the aligned tuples in the aligned example. Since in the FCG-based approach the main improvement is the refinement of relation labeling, we do not compare them in the no-relation case because the two results are very close. For the no-relation case, the F-measure is 0.7197 and the correctness of whole conceptual graphs is 0.4948. If relation names are considered, the F-measure is 0.6818 and 0.6683 using or not using FCG respectively. The FCG approach shows some improvement over the original approach.

Table 3. Results with Example Base of 220 Sentences

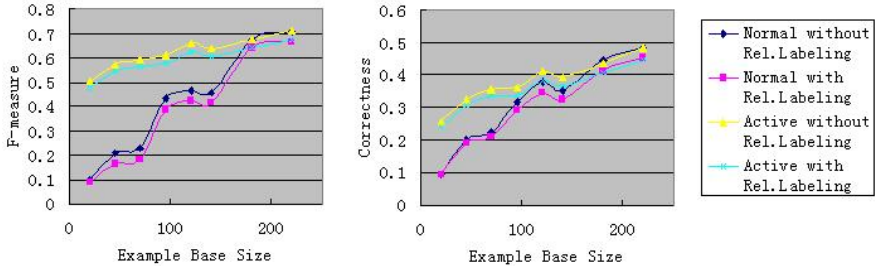
Rel.	Strategy	P	R	$F_{\beta=1}$	Correctness
No	/	0.7048	0.7354	0.7197	0.4948
Yes	No FCG	0.6545	0.6828	0.6683	0.4350
Yes	With FCG	0.6677	0.6967	0.68189	0.4580

Several similarity models are used in this approach. The models are compared indirectly according to the performance of our system in Table 4. In this table, Dict. means the dictionary-based approach. Hownet means the Hownet based approach. We use a simple linear combination to combine the two models into a hybrid model. Table 4 indicates that the Hownet-based approach outperforms the dictionary-based one, but combining the two models gives the best result. One interesting thing is that in a standard evaluation set for word similarities [16], the dictionary-based approach outperforms taxonomy-based approaches like Hownet. The contradiction is due to the fact that for the word similarity task, the gold data reflects human cognition, which is an ensemble of all factors. However, for this particular application of query analysis, the *is-a* relation between two concepts, which is captured by a hierarchy, is more important. Therefore, the Hownet-based approach offers better result.

We also try active example selection in our experiment. The results are shown in Fig. 2 along with the results of normal (random) selection method. From the figure, we see that when we label just a few examples, active selection provides far better results than normal selection method. The difference in results narrows as we label more data, because the size of unlabeled data is limited in current experiment. Still, we can see that active selection is helpful for our task, especially when we will have abundant data in the future.

Table 4. Results With Different Similarity Strategies

Rel. Strategy	P	R	$F_{\beta=1}$	Correctness
No Dict.	0.6702	0.6960	0.6828	0.4210
Yes Dict.	0.5786	0.6008	0.5894	0.3220
No Hownet	0.6864	0.7280	0.7065	0.4810
Yes Hownet	0.6527	0.6922	0.6718	0.4547
No Hybrid	0.7048	0.7354	0.7197	0.4948
Yes Hybrid	0.6677	0.6967	0.6818	0.4580

**Fig. 2.** Comparison of Normal Example Selection and Active Example Selection

6 Conclusion

In this paper the authors suggest an example-based method to analyze queries of Wh-questions in Chinese to conceptual graphs. The analysis is analogue to example-based machine translation. Unlike the previous work, the authors suggest a novel annotation layer: functional conceptual graphs to capture the transformation between a sentence and the annotated conceptual graph. Applied to the example-based approach, it can help to refine the labeling of relations. Moreover, an active example selection algorithm is proposed to enhance performance with just a few annotated examples.

Though current system only attacks three kinds of queries, it can be easily extended to other query types if the queries in one type share some similar query focuses, which is true for most queries since the users always search for similar things in similar words. For future works, we are to apply our approach to larger set of user queries. Another crucial problem is the role of similarity models. Though we give a simple comparison on different similarity models, it remains an interesting question that what kind of similarity model is suitable for semantic analysis. Current similarity models study similarity in an isolated manner, which is not suitable for applications. Maybe it is possible to use machine learning techniques to derive a new similarity model from existing ones according to requirements.

References

1. Contreras, J., Benjamins, V., Blazquez, M., Losada, S., Salla, R., Sevilla, J., Navarro, D., Casillas, J., Mompo, A., Paton, D.: A semantic portal for the International Affairs sector. LNCS, pp. 203–215 (2004)
2. Guha, R., McCool, R., Miller, E.: Semantic search. In: Proceedings of the twelfth international conference on World Wide Web, pp. 700–709 (2003)
3. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. *Computational Linguistics* 28(3), 245–288 (2002)
4. Narayanan, S., Harabagiu, S.: Question answering based on semantic structures. In: Proceedings of the 20th international conference on Computational Linguistics (2004)
5. Xue, N., Palmer, M.: Automatic Semantic Role Labeling for Chinese Verbs. In: Proc. IJCAI 2005 (2005)
6. Guo, S.: Preface. In: *Tongyici Cilin -A Thesaurus of Chinese* (1996)
7. Sowa, J.: *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., Boston (1984)
8. Velardi, P., Pazienza, M., Giovanetti, M.: *Conceptual Graphs for the Analysis and Generation of Sentences*. IBM Journal of Research and Development 32(2), 251–267 (1988)
9. Nicolas, S., Moulin, B., Mineau, G.: *sesei: A cg-based filter for internet search engines*. In: Ganter, B., de Moor, A., Lex, W. (eds.) ICCS 2003. LNCS, vol. 2746, pp. 362–377. Springer, Heidelberg (2003)
10. Zhong, J., Zhu, H., Li, J., Yu, Y.: *Conceptual Graph Matching for Semantic Search*. In: Priss, U., Corbett, D.R., Angelova, G. (eds.) ICCS 2002. LNCS (LNAI), vol. 2393, p. 92. Springer, Heidelberg (2002)
11. Liu, H., Zhao, J., Lu, R.: *An Example-based Approach to the Semantic Analysis of Questions*. In: The 11th China Conference on Machine Learning, CCML 2008, Dalian, China (2008)
12. Nagao, M.: *A framework of a mechanical translation between Japanese and English by analogy principle*. In: Proc. of the international NATO symposium on Artificial and human intelligence table of contents, pp. 173–180 (1984)
13. Hu, Y., Ruzhan Lu, L.H.: *Text Retrieval Oriented Auto-construction of Conceptual Relationship*. *Journal of Chinese Information Processing* (2007)
14. Maedche, A., Staab, S.: *Discovering conceptual relations from text*. In: ECAI, pp. 321–325 (2000)
15. Liu, Q., Li, S.: *Computation of Semantic Similarity between Words based on Hownet*. *Computational Linguistics and Chinese Language Processing* 7(2), 59–76 (2002)
16. Liu, H., Zhao, J., Lu, R.: *Computing Semantic Similarities based on Machine Readable Dictionaries*. In: The First International Workshop on Semantic Computing and Systems, WSCS 2008, Huangshan, China (2008)
17. Cohn, D., Ghahramani, Z., Jordan, M.: *Active Learning with Statistical Models*. Arxiv preprint cs.AI/9603104 (1996)

Checking Satisfactions of XML Referential Integrity Constraints*

Md. Sumon Shahriar and Jixue Liu

Data and Web Engineering Lab
School of Computer and Information Science
University of South Australia, SA-5095, Australia
shamy022@students.unisa.edu.au, jixue.liu@unisa.edu.au

Abstract. Recently we proposed referential integrity constraints for XML. In defining two important referential constraints namely XML inclusion dependency and XML foreign key, we considered ordered XML data model to capture the correct semantics of data when tuples are to be produced. In this paper, we report on the performances of checking both XML inclusion dependency and XML foreign key. We show that both these constraints can be checked in linear time in the context of number of tuples and the number of paths.

1 Introduction

In relational data model, integrity constraints are well studied and established [1,2]. The important integrity constraints include key constraints, functional dependency constraints and referential integrity constraints. The constraints are used in data integrity and semantics, normalization, data integrations [3], data warehousing and data mining techniques [4].

In recent years, XML [22] is massively used as data representation and storage formation over the world wide web and together with relational data storage. Now the use of XML for many data-centric activities such as data integration, data warehousing [6,5] is note worthy. With the increasing use of XML, the research for XML in database perspective is getting much attention. One such area is the constraints mechanism for XML [7,8,9,10].

The most important constraints those are investigated in XML are XML keys [17,9,10,20], XML functional dependencies [18,19] and XML multi valued dependencies [21]. Like inclusion dependencies in relational data model, inclusion dependencies for XML is also investigated in [11,12]. In both [11,12], XML inclusion dependencies are defined in the context of XML document without XML Document Type Definition(DTD). In [11], the definition of XML inclusion dependency can't capture the correct semantics if the number of paths is more than two and later this problem is addressed in the report [12].

* This research supported with Australian Research Council(ARC) Discovery Project(DP) Fund.

Recently, the referential integrity constraints for XML namely XML inclusion dependency(XID) and XML foreign key(XFK) [13] are defined on the XML DTD [22]. Though the DTD has the ID and IDREF definition for denoting key and referential integrity, the drawbacks of ID and IDREF are well recognized as their scope is the entire document and behave as object identifier. When defining XFK, we use the XID and the definition of XML key proposed in [14]. The satisfactions of XID, XML key and XFK are defined on the ordered model of XML document. We use the concept **tuple** that produces semantically correct values when the satisfactions for both XID and XFK are checked. This property is not achievable from both ID and IDREF of XML DTD and *Key* and *KeyRef* of XML Schema [23].

In this paper, we study the implementation and performance of checking XID and XFK proposed in [13]. The study is motivated by the fact that like in the relational database, the implementation of XFKs is critical to the quality of data in the XML database. Every time when there is a new instance for the database, we like to check the constraints against the new instance to ensure proper data is added or removed from the database. At the same time, the performance of the implementation is important to the efficiency of the database. Different ways of implementing the same mechanism will result in different performances. To a database management system, the efficiency of all processes is always critical to the success of the system.

In the literature, the checking of XML key and foreign keys using SAX was studied in [15] based on the proposal presented in [17]. An indexing based on paths on key is used in checking and the performance was shown as linear. Another study in [16] showed the XML key satisfaction checking using XPath based on DOM [24]. The study showed the checking of XML key can be done in polynomial time. We also use DOM (contrasting the use of SAX in [15]) for parsing XML document, but our implementation is different from the studies [15],[16] because we use a novel method of pairing the close values of elements to capture the correct semantics in tuple generation while parsing the document.

This paper is organized as follows. In Section 2, we give the basic definitions and notation of XID and XFK. The performances of checking XID and XFK are given in Section 3 and Section 4 respectively. In Section 5, we conclude with some remarks.

2 Basic Definitions and Notation

In this section, we review some basic definitions and notation proposed in [13] that is critical to guarantee the self-containment of the paper. We first introduce the DTD and paths on DTD using examples.

A DTD is defined in our notation as $D = (EN, \beta, \rho)$ where EN contains element names, ρ is the root of the DTD and β is the function defining the types of elements. For example, the DTD D in Fig. 1(a) is represented in our notation as $\beta(A) = [U^+ \times W^+]$, $\beta(U) = [B \times M^+ \times E]$, $\beta(M) = [C^* \times D^*]$, $\beta(W) = [F \times N^+ \times I]$, $\beta(N) = [G^* \times H^*]$, $\beta(B) = \beta(C) = \beta(D) = \beta(E) = Str$, $\beta(F) = \beta(G) = \beta(H) =$

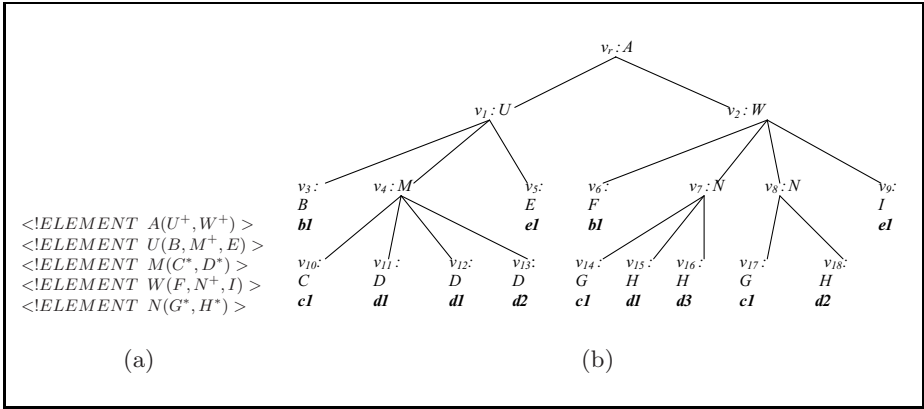


Fig. 1. (a) An XML DTD D and (b) An XML document T

$\beta(I) = Str$ where Str is $\#PCDATA$, $EN = \{A, U, B, C, D, E, M, W, F, G, H, I, N, Str\}$, and $\rho = A$. An element name and a pair of squared brackets $[]$ each, with its multiplicity, is called a component. For example, $[U^+ \times W^+]$, $[B \times M^+ \times E]$ are two components. A conjunctive or disjunctive sequence of components, often denoted by g , is called a structure. For example, the structure $g_1 = [B \times M^+ \times E]$ is a conjunctive sequence and $g_1 = [A|B]^+$ is a disjunctive sequence. A structure is further decomposed into substructures such as g_1 is decomposed into g_a and g_b where $g_a = B^+$ and $g_b = M^+ \times E$.

Now we define paths on the DTD. In Fig. 1(a), we say $U/M/C$ is a simple path and $A/U/M/C$ is a complete path. A complete path starts with the root element of the DTD. The function $beg(A/U/M)$ returns A , $last(A/U/M/C)$ returns C and $par(M)$ returns U .

We now define the XML inclusion dependency.

Definition 1 (XML Inclusion Dependency). An XML inclusion dependency over the DTD D can be defined as $\Upsilon(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$ where Q is a complete path called **selector**, P_i is a simple path called **dependent** path where $\beta(last(P_i)) = Str$ and R_i is a simple path called a **referenced** path where $\beta(last(R_i)) = Str$, Q/P_i and Q/R_i are complete paths.

An XID following the above definition is valid, denoted as $\Upsilon \sqsubseteq D$.

For example, consider the XID $\Upsilon(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ on the DTD D in Figure 1(a). The XID follows the definition 1.

Before defining XID satisfaction, we introduce some definitions and notation using examples.

An XML document is represented as an XML tree $T = (v : e (T_1 T_2 \dots T_f))$ if element e encloses other elements or $T = (v : e : txt)$ if e encloses the string value txt where v is the node identifier which is omitted when the context is clear, e is the element name labeled on the node, $T_1 \dots T_f$ are subtrees. For example,

in Fig. [II\(b\)](#), the document T is represented as $T_{v_r} = (v_r : A(T_{v_1}T_{v_2}))$. Then $T_{v_1} = (v_1 : U(T_{v_3}T_{v_4}T_{v_5}))$, $T_{v_2} = (v_2 : A(T_{v_6}T_{v_7}T_{v_8}T_{v_9}))$, $T_{v_3} = (v_3 : B : b1), \dots$ Other subtrees can be expressed in the same way. We say $T_{v_{11}} =_v T_{v_{12}}$ where $T_{v_{11}} = (v_{11} : D : d1)$ and $T_{v_{12}} = (v_{12} : D : d1)$.

Now we give an example to show the important concept *hedge* which is a sequence of adjacent subtrees for a type structure. Consider the structure $g_1 = [C^* \times D^*]$ in Fig. [II\(a\)](#). The trees $T_{v_{10}}T_{v_{11}}T_{v_{12}}T_{v_{13}}$ form a hedge conforming to g_1 under node v_4 . However, when we consider $g_2 = [G^* \times H^*]$, there are two sequence conforming to g_2 : $T_{v_{14}}T_{v_{15}}T_{v_{16}}$ for node v_7 and $T_{v_{17}}T_{v_{18}}$ for node v_8 . To reference various structures and their conforming sequences, we introduce the concept *hedge*, denoted by H^g , which is a sequence of trees conforming to the structure g .

Now we introduce two concepts *minimal structure* and *minimal hedge*. A minimal structure is the one that encloses two or more given elements with the bracket '[]'. Consider $\beta(A) = [B \times M^+ \times E]$ and $\beta(M) = [C^* \times D^*]$ for D in Fig. [II\(a\)](#). The minimal structure of B and C is $g_3 = [B \times M^+ \times E]$ meaning that both elements B and E is encompassed within the outermost '[]' bracket and C in $\beta(M)$. Thus the minimal hedge conforming to g_3 is $H_1^{g_3} = T_{v_3}T_{v_4}T_{v_5}$ for node v_1 in the document T in Fig. [II\(b\)](#). But the minimal structure of C and D is $g_2 = [C^* \times D^*]$. So the the minimal hedges conforming to g_2 are $H_1^{g_2} = T_{v_{10}}T_{v_{11}}T_{v_{12}}T_{v_{13}}$ for node v_4 in T .

We then use minimal structure and minimal hedge to produce *tuple* for the paths in $\mathcal{Y}(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$. We say $P(\mathbf{P-tuple})$ for paths P and $R(\mathbf{R-tuple})$ for paths R . Consider an XID $\mathcal{Y}(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ on the DTD D in Fig. [II\(a\)](#). Here, the selector path is A , the paths are $U/B, U/M/C, U/M/D, U/E$ are dependent paths(P-paths) and $W/F, W/N/G, W/N/H, W/I$ are referenced paths(R-paths).

First we produce the P-tuples. The minimal structure for element names B (note that $last(U/B) = B$), C , D and E is $g_4 = [B \times M^+ \times E]$ because B and E are enclosed in the '[]' and $\beta(M) = [C^* \times D^*]$. The minimal hedges for g_4 is $H_1^{g_4} = T_{v_3}T_{v_4}T_{v_5}$ under node v_1 . The P-tuples for the hedge $H_1^{g_4}$ are $F_1[P] = (T_{v_3}T_{v_{10}}T_{v_{11}}T_{v_5}) = ((v_3 : B : b1)(v_{10} : C : c1)(v_{11} : D : d1)(v_5 : E : e1))$, $F_2[P] = (T_{v_3}T_{v_{10}}T_{v_{12}}T_{v_5}) = ((v_3 : B : b1)(v_{10} : C : c1)(v_{12} : D : d1)(v_5 : E : e1))$, $F_3[P] = (T_{v_3}T_{v_{10}}T_{v_{13}}T_{v_5}) = ((v_3 : B : b1)(v_{10} : C : c1)(v_{13} : D : d2)(v_5 : E : e1))$. We say the P-tuples $F_1 =_v F_2$ because the values for corresponding trees in the tuples are the same.

We now produce R-tuples. The minimal structure for element names F (note that $last(W/F) = F$), G , H and I is $g_5 = [F \times N^+ \times I]$ because the elements F and I are in the '[]' and $\beta(N) = [G^* \times H^*]$. The minimal hedge for g_5 is $H_1^{g_5} = T_{v_6}T_{v_7}T_{v_8}T_{v_9}$ for node v_2 . We observe that there are two hedges for $[G^* \times H^*]$ in node v_1 . Thus we first need to produce pair-wise values for G, H elements and then we need to combine those pair-wise values with F and I elements to make the semantics correct. Thus R-tuples are $F_1[R] = (T_{v_6}T_{v_{14}}T_{v_{15}}T_{v_9}) = ((v_6 : F : b1)(v_{14} : G : c1)(v_{15} : H : d1)(v_9 : I : e1))$, $F_2[R] = (T_{v_6}T_{v_{14}}T_{v_{16}}T_{v_9}) = ((v_6 : F : b1)(v_{14} : G : c1)(v_{16} : H : d3)(v_9 : I : e1))$, $F_3[R] = (T_{v_6}T_{v_{17}}T_{v_{18}}T_{v_9}) = ((v_6 : B :$

$b1)(v_{17} : G : c1)(v_{18} : H : d2)(v_9 : I : e1)$). We see that all R-tuples are value distinct.

Now we are ready to define XID satisfaction.

Definition 2 (XML Inclusion Dependency Satisfaction). *An XML document T satisfies an XML inclusion dependency $\Upsilon(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$, denoted as $T \prec \Upsilon$ if there exists a P-tuple in T , then there must be a R-tuple where $(T^{P_i}) =_v (T^{R_i})$ and $i \in [1, \dots, n]$.*

Now consider the XID $\Upsilon(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ on the DTD D in Fig. [1\(a\)](#). We have already showed how to produce tuples for paths in P and R . We see that for all P-tuples, there exists a R-tuple. Thus the XID Υ is satisfied by the document T in Fig. [1\(b\)](#).

Now we define XML foreign key(XFK).

Definition 3 (XML Foreign Key). *Given an XID $\Upsilon(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$ on the DTD, we define XFK as $F(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$ if there is an XML key as $\mathbb{k}(Q, \{R_1, \dots, R_n\})$.*

For example, we define an XFK as $F(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ on the DTD D in Fig. [1\(a\)](#).

Before defining XFK satisfaction, we briefly describe the XML key satisfaction [14](#). An XML key $\mathbb{k}(Q, \{R_1, \dots, R_n\})$ is satisfied by the document T if all the R-tuples are value distinct in T . For example, we see that the XML key $\mathbb{k}(A, \{W/F, W/N/G, W/N/H, W/I\})$ on the DTD D in Fig. [1\(a\)](#) is satisfied by the document T in Fig. [1\(b\)](#) because all R-tuples are value distinct in the document T .

Definition 4 (XML Foreign Key Satisfaction). *An XML document T satisfies the XFK $F(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$ denoted as $T \prec F$ if both XID $\Upsilon(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$ and XML key $\mathbb{k}(Q, \{R_1, \dots, R_n\})$ are satisfied by the document T .*

We see that the XFK $F(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ is satisfied by the document T because both XID $\Upsilon(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ and the XML key $\mathbb{k}(A, \{W/F, W/N/G, W/N/H, W/I\})$ are satisfied by the document T .

We have just finished the definitions for XID and XFK. Now we give the experimental results on checking both XID and XFK satisfactions. All experiments are implemented in Java using a PC with Intel(R) Centrino Duo CPU T2050 at 1.60GHz, 1.49GB RAM and Microsoft Windows XP.

3 Checking XID Satisfactions

As mentioned before, there are two major tasks in checking XID satisfactions: (a) generation P- tuples and R-tuples and (b) checking the inclusion of P-tuples in R-tuples. First, we present the algorithms of generation of tuples.

In tuple generation, we accomplish two tasks: parsing the document and pairing the values of elements to produce tuples while parsing. Here the term pairing means the process of computing the product of relevant hedges. For example, if the relevant hedges are $H_a = T_1T_2$, $H_b = T_3T_4$ and $H_c = T_5T_6$, pairing produces the tuples (T_1, T_3, T_5) , (T_1, T_3, T_6) , (T_1, T_4, T_5) , (T_1, T_4, T_6) , (T_2, T_3, T_5) , (T_2, T_3, T_6) , (T_2, T_4, T_5) , and (T_2, T_4, T_6) . Product calculation itself is not difficult, but in the process of pairing, product calculation has to be combined with parsing.

The subsection presents two algorithms. The algorithm 1 shows the parsing and the algorithm 2 shows the pairing and tuple generation. In parsing, we first find the nodes QN for the selector path Q. We then proceed to find the occurrences of elements for paths dependent(P paths) in order of DTD under a selector node. Note that paths in dependent of an XID can appear as a set. But we order the paths of dependent of an XID according to the order of the elements of DTD that involve XID(we omit this process from the algorithm 1 for simplicity). We keep track of the occurrences of the elements which are the last elements of dependent paths so that the pairings can be done to produce P-tuples.

In the same way, we produce the R-tuples for paths referenced of an XID.

Data: An XML document T , An XID $\gamma(Q, (\{P_1, \dots, P_n\} \subseteq \{R_1, \dots, R_n\}))$

Result: A set of P-tuples, $F[P] = (T^{P_1} \dots T^{P_n})$ and

A set of R-tuples, $F[R] = (T^{R_1} \dots T^{R_n})$

Let QN= all Q nodes in T

foreach node in QN do

 | **FIND_TUPLES**($\{P_1, \dots, P_n\}, \text{node}$);

end

foreach node in QN do

 | **FIND_TUPLES**($\{R_1, \dots, R_n\}, \text{node}$);

end

//procedure

FIND_TUPLES($\{X_1, \dots, X_n\}, \text{node}$)

{

if (all of X_1, \dots, X_n occurred with order and any pairing before) **then**

 | **MAKE_TUPLES**($\text{array_X}_1[], \dots, (\text{pair}_{r_s}[]), \dots, \text{array_X}_n[]$);

end

foreach j=1 to n do

 | **foreach k=j to n-1 do**

 | Check any need to pair as **PAIRING**($\text{array_X}_j[], \text{array_X}_{k+1}[]$);

 | **end**

end

 Store value for path X_i to $\text{array_X}_i[]$ according to the order of the fields and keep track of order;

}

Algorithm 1. Parsing the document

```

//procedure
PAIRING(array_Xr[], array_Xs[])
{
  foreach i=1 to sizeof(array_Xr[]) do
    foreach j=1 to sizeof(array_Xr+1[]) do
      ...
      foreach k=1 to sizeof(array_Xs[]) do
        pair_rs[][1]=array_Xr[i];
        pair_rs[][2]=array_Xr+1[j];
        ...
        pair_rs[][s-r]=array_Xs[k];
      end
    end
  end
}
//procedure
MAKE_TUPLES(array_X1[], ..., (pair_rs[]), ..., array_Xn[])
{
  foreach i=1 to sizeof(array_X1[]) do
    foreach j=1 to sizeof(pair_rs[]) do
      foreach k=1 to sizeof(array_Xn[]) do
        tuple_1...n[][1] = array_X1[i];
        ...
        tuple_1...n[][r] = pair_rs[j][1];
        ...
        tuple_1...n[][s] = pair_rs[j][s-r];
        ...
        tuple_1...n[][n] = array_Xn[k];
      end
    end
  end
}

```

Algorithm 2. Pairing of values for fields and generation of tuples

We give an example to illustrate how to produce P-tuples and R-tuples using the above algorithms. Consider the XID $\Upsilon(A, (\{U/B, U/M/C, U/M/D, U/E\} \subseteq \{W/F, W/N/G, W/N/H, W/I\}))$ on the DTD D in Fig. 1(a). Here, the selector path is A , the paths $U/B, U/M/C, U/M/D, U/E$ are dependent paths(P-paths) and $W/F, W/N/G, W/N/H, W/I$ are referenced paths(R-paths).

In algorithm 1, we take the document T and the XID Υ as inputs. First, we get the Q node v_r . We then take the P-paths to produce P-tuples for node v_r . We get the value $b1$ for node v_3 using path U/B , value $c1$ for node v_{10} using path $U/M/C$, values $d1, d1, d2$ for nodes v_{11}, v_{12}, v_{13} using path $U/M/D$ and value $e1$ for node v_5 using path U/E . There is no need to call PAIRING procedure as there is no repeating occurrences of a set of elements in the document. So we call the procedure MAKE_TUPLES to produce P-tuples. So we get the P-tuples $F_1[P] = ((b1)(c1)(d1)(e1))$, $F_2[P] = ((b1)(c1)(d1)(e1))$ and $F_3[P] = ((b1)(c1)(d2)(e1))$.

We omit the node identifier v and element label e from the tuples because we consider the value comparison in satisfactions.

Now we show how to produce R-tuples for paths $W/F, W/N/G, W/N/H, W/I$. For the selector node v_r , the value for node v_6 using path W/F is $b1$. Then we get the value $c1$ for node v_{14} using path $W/N/G$ and the values $d1, d3$ for nodes v_{15}, v_{16} using path $W/N/H$. We again see the repeating occurrence of values $c1$ for node v_{17} using path $W/N/G$ and the value $d2$ for node v_{18} using path $W/N/H$. Thus we call the procedure PAIRING for the values for the elements G, H . So we get the pair-wise values $(c1, d1), (c1, d3)$ and $(c1, d2)$. At last, we call the procedure MAKE_TUPLES with values of F element, pair-wise (G, H) values and I value. So we get the R-tuples $F_1[R] = ((b1)(c1)(d1)(e1))$, $F_2[R] = ((b1)(c1)(d3)(e1))$ and $F_3[R] = ((b1)(c1)(d2)(e1))$.

We have shown how to generate P-tuples and R-tuples. We are now ready to check the satisfactions of XID. In checking satisfactions of XID, we use hashing technique using Java HashTable. After generating tuples, we put the values of R-tuples as *key*¹ in the hashtable. We then take each P-tuple and check the hashtable to see whether there exists a R-tuple that is value equivalent to that P-tuple. If there exists a R-tuple for each P-tuple in the hashtable, we say that the XID is satisfied.

We now present the experimental results of checking XID satisfactions. We take the DTD D in Fig. 1(a) and we generate XML documents with different structures and different sizes. By different structures, we mean *repeating* structure and *non-repeating* structure. By repeating structure, we mean the multiple occurrences of elements in the hedge and by non-repeating structure, we mean the single occurrence of the elements in the hedge in the document. With single occurrence, an element or a group of elements appears only once while with multiple occurrence, an element or a group of elements appear more than one time in a hedge.

In the case of repeating structure, elements need to be combined from different occurrences to form a tuple. For the same number of tuples, if the structure is non-repeating, we need larger number of elements in the document, which means larger document size. In contrast, if the structure is repeating, because of production, a small document will make the number of tuples.

We first study the experimental results of XID satisfactions for the documents with non-repeating structure in Fig. 2 and Fig. 3. For non-repeating structure, the complexity of tuple generation is $O(2n|e|)$ where n is the number of fields and $|e|$ is the average number of occurrences of an element in a hedge. The cost of parsing is $n|e|$. As we use breadth first search in traversing the document, thus we need to traverse $n|e|$ element nodes. In this case, we assume all element nodes are at the same level under a parent(context) node. The cost of pairing is $n|e|$ for close elements in the hedge. This is because in each pair, there is only one occurrence of each element. We note here that the complexity of hashing is nearly constant and linear but it is slightly increasing in most checking. However,

¹ This key is in Java Hashtable(Key,Value), not key constraints in XML.

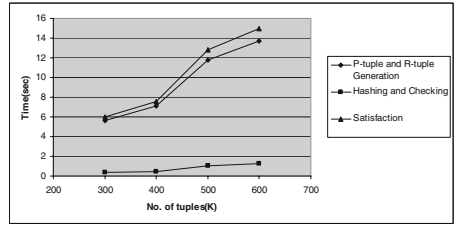
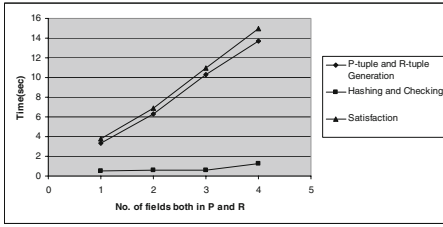


Fig. 2. XID satisfaction time when the number of tuples is fixed to 300K for P and 300K for R

Fig. 3. XID satisfaction time when the number of paths is fixed to 4 in both P and R

sometimes, the time in hashing is increased in non-linear fashion due to the Java hashtable management.

In Fig. 2, the XID satisfaction time is shown where the number of tuples is fixed to 300K both for P-tuples and R-tuples. Thus the total number of tuples in the document is 600K. The satisfaction time for XID is the sum of the tuple generation time and the hashing time of R-tuples and then the inclusion checking time of P-tuples in R-tuples. Both tuple generation time and hashing time with inclusion checking time are linear, but the tuple generation time is increasing because the number of paths is increased in XID whereas the the hashing with inclusion checking time is slightly increasing and this is due to Java Hashtable management to maintain the values for increasing number of paths. We see that tuple generation time much higher than the hashing with inclusion checking time.

We show the XID satisfaction time where we fix the number of paths to 4 both in P and R in Fig. 3. Here, the XID satisfaction time is also linear because the tuple generation time and the hashing and the inclusion checking time are also linear. The tuple generation time is increasing because the number of tuples to be produced is increasing. The hashing and the inclusion checking is slightly increasing due to Java Hashtable management to maintain values of the increasing number of tuples.

We now study the experimental results of XID satisfactions for documents with repeating structure in Fig. 4 and Fig. 5. For the repeating structure in the document, the complexity of tuple generation is $O(n \sqrt[n]{|e|} + |e|)$ where n is the number of fields and $|e|$ is the average number of occurrences of an element in a hedge. The cost $n \sqrt[n]{|e|}$ is for parsing using breadth first search. The cost $|e|$ is for pairing because we do the production of elements.

In Fig. 4, the XID satisfaction time is shown where we fix the number of P-tuples to 300K and R-tuples to 300K and we vary the number of paths in P and R. We start with the number of paths two because the pairing has to be done with at least two paths. Here, the P-tuple and R-tuple generation time is smaller than the hashing time and the inclusion checking time. This is because the parsing time is smaller for a document with repeating structure than that of a document with non-repeating structure. However, the satisfaction time is

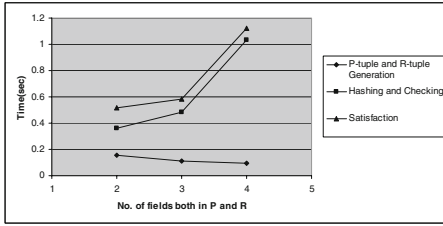


Fig. 4. XID satisfaction time when the number of tuples is fixed to 300K for P and 300K for R

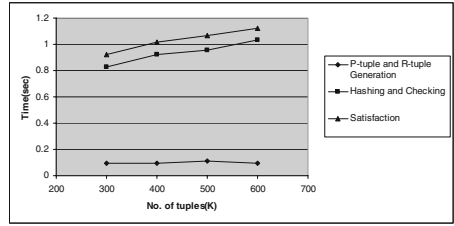


Fig. 5. XID satisfaction time when the number of paths is fixed to 4 in both P and R

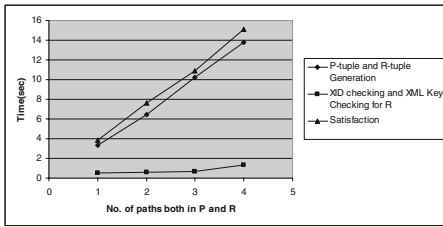


Fig. 6. XFK satisfaction time when the number of tuples is fixed to 300K for P and 300K for R

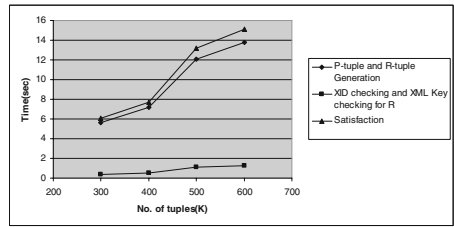


Fig. 7. XFK satisfaction time when the number of paths is fixed to 4 in both P and R

linear considering the tuple generation time and the hashing and the inclusion checking time are linear. We also see that the hashing time is increasing while the tuple generation time is decreasing linearly.

We show the XID satisfaction time in Fig. 5 where we fix the number of paths to 4 but we vary the number of tuples. Like Fig. 4, the XID satisfaction time is linear with the increasing number of tuples to be checked. The tuple generation time is significantly smaller than the hashing and the inclusion checking time.

4 Checking XFK Satisfaction

In this section we study the performance of checking XML foreign key satisfaction. In checking XFK satisfaction, we need two checking: checking inclusion of P-tuples in R-tuples(XID checking) and whether R-tuples are distinct(XML key checking). We already showed how to check XID in the previous section. We only add the XML key checking for paths R to check the satisfaction of XFK. In checking XFK, we only show the experimental results for non-repeating structure of the document.

In Fig. 6, the XFK satisfaction time is shown where we fix the number of tuples to 300K both for P-tuples and R-tuples but we vary the number of paths in P and R. We see that tuple generation time is linear and it is increasing.

But the XID of P-tuples in R-tuples checking and the XML key checking of R-tuples is much smaller than the tuple generation time, but it is linear and slightly increasing. Thus the XFK satisfaction time is linear with the increasing number of paths in both P and R.

We show the XFK satisfaction time in Fig. 7 where we fix the number of paths to 4 but we vary the number of tuples. The XFK satisfaction time is linear and increasing. The tuple generation time is more increasing than the checking of inclusion of P-tuples in R-tuples and the checking of distinctness of R-tuples.

5 Conclusions

We have showed the performances of checking XML inclusion dependency and XML foreign key satisfactions. A novel algorithm for tuple generation is used in checking satisfactions. Both XID and XFK satisfactions can be checked in linear time with the increasing number of tuples in the document and the increasing number of paths in XID and XFK.

References

1. Abiteboul, S., Hull, R., Vianu, V.: *Foundations of Databases*. Addison-Wesley, Reading (1995)
2. Ramakrishnan, R., Gehrke, J.: *Database Management Systems*. McGraw-Hill Higher Education, New York (2003)
3. Cali, A., Calvanese, D., De Giacomo, G., Lenzerini, M.: Data Integration under Integrity Constraints. In: Pidduck, A.B., Mylopoulos, J., Woo, C.C., Ozsu, M.T. (eds.) CAiSE 2002. LNCS, vol. 2348, pp. 262–279. Springer, Heidelberg (2002)
4. Bykowski, A., Daurel, T., Meger, N., Rigotti, C.: Integrity Constraints Over Association Rules. In: Meo, R., Lanzi, P.L., Klemettinen, M. (eds.) DSDMA 2004. LNCS (LNAI), vol. 2682, pp. 306–323. Springer, Heidelberg (2004)
5. Fankhouser, P., Klement, T.: XML for Datawarehousing Chances and Challenges. In: Kambayashi, Y., Mohania, M., Wöß, W. (eds.) DaWaK 2003. LNCS, vol. 2737, pp. 1–3. Springer, Heidelberg (2003)
6. Zamboulis, L.: XML Data Integration by Graph Restructuring. In: Williams, H., MacKinnon, L.M. (eds.) BNCOD 2004. LNCS, vol. 3112, pp. 57–71. Springer, Heidelberg (2004)
7. Buneman, P., Fan, W., Simeon, J., Weinstein, S.: Constraints for Semistructured Data and XML. *SIGMOD Record*, 47–54 (2001)
8. Fan, W.: XML Constraints: Specification, Analysis, and Applications. In: DEXA, pp. 805–809 (2005)
9. Fan, W., Simeon, J.: Integrity constraints for XML. In: PODS, pp. 23–34 (2000)
10. Fan, W., Libkin, L.: On XML Integrity Constraints in the Presence of DTDs. *Journal of the ACM* 49, 368–406 (2002)
11. Vincent, M.W., Schrefl, M., Liu, J., Liu, C., Dogen, S.: Generalized inclusion dependencies in XML. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 224–233. Springer, Heidelberg (2004)
12. Karlinger, M., Vincent, M., Schrefl, M.: On the Definition and Axiomatization of Inclusion Dependency for XML, Technical Report, No. 07/02, Johannes Kepler University (2007)

13. Shahriar, Md., S., Liu, J.: On Defining Referential Integrity for XML. In: IEEE International Symposium of Computer Science and Its Applications (CSA), pp. 286–291 (2008)
14. Shahriar, Md., S., Liu, J.: On Defining Keys for XML. In: IEEE CIT 2008, Database and Data Mining Workshop, DD 2008, pp. 86–91 (2008)
15. Liu, Y., Yang, D., Tang, S., Wang, T., Gao, J.: Validating key constraints over XML document using XPath and structure checking. *Future Generation Computer Systems* 21(4), 583–595 (2005)
16. Chen, Y., Davidson, S.B., Zheng, Y.: XKvalidator: a constraint validator for XML. In: CIKM, pp. 446–452 (2002)
17. Buneman, P., Davidson, S., Fan, W., Hara, C., Tang, W.C.: Keys for XML. *WWW* 10, 201–210 (2001)
18. Vincent, M.W., Liu, J.: Functional dependencies for XML. In: Zhou, X., Zhang, Y., Orłowska, M.E. (eds.) APWeb 2003. LNCS, vol. 2642, pp. 22–34. Springer, Heidelberg (2003)
19. Arenas, M., Libkin, L.: A Normal Form for XML documents. In: ACM PODS, pp. 85–96 (2002)
20. Hartmann, S., Köhler, H., Link, S., Trinh, T., Wang, J.: On the Notion of an XML Key. In: Schewe, K.-D., Thalheim, B. (eds.) SDKB 2008. LNCS, vol. 4925, pp. 103–112. Springer, Heidelberg (2008)
21. Vincent, M.W., Liu, J.: Multivalued Dependencies in XML. In: James, A., Younas, M., Lings, B. (eds.) BNCOD 2003. LNCS, vol. 2712, pp. 4–18. Springer, Heidelberg (2003)
22. Bray, T., Paoli, J., Sperberg-McQueen, C.M.: Extensible Markup Language (XML) 1.0., World Wide Web Consortium (W3C) (February 1998), <http://www.w3.org/TR/REC-xml>
23. Thompson, H.S., Beech, D., Maloney, M., Mendelsohn, N.: XML Schema Part 1:Structures, W3C Working Draft (April 2000), <http://www.w3.org/TR/xmlschema-1/>
24. Java Document Object Model (DOM), <http://java.sun.com/j2se/1.4.2/docs/api/orgw3c/dom/package-summary.html>

A Verification Method of Hyponymy between Chinese Terms Based on Concept Space

Lei Liu¹, Sen Zhang¹, Lu Hong Diao¹, Shu Ying Yan², and Cun Gen Cao²

¹ College of Applied Sciences, Beijing University of Technology

² Institute of Computing Technology, Chinese Academy of Sciences
{liuliu_leilei, zhansen, diaoluhong}@bjut.edu.cn,
{yanshuying, cgcao}@ict.ac.cn

Abstract. Hyponymy is a basic semantic relation. There are still many error relations in the automatic acquired hyponymy relations from free text. They will affect the building of ontologies. We present an iterative method of hyponymy verification between Chinese terms based on concept space. We give the definition of concept space about a group of candidate hyponymy relations initially. Then a set of hyponymy features based on concept space are acquired. These features are changed into production rules. Finally the candidate hyponymy relations are iterative verified with those production rules. Experimental results demonstrate good performance of the method.

1 Introduction

Hyponymy is a semantic relation between concepts. Given two concepts x and y , there is the hyponymy between x and y if the sentence “ x is a (kind of) y ” is acceptable. x is a hyponym of y , and y is a hypernym of x . Hyponymy is also called as subordination, or “isa” relation. This relation is irreflexive, transitive and asymmetrical under the condition of the same sense. We denote a hyponymy relation as $hr(x, y)$. For example, $hr(\text{apple}, \text{fruit})$.

Hyponymy acquisition is a more interesting and fundamental because hyponymy relations play a crucial role in various NLP (Natural Language Processing) systems, such as systems for information extraction, information retrieval, and dialog systems. Hyponymy relations are important in accuracy verification of ontologies, knowledge bases and lexicons [1].

The types of input used for hyponymy relation acquisition are usually divided into three kinds: the structured text (e.g. database), the semi-structured text (e.g. dictionary), and free text (e.g. Web pages)[2]. Human knowledge is mainly presented in the format of free text at present, so processing free text has become a crucial yet challenging research problem.

There are still many error relations in the acquired hyponymy relations from free text. They will affect the building of ontologies. In this paper, for the error relations in the phase of acquiring hyponymy, we present an iterative method of hyponymy verification based on concept space. Experimental results show that the method is adequate of verifying the hyponymy relations from Chinese free text. The rest of the

paper is organized as follows. Section 2 describes related work of automatic hyponymy acquisition, section 3 section gives the definition of concept space and analyzes the hyponymy features for this work, section 4 presents how to change these features into a set of production rules, and how to iterative verify candidate hyponymy relations, section 5 conducts a performance evaluation of the proposed method, and finally section 6 concludes the paper.

2 Related Work

There are two main approaches for automatic/ semi-automatic hyponymy acquisition. One is pattern-based (also called rule-based), and the other is statistics-based. The former uses the linguistics and natural language processing techniques (such as lexical and parsing analysis) to obtain hyponymy patterns, and then makes use of pattern matching to acquire hyponymy, and the latter is based on corpus and statistical language model, and uses clustering algorithm to acquire hyponymy.

At present the pattern-based approach is dominant, and its main idea is the hyponymy can be extracted from text as they occur in detectable syntactic patterns. The so-called patterns include special idiomatic expressions, lexical features, phrasing features, and semantic features of sentences. Patterns are acquired by using the linguistics and natural language processing techniques.

There have been many attempts to develop automatic methods to acquire hyponymy from text corpora. One of the first studies was done by Hearst [3]. Hearst proposed a method for retrieving concept relations from unannotated text (Groslier's Encyclopedia) by using predefined lexico-syntactic patterns, such as

...NP ₁ is a NP ₂ ...	---hr(NP ₁ , NP ₂)
...NP ₁ such as NP ₂ ...	---hr(NP ₂ , NP ₁)
...NP ₁ {, NP ₂ }*{,} or other NP ₃ ...	---hr (NP ₁ , NP ₃), hr (NP ₂ , NP ₃)

Other researchers also developed other ways to obtain hyponymy. Most of these techniques are based on particular linguistic patterns.

Caraballo used a hierarchical clustering technique to build a hyponymy hierarchy of nouns like the hypernym-labeled noun hierarchy of WordNet from text [4]. Nouns are clustered into a hierarchy using data on conjunctions and appositives appearing in text corpus. The internal nodes of the resulting tree are labeled by the syntactic constructions from Hearst.

Maedche et al. propose to learn the ontology using as a base a core ontology that is enriched with new specific domain concepts. The hyponymy is acquired iteratively by the use of natural language analysis techniques in ontology learning [5].

Sánchez presented a novel approach that adapted to the Web environment, for composing taxonomies in an automatic and unsupervised way [6].

Elghamry showed how a corpus-based hyponymy lexicon with partial hierarchical structure for Arabic can be created directly from the Web with minimal human supervision. His method bootstraps the acquisition process by searching the Web for the lexico-syntactic patterns [7].

3 Concept Space and Features

In our research, the problem of hyponymy verification is described as follows:

Given a set of candidate hyponymy relations acquired based on pattern or statistics, we denoted these relations as $CHR = \{(c_1, c_2), (c_3, c_4), (c_5, c_6), \dots\}$, where c_i is the concept of constituting candidate hyponymy relation. The problem of hyponymy verification is how to identify correct hyponymy relations from CHR using some specific verify methods. Here we present an iterative method of hyponymy verification based on concept space.

Definition 2: The concept space is a directed graph $G = (V, E, W)$ where nodes in V represent concepts of the hyponymy and edges in E represent relationships between concepts. A directed edge (c_1, c_2) from c_1 to c_2 corresponds to a hyponymy from concept c_1 to concept c_2 . Edge weights in E are used to represent varying degrees of certainty factors.

For a node c in a graph, we denote by $I(c)$ and $O(c)$ the set of in-neighbors and out-neighbors of v , respectively. Individual in-neighbors are denoted as $I_i(c)$, for $1 \leq i \leq |I(c)|$, and individual out-neighbors are denoted as $O_i(c)$, for $1 \leq i \leq |O(c)|$.

Firstly we acquire a group of candidate hyponymy relations using Chinese lexico-syntactic patterns from Chinese corpus which is not restricted to domain-specific. When a candidate hyponymy is correct or error, it often satisfies some features. We combine the semantic features, context features and space structure features of hyponymy together. If a candidate hyponymy satisfies a certain threshold with matching those features, we think that it is a real hyponymy. The features of hyponymy are defined as follows:

Definition 2: The feature of hyponymy is a 3-tuple $ISAF = \{SMF, CTF, STF\}$, where SMF is a group of semantic features, CTF is a group of context features and STF is a group of space structure features.

3.1 Semantic Features

SMF is constructed by the assumption that x and y are semantically similar in $hr(x, y)$. A candidate hyponymy can be verified by computing the semantic similar measure between x and y . SMF is subdivided into three features, i.e. $SMF = \{WF, SF, AF\}$, where WF represents word-formation feature, SF represents synonymous word feature and AF represents attribute feature.

(1) WF (word-formation feature)

As we know, Chinese is a language different from any western language [8]. A Chinese sentence consists of a string of characters which do not have any space or delimiter in between; Chinese word order is strict; Chinese lacks morphological change, and has no the explicit variety tag of plural, possessive and part of speech. A concept consists of one or several certain sequence Chinese characters. To some extent, Chinese characters can appear the semantic feature of concept.

So for each pair of candidate hyponymy (c_1, c_2) , we assume the common substrings between c_1 and c_2 could imply the semantic similar measure between them. If there exists some substrings, the position (such as prefix and suffix), length and amount of substrings will provide the evidence for the existence of a hyponymy.

Given a candidate hyponymy (c_1, c_2) , where $c_1=a_1a_2a_3\dots a_n$, $c_2=b_1b_2b_3\dots b_m$, the a_i and b_i both represent single character. We can give some features as follows:

$\text{CoChar}(c_1, c_2) = \{a_1, a_2, a_3, \dots, a_n\} \cap \{b_1, b_2, b_3, \dots, b_m\}$, $|\text{CoChar}(c_1, c_2)|$ denote the number of common character of c_1 and c_2 .

Obviously, the above definition doesn't consider the sequence of character. So we add the following definitions.

$\text{CoPrefix}(c_1, c_2) = a_1a_2\dots a_i$, where $a_1a_2\dots a_i = b_1b_2\dots b_i$, $a_{i+1} \neq b_{i+1}$, $i \leq n$, $i \leq m$. Especially, if $\text{CoPrefix}(c_1, c_2) = c_2$, then c_2 is the prefix c_1 ; if $\text{CoPrefix}(c_1, c_2) = c_1$, c_1 is the prefix c_2 .

$\text{CoSuffix}(c_1, c_2) = a_j\dots a_{n-1}a_n$, where $a_j\dots a_{n-1}a_n = b_{m-j}\dots b_{m-1}b_m$, $a_{n-j-1} \neq b_{m-j-1}$, $i < n$, $i < m$. Especially, if $\text{CoSuffix}(c_1, c_2) = c_2$, then c_2 is the suffix c_1 ; if $\text{CoSuffix}(c_1, c_2) = c_1$, c_1 is the suffix c_2 .

For example:

$\text{CoSuffix}(\text{定时炸弹, 炸弹}) = \text{炸弹}$ ($\text{CoSuffix}(\text{Time bomb, bomb}) = \text{bomb}$)

$\text{CoChar}(\text{话费优惠业务, 业务套餐}) = \{\text{业务}\}$

$(\text{CoChar}(\text{cheap-charge-of-calls service, service}) = \{\text{service, plan}\})$

$\text{CoPrefix}(\text{诗人屈原, 诗人}) = \text{诗人}$ ($\text{CoPrefix}(\text{poet QuYuan, poet}) = \text{poet}$)

(2) SF (synonymous word feature): We can use Cilin(a dictionary of Chinese synonymous words) to compute the semantic similarity of candidate (c_1, c_2) . Cilin provides the mandarin synonym sets in a hierarchical structure. It contains approximately 70,000 Chinese words, and describes a five levels semantic hierarchy from common word to concrete word [9]. The fifth level is for the basis of synonym feature words. We denote $\text{Syn}(c_1, c_2)$ as the common synonymous words between c_1 and c_2 . “医生”(doctor) and “大夫”(doctor) are synonymous word in the below example.

$\text{Syn}(\text{主治医生, 大夫}) = \{\text{医生|大夫}\}$

$(\text{Syn}(\text{doctor in charge of a case, doctor}) = \{\text{doctor|doctor}\})$

(3) AF (attribute feature): The attributes of concept can be used to discriminate different concept. If two concepts have the same attributes, they should be semantic similar. The attributes of concept can be acquired by an attribute acquiring system [10]. We denote common attributes as $\text{CoAttr}(c_1, c_2)$. For example:

$\text{CoAttr}(\text{黄河, 河流}) = \{\text{上游}\}$ ($\text{CoAttr}(\text{yellow river, river}) = \{\text{upriver}\}$)

$\text{CoAttr}(\text{比利时, 国家}) = \{\text{面积, 首都, 人口}\}$

$(\text{CoAttr}(\text{Belgium, country}) = \{\text{area, capital, population}\})$

3.2 Context Features

Here we verify hyponymy using contextual knowledge. The co-occurrence context features are subdivided into two features, CTF= {FF, DF}.

(1) FF (frequency features): If candidate (c_1, c_2) appears frequently in a kind of hyponymy pattern or in various hyponymy patterns, the probability of $hr(c_1, c_2)$ is higher. The type number of pattern that can acquire (c_1, c_2) is denoted by $\text{lpf}(c_1, c_2)$. The total of number of pattern that can acquire (c_1, c_2) is denoted by $\text{lef}(c_1, c_2)$. For example:

lpf(爱因斯坦, 著名科学家) = 3
 (lpf(Einstein, famous scientist) = 3)
 lef(摩托车, 机动车) = 56
 (lef(motorcycle, machine) = 56)

(2)DF (domain features): Our corpus comes from Web and includes some error knowledge. We can acquire many error hyponymy, such as (美丽, 罪) ((beauty, evil)). If candidate (c_1, c_2) appears in a certain scientific domain-specific context, (c_1, c_2) may be a true piece of scientific knowledge; otherwise it may be a pair of general concepts and may not have any value. The domain-specific context is discriminated with a domain dictionary [2]. Given a group of context $CT(c_1, c_2) = \{ct_1, ct_2, \dots, ct_n\}$, where ct_i is the i context of (c_1, c_2) , $fw(ct_i)$ is the number of domain word in ct_i , and $length(ct_i)$ is the byte length of ct_i . The classify formula is as follows.

$$Classify(c_1, c_2) = \frac{\sum_{i=1}^n fw(ct_i)}{\sum_{i=1}^n length(ct_i)} \times 1000 \quad (1)$$

For example:

Classify(木瓜, 水果) = 19.5 (Classify(pawpaw, fruit) = 19.5)
 Classify(钠, 元素) = 52.1 (Classify(natrium, element) = 52.1)

3.3 Space Structure Features

When a group of candidate hyponymy relations are correct or error, they often satisfy some space structure feature. In space structure analysis, we use the coordinate relation between concepts. The coordinate relations are acquired using a set of coordinate relation patterns including “、”. Chinese dunhao “、” is a special kind of Chinese comma used to set off items in a series. For example:

In a sentence of matching a coordinate pattern, if there exists concept c_1 and concept c_2 divided by “、”, then c_1 and c_2 are coordinate, denoted as $cr(c_1, c_2)$. An example is as shown below.

农作物主要有{水稻} c_1 、{玉米} c_2 、{红薯} c_3 、{烟叶} c_4 等

(The farm crop mainly includes paddy rice, corn, sweet potato, tobacco leaves etc..)

$cr(\text{水稻, 玉米, 红薯, 烟叶})$ ($cr(\text{paddy rice, corn, sweet potato, tobacco leaves})$) is acquired from the above example.

We can add some space structure features on the basis of above coordinate relations. A few important features are as follows:

Structure 1: $(c_1, c_2), (c_2, c_3), (c_1, c_3)$. For example:

(番茄, 蔬菜), (番茄, 食品), (蔬菜, 食品)
 ((tomato, vegetable), (tomato, food), (vegetable, food))

Structure 2: $(c_1, c_2), (c_2, c_3), (c_3, c_1)$. For example:

(游戏, 生活), (生活, 童话), (童话, 游戏)
 ((game, life), (life, fairy tale), (fairy tale, game))

Structure 3: $(c_1, c), (c_2, c), \dots, (c_m, c), cr(c_1, c_2, \dots, c_m), \exists c_i \in \{c_1, c_2, \dots, c_m\}, CoSuffix(c_i, c)$. For example:

$c=马$, $cr(千里马, 骏马, 老骥, 白驹)$, $CoSuffix(千里马, 马)$, $CoSuffix(骏马, 马)$
 (c=horse, cr(swift horse, courser, nag, white horse), $CoSuffix(swift\ horse, horse)$,
 $CoSuffix(courser, horse)$)

Structure 4: $(c_1, c), (c_2, c), \dots, (c_m, c), cr(c_1, c_2, \dots, c_m), \exists c_i \in \{c_1, c_2, \dots, c_m\}, lpf(c_i, c) \geq 1$. For example:

$c=职业$, $cr(警察, 医生, 歌手, 护士)$, $lpf(歌手, 职业)=3$
 (c=profession, cr(policeman, doctor, singer, nurse), $lpf(singer, profession)=3$)

Structure 5: $(c_1, c), (c_2, c), \dots, (c_m, c), cr(c_1, c_2, \dots, c_m),$
 $(c'_1, c), (c'_2, c), \dots, (c'_n, c), cr(c'_1, c'_2, \dots, c'_n),$
 $\{c_1, c_2, \dots, c_m\} \cap \{c'_1, c'_2, \dots, c'_n\} \neq \emptyset$.

For example:

$c=职业$, $cr(时装, 休闲装, 礼服)$, $cr(裤装, 时装, 休闲服)$ $\{c_1, \dots, c_m\} \cap \{c'_1, \dots, c'_n\} = \{时装, 休闲装\}$
 (c= clothing, cr(fashionable dress, sportswear, full dress), cr(trousers, fashionable dress, sportswear), $\{c_1, \dots, c_m\} \cap \{c'_1, \dots, c'_n\} = \{fashionable\ dress, sportswear\}$)

Structure 6: $(c_1, c), (c_2, c), \dots, (c_m, c), (c'_1, c'), (c'_2, c'), \dots, (c'_n, c'), \{c_1, c_2, \dots, c_m\} \cap \{c'_1, c'_2, \dots, c'_n\} \neq \emptyset$.

For example:

$c=食品$, $\{c_1, \dots, c_m\} = \{牛肉饼, 蛋糕, 面包, 奶油\}$, $c'=产品$, $\{c'_1, \dots, c'_n\} = \{牛肉饼, 牛肉干, 牛肉汤\}$, $\{c_1, \dots, c_m\} \cap \{c'_1, \dots, c'_n\} = \{牛肉饼\}$
 (c= foodstuff, $\{c_1, \dots, c_m\} = \{hamburger, cake, bread, butter\}$, $c'=product$, $\{c'_1, \dots, c'_n\} = \{hamburger, beef\ jerky, brewis\}$, $\{c_1, \dots, c_m\} \cap \{c'_1, \dots, c'_n\} = \{hamburger\}$)

Structure 7: $(c, c_1), (c, c_2), \dots, (c, c_m), (c, c'_1), (c, c'_2), \dots, (c, c'_n)$ $\{c_1, c_2, \dots, c_m\} \cap \{c'_1, c'_2, \dots, c'_n\} \neq \emptyset$.

For example:

$c=西红柿$, $\{c_1, \dots, c_m\} = \{植物, 蔬菜, 食品, 果实\}$, $c'=茄子$, $\{c'_1, \dots, c'_n\} = \{蔬菜, 食品, 食材\}$, $\{c_1, \dots, c_m\} \cap \{c'_1, \dots, c'_n\} = \{蔬菜, 食品\}$
 (c=tomato, $\{c_1, \dots, c_m\} = \{plant, vegetable, foodstuff, fruit\}$, $c'=aubergine$, $\{c'_1, \dots, c'_n\} = \{vegetable, foodstuff, food\ for\ cooking\}$, $\{c_1, \dots, c_m\} \cap \{c'_1, \dots, c'_n\} = \{vegetable, foodstuff\}$)

4 Iterative Hyponymy Verification

4.1 Production Rules

Because the above features are all uncertainty knowledge, they must be converted into a set of production rules used in uncertainty reasoning. Here we use CF (certainty factors) that is the most common approach in rule-based expert system. The CF formula is as follows:

$$CF(CHR, f) = \begin{cases} \frac{P(CHR|f) - P(CHR)}{1 - P(CHR)}, & P(CHR|f) \geq P(CHR) \\ \frac{P(CHR|f) - P(CHR)}{P(CHR)}, & P(CHR|f) < P(CHR) \end{cases} \quad (2)$$

Where CHR is a set of candidate hyponymy, which has a precision $P(CHR)$. $P(CHR|f)$ is the precision of a subset of CHR satisfying feature f . CF is a number in the range from -1 to 1. If there exists $CF(CHR, f) \geq 0$, then we denote f as positive feature and $CF(CHR, f)$ denotes the support degree of feature f ; if there exists $CF(CHR, f) < 0$, then we denote f as negative feature and $CF(CHR, f)$ denotes the no support degree of feature f . We take word-formation feature and space structure feature as example.

If $P(CHR)=0.69$, the precision of candidate hyponymy relations satisfying the feature $|CoSuffix(c_1, c_2)|=c_2$ is 98%, namely $P(CHR|f)=0.98$, then the result of CF is $(0.98-0.69)/(1-0.69)=0.94$. The f is a positive feature.

If $P(CHR)=0.69$, the precision of candidate hyponymy relations satisfying the structure feature 2 is 23%, namely $P(CHR|f)=0.23$, the result of CF is $(0.23-0.69)/0.69=-0.67$. The f is a negative feature.

After those features are converted into a set of production rules, we can carry uncertainty reasoning in concept space.

4.2 Algorithm

The iterative verification of hyponymy can be realized by a production system. The rule database of production system is composed of the above production rules. Here we mainly focus on the control mechanism of the whole iterative verification, that is to say, how to select the rules which can be activated and use these rules to update the CF value of candidate hyponymy. The basic control process is shown in Algorithm 1.

Algorithm 1. The iterative verification of hyponymy

Input: a set of candidate hyponymy relations CHR in concept space, the set of production rules Rule, the initial judgment threshold α , the incremental threshold β , the terminal threshold γ ;

Output: the set of correct hyponymy HR, the set of error hyponymy FR.

Step1: For each hyponymy $r \in CHR$, set its certainty factor $CF(r)$ to be 0;

Step2: For each hyponymy $r \in CHR$, continue Step3 - Step4□

Step3: Find the production rules $rulelist \subseteq Rule$ which r can satisfy.

Step4: Execute all the rules in $rulelist$ and modify the CF of r . The execution orders of rules may lead to the different CF. So these rules can be executed in order according to the descending sort of certainty factors.

Step5: If for each hyponymy $r \in \text{CHR}$, $\text{CF}(r) < \alpha$ is satisfied, then goto Step8;

Step6: For each hyponymy $r \in \text{CHR}$, If $\text{CF}(r) < \alpha$ is satisfied, then $\text{CHR} = \text{CHR} - \{r\}$, $\text{FR} = \text{FR} \cup \{r\}$.

Step7: For each hyponymy $r \in \text{CHR}$, set $\text{CF}(r) = 0$ and goto Step2.

Step8: If $\alpha < \gamma$, then $\alpha = \min\{\alpha + \beta, \gamma\}$, for each $r \in \text{CHR}$, set $\text{CF}(r) = 0$ and goto Step2. If $\alpha \geq \gamma$, then move each $r \in \text{CHR}$ to HR.

Step9: return HR and FR.

In algorithm 1, if $\text{CF}(r) < \alpha$, r is an error hyponymy. When the error hyponymy r is deleted from CHR, the space structure of CHR has changed. So it need to rerun the rules until the number of hyponymy which satisfy $\text{CF}(r) < \alpha$ to be 0. Then the first inside cycle is terminated.

Next, α is updated using β . If $\alpha < \gamma$, continue the next inside cycle. The outside cycle is end until $\alpha \geq \gamma$, and the algorithm terminates. We adopt two-layer cycle instead of directly using γ as the judgment threshold. Because the certainty factors of some correct hyponymy may decrease by error hyponymy which activates the space feature rules. So it is necessary to remove the hyponymy relations with the lowest CF in every cycle. It can reduce the removal of correct hyponymy.

5 Evaluation

5.1 Evaluation Method

We adopt three kinds of measures: R (Recall), P (Precision), and F (F-measure). They are typically used in information retrieval.

Let h be the total number of correct hyponymy relations in the CHR.

Let h_1 be the total number of hyponymy relations in the classify set.

Let h_2 be the total number of correct hyponymy relations in the classified set.

(1) Recall is the ratio of h_2 to h , i.e. $R = h_2/h$

(2) Precision is the ratio of h_2 to h_1 , i.e. $P = h_2/h_1$

(3) F-measure is the harmonic mean of precision and recall. It is high when both precision and recall are high. $F = 2RP/(R+P)$

5.2 Experimental Results

Because the language and corpus is different from other work, it is difficult in the comparison of the proposed method with previous approaches. We used about 15GB of raw corpus from the Chinese Web pages. Raw corpus is preprocessed in a few steps, including word segmentation, part of speech tagging, and splitting sentences according to periods. Then we acquired candidate hyponymy relations CHR from processed corpus by matching Chinese hyponymy patterns. We manually evaluated the initial set CHR and the classified sets. The detailed result is shown in Table 1.

As we can see from table 1, there are 9,609 relations in FR finally after 4 outside cycle and 12 inside cycle. Because FR saves error relations, its recall and precision must be very low. FR has the precision of 7.9% and recall of 1.1%. That is to say, our methods can throw away many error hyponymy relations under the condition of skipping a

Table 1. The result of Iterative Verification

Input: CHR 101,346 P(CHR)=69% $\alpha=-0.8$ $\beta=0.3$ $\gamma=0$			
Threshold		Result	
		CHR	The increment of FR
$\alpha=-0.8$			
	1 cycle	99,804	1542
	2 cycle	99,651	153
	3 cycle	99,620	31
	4 cycle	99,618	2
$\alpha=-0.5$			
	1 cycle	96,994	2,624
	2 cycle	96,728	266
	3 cycle	96,723	5
$\alpha=-0.2$			
	1 cycle	94,166	2,557
	2 cycle	93,753	413
	3 cycle	93,740	13
$\alpha=0$			
	1 cycle	91,748	1,992
	2 cycle	91,737	11
HR: 91,737 (90.5%) P(HR)=75.4%, R(HR)=98.9%, F(HR)=85.6% FR: 9,609 (9.5%) P(FR)=7.9%, R(FR)=1.1%, F(FR)=1.9%			

few correct relations. HR saves correct relations and has the precision of 75.4% and recall of 98.9%. If we want to increase the precision of HR, we can augment γ value. For analyzing the influence of threshold g , we choose several different values. The detailed result is shown in Table 2.

Table 2. The influence of threshold γ ,

Input: CHR 101,346 P(CHR)=69% $\alpha=-0.8$ $\beta=0.3$				
γ	The result of verified hyponymy			
	The number	P	R	F
$\gamma=0$	91,737(90.5%)	75.4%	98.9%	0.856
$\gamma=0.2$	69,023 (68.1%)	80.2%	79.2%	0.797
$\gamma=0.4$	54,105(53.4%)	86.7%	67.0%	0.756
$\gamma=0.6$	29,293(28.9%)	91.2%	38.2%	0.538
$\gamma=0.8$	20,370(20.1%)	94.3%	27.5%	0.426

As we can see from table 2, there are 101,346 hyponymy relations initially. With the increase of threshold γ , the precision is also increase. If we want to increase the precision, we can augment γ value. For example, when $\gamma=0.8$, the precision is up to 94.3%, and but its recall decreases to 27.5%. That is to say, when threshold γ is a small value, our methods can throw away many error hyponymy relations under the condition of skipping a few correct relations. But when threshold γ is a large value, our methods can throw away many error hyponymy relations and also skip many correct relations at same time.

6 Conclusion

In this paper, we present an iterative method of hyponymy verification based on concept space. It initially acquires a set of hyponymy features between Chinese terms based on the structure of hyponymy. Experimental results demonstrate good performance of the method. Our methods can throw away many error hyponymy relations under the condition of skipping a few correct relations. It will benefit to the building of ontologies, knowledge bases and lexicons.

There are still some inaccurate relations in the result. It is necessary for us to solve some problems, such as the polysemy and synonymy of concept word, the relativity of context of hyponymy and so on. These problems maybe cause incorrect space structure among hyponymy relations. So more sophisticated verification methods are needed. In future, we will combine some methods (such as word sense disambiguation, the morpheme analysis, web page tag etc.) to the further verification of hyponymy.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant No.60573064, and 60773059; the National 863 Program under Grant No. 2007AA01Z325, and the Beijing University of Technology Science Foundation (grant nos. X0006014200803, 97006017200701).

References

1. Beeferman, D.: Lexical discovery with an enriched semantic network. In: Proceedings of the Workshop on Applications of WordNet in Natural Language Processing Systems, ACL/COLING, pp. 358–364 (1998)
2. Cao, C., Shi, Q.: Acquiring Chinese Historical Knowledge from Encyclopedic Texts. In: Proceedings of the International Conference for Young Computer Scientists, pp. 1194–1198 (2001)
3. Hearst, M.A.: Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database and Some of its Applications, pp. 131–153. MIT Press, Cambridge (1998) (to appear)
4. Maedche, A., Pekar, V., Staab, S.: Ontology Learning Part One—On Discovering Taxonomic Relations from the Web. In: Web Intelligence, pp. 301–322. Springer, Heidelberg (2002)
5. Caraballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 120–126 (1999)
6. Maedche, A., Pekar, V., Staab, S.: Ontology Learning Part One—On Discovering Taxonomic Relations from the Web. In: Web Intelligence, pp. 301–322. Springer, Heidelberg (2002)
7. Sánchez, D., Moreno, A.: Patterned automatic taxonomy learning from the Web. *AI Communications* 21(3), 27–48 (2008)
8. Elghamry, K.: Using the Web in Building a Corpus-Based Hypernymy-Hyponymy Lexicon with Hierarchical Structure for Arabic. Faculty of Computers and Information, pp. 157–165 (2008)

9. Zhang, C.-x., Hao, T.-y.: The State of the Art and Difficulties in Automatic Chinese Word Segmentation. *Journal Of System Simulation* 17(1), 138–143 (2005)
10. Mei, J.J., Zhu, Y.M., Gao, Y.Q., Yin, H.X.: *Tongyici Cilin (Dictionary of Synonymous Words)*. Shanghai Cishu Publisher, China (1983)
11. Tian, G., Cao, C., Liu, L., Wang, H.: MFC: A Method of Co-referent Relation Acquisition from Large-scale Chinese Corpora. In: *ICNC 2006-FSKD 2006*, Xi'an, China (2006)

Sharing Mobile Multimedia Annotations to Support Inquiry-Based Learning Using MobiTOP*

Khasfariyati Razikin¹, Dion Hoe-Lian Goh¹, Yin-Leng Theng¹,
Quang Minh Nguyen¹, Thi Nhu Quynh Kim¹, Ee-Peng Lim², Chew Hung Chang³,
Kalyani Chatterjea³, and Aixin Sun⁴

¹ Wee Kim Wee School of Communication & Information, Nanyang Technological University
{khasfariyati, ashlgoh, tyltheng, qmnguyen, ktng}@ntu.edu.sg

² School of Information Systems, Singapore Management University
eplim@smu.edu.sg

³ National Institute of Education, Nanyang Technological University
{chewhung.chang, kalyani.c}@nie.edu.sg

⁴ School of Computer Engineering, Nanyang Technological University
axsun@ntu.edu.sg

Abstract. Mobile devices used in educational settings are usually employed within a collaborative learning activity in which learning takes place in the form of social interactions between team members while performing a shared task. We introduce **MobiTOP** (**M**obile **T**agging of **O**bjects and **P**eople), a geospatial digital library system which allows users to contribute and share multimedia annotations via mobile devices. A key feature of MobiTOP that is well suited for collaborative learning is that annotations are hierarchical, allowing annotations to be annotated by other users to an arbitrary depth. A group of student-teachers involved in an inquiry-based learning activity in geography were instructed to identify rock types and associated landforms by collaborating with each other using the MobiTOP system. The outcome of the study and its implications are reported in this paper.

1 Introduction

As mobile devices increase in popularity and functional features, it not surprising that they have been adopted for use in education [17]. In such settings, they are usually employed within a collaborative learning activity where learning takes place in the form of social interactions between the team members while executing a shared task [6]. Mobile devices are suited for collaborative learning because they allow students to take control of the hardware without being impeded by cumbersome instruments [5]. The learning activity could take place indoors (e.g. [21]) and/or outdoors (e.g. [16]) and could be highly relevant to subjects such as geography (e.g. [19]), biology (e.g. [20]) and history (e.g. [4]).

Mobile devices are best used in learning situations as a tool to support group activity as such learning activities involve sharing of students' interpretations of the situation

* This work is partly funded by A*STAR grant 062 130 0057.

and the environment with one another [18]. These devices make it possible for students to communicate with others and engaging with the environment without the need to constantly look at the screen [14], thus facilitating collaborative learning.

Within the area of geography inquiry, mobile devices have been deployed in field-work studies. Fieldwork has become an essential part of geography learning as it enables students to apply what they have learnt in the classroom in authentic outdoor settings [11] by active engagement and collaboration. Indeed, both classroom learning and field-based learning are complementary in geography education, to enrich the student's learning experience [3]. Moreover, students can use mobile devices to conduct social interactions that are no longer confined to those in the field but also extended to those who are at other locations (e.g. [1], [7]).

In this paper, we introduce **MobiTOP (Mobile Tagging of Objects and People)**, a geospatial digital library system that allows users to contribute and share geospatial multimedia annotations. A key feature of MobiTOP suitable for collaborative learning is the hierarchical annotations that allow annotations to be annotated by other users to an arbitrary depth, essentially, creating threads of discussions. MobiTOP served as a platform for a geography study conducted by a group of student-teachers. The goal was to identify rock type and associated landforms for an assignment. The students communicated through the hierarchical annotations afforded by MobiTOP, each comprising textual information and images. We highlight the experiences of the students and the lessons learnt. Data was collected from observations made during the study and questionnaires that were distributed to the students after the exercise.

The remainder of this paper will elaborate on the design, development, deployment and evaluation of MobiTOP. Section 2 presents the MobiTOP system. This will be followed by the description of the geography inquiry. Sections 4 and 5 highlight the observations made and the results of the evaluation from the study. The paper concludes with a discussion on the implications of our findings together with the possible areas of future work.

2 The MobiTOP System

MobiTOP is built upon an earlier geospatial digital library system known as G-Portal ([10], [12]) which supports the identification, classification and organization of geospatial and geo-referenced content on the Web, and the provision of digital services such as searching and visualization. MobiTOP offers an updated AJAX-based user interface as opposed a Java-applet interface to facilitate more widespread use, and enhanced mobile user interfaces. Another key difference is its hierarchical multimedia annotation support which allows users to create, share and organize media-rich annotations any-time, anywhere. In MobiTOP, annotations consist of locations, images and other multimedia, as well as textual details augmented by tags, titles and descriptions. Tags are freely assigned keywords [15] that are not limited by any taxonomy, ontology or controlled vocabularies. MobiTOP employs a client/server architecture (Figure 1) and consists of a single server and two independent mobile and web clients. At the server, the annotation database stores the annotations with the georeferenced locations and attached media.

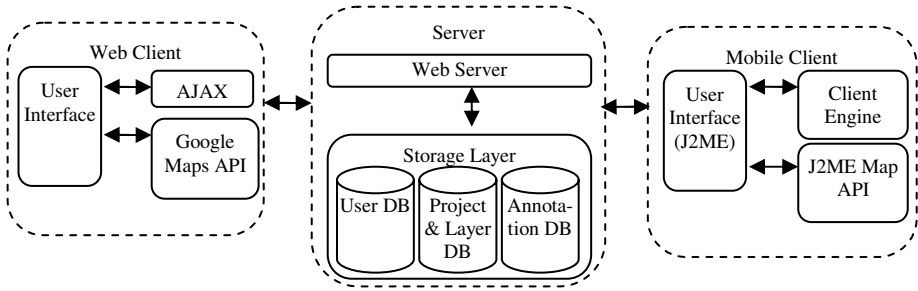


Fig. 1. Architecture of MobiTOP System

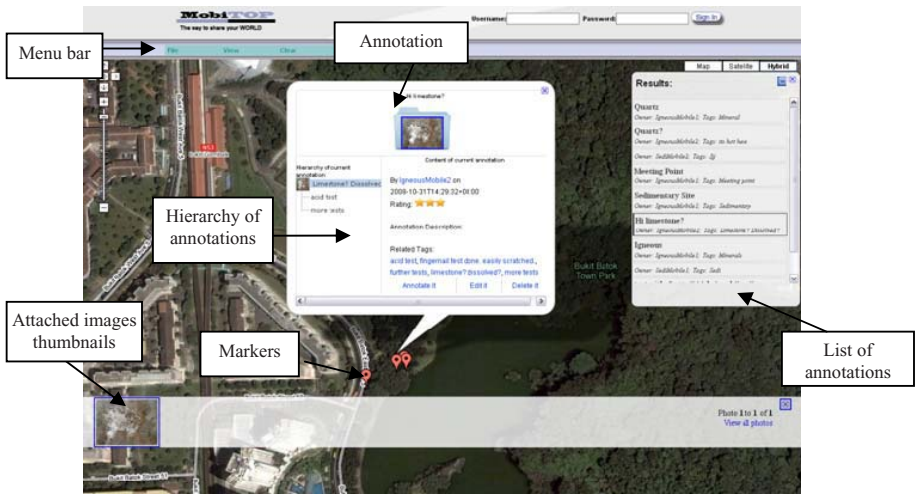


Fig. 2. User Interface on the web client

The Web client displays a map-based visualization for exploring contributed annotations using the Google Maps™ API. It also uses an AJAX library to enrich the user experience with sophisticated interactions. In Figure 2, the top bar displays the different functions available to the users, such as viewing all annotations on the map, and enabling the notification of updates. These georeferenced annotations consist of title, tags, description and photos in addition to the latitude and longitude. Each marker on the map represents a root annotation, and indicates that there is at least one annotation on that location. The panel on the right lists the available root annotations. Selecting one of the annotations in the list will display the content of the annotation and the sub-annotations associated with it on the map. Sub-annotations are displayed in a modified tree-view structure which displays only three levels of the hierarchy, consisting of the current annotation, parent and children, at any one time. This design is meant to reduce clutter on the interface as well as minimize information overload when many annotations are contributed. Further, a folder icon indicates that there are images attached to the annotation. Thumbnails of the attached images are shown in a

banner displayed at the bottom of the map. The user interface was designed based on the outcome of a participatory design workshop where potential end users of MobiTOP took part as designers [9].

MobiTOP’s mobile client supports a map-based visualization for exploring hierarchical geospatial annotations and location-based mobile annotating. The mobile client was primarily developed for Nokia N95 8GB smart phones (Figure 3). The client uses the global positioning system (GPS) feature available in the phone to determine the current location of the user.

The client’s functions are logically organized into tabs (Figure 3). The left and right direction keys are used to navigate between the tabs. Users are able to create root annotations by accessing a form directly or by selecting a location on the map. At the form, the user can capture images using the phone’s camera or select an existing image to be attached to the annotation. In order to annotate an existing annotation, the user selects the parent from the hierarchy before creating a new child annotation.



Fig. 3. MobiTOP mobile client



Fig. 4. Map interface of mobile client

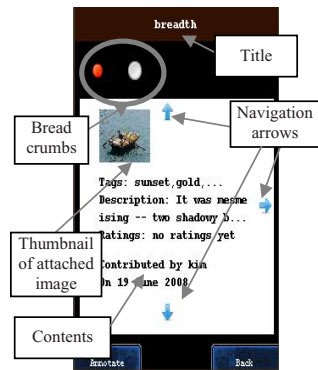


Fig. 5. Mobile client hierarchical visualization

Similar to the Web client, the mobile client displays annotations in a hierarchy. However, the challenge is to fit the information-packed visualization onto a small screen of the mobile phone. At the same time, the design should support a seamless integration with the Web client. Here, the design for the mobile client was again based on the outcome of a participatory design workshop [9]. Figure 4 shows the map-based visualization with a summary of a selected root annotation, while Figure 5 shows the interface for navigating a hierarchy of annotations. Users can use the navigation keys to open other annotations associated with this current annotation: parent (Up), children (Down), sibling (Left/Right). Based on existing mental models of navigation, users would be able to map the association of using the direction keys for navigating to the other annotations [2]. Bread crumbs are used to indicate the current level of the annotation in the hierarchy helping users to keep track of their locations in the annotation hierarchy. Here, the first dot denotes the root level, while subsequent dots indicate the respective level in the hierarchy.

3 Geographical Study of Rocks Using MobiTOP

A class of geography student-teachers from a teacher training institute was involved in the field study. One of their learning goals was to examine the type of rocks at a given location. Additionally, they had to explain the landform which was present at the site by identifying the type of rocks. The students were divided into two groups which were to identify two different types of rocks: igneous and sedimentary. Each group was assigned the task of identifying one of the rock types and comprised of field investigators and lab investigators. The field investigators were to go to the actual site to collect information and collaborate with the lab investigators who were in the classroom. Altogether 12 students were involved. They were divided evenly into the two groups. Four members of each group were assigned to be field investigators and the rest were lab investigators. The students' ages ranged from 23 to 26 and consisted of three males and nine females.

The students were introduced to the MobiTOP system prior to the actual field study. The concept of hierarchical annotations and their possible use in the fieldwork study were also explained to them. The roles and their respective tasks were highlighted as well. In the actual fieldwork exercise, four mobile devices were deployed, with each field team having two mobile phones. The field investigators within each group were not told of the specific types of activities they had to perform to meet the objectives of the study. Tools, such as a geologic hammer, magnifying glass and a small bottle of diluted hydrochloric acid, were also provided.

The lab investigators were stationed at the geography lab at the institute. Communication between the field and lab investigators was achieved through the MobiTOP system. The field investigators in each group collaborated, utilized the tools on the rocks and created annotations of their findings before uploading them to the MobiTOP server. As part of the creation of annotations, they took photographs as evidence to substantiate their findings. When the lab investigators received the new annotations, they replied with possible tests that could be easily conducted, description of the rocks, or the possible types of rocks. At both field and lab, the students' interactions with the application were recorded for later analysis.

Finally, based on the information gathered from the fieldwork, the group members were to consolidate and present their findings. Questionnaires were distributed to evaluate the application afterwards. The different groups of investigators were given different sets of questionnaire depending on the type of application that they had used (mobile or Web). The first part of the survey consisted of questions related to the demographic profile and the frequency of using various Web and mobile services. The second part asked the students to rate the usability and usefulness of the applications that they had used during the fieldwork. In this section, they were also asked open ended questions about the applications' features.

4 Observations

Video recordings that were made during the study served as a guide to elicit the students' usage patterns of the MobiTOP system. The recordings also provided evidence on the student's behaviors. Here, the recordings were analyzed and categorized according to various themes, described in the following sections.

4.1 Aesthetics and Layout

- **Color and icons.** Investigators commented that both mobile and Web clients had aesthetically pleasing designs. The Web client's main focus was the map and it seemed that the lab investigators were familiar with map navigation. For the mobile client, the icons that appear on the navigation tab made it easy for the field investigators to understand the available functions.
- **Tabbed navigation.** The different functions in the mobile client were organized using tabs. This made it easy for the field investigators to navigate to the different functions available. Tabbed navigation was proposed as all the tabs fit into one row of the screen so that it remained clutter free. At the same time, more controls could be fit into a single screen.

4.2 Navigation/Browsing of Annotations

- **Notification of annotation updates.** It was observed that the lab investigators were not aware of new annotations that were uploaded by the field investigators. This is because the original design of the Web client does not update the annotations on the map automatically. Instead, the user is required to click on an update notification message to refresh the list of annotations. This was done to give the users more control over the visualization. Specifically, a constraint in the Google Maps API meant that if automatic updates were supported, any open annotation window would be closed whenever the map was refreshed with new annotation markers, possibly causing surprise to users in the midst of reading an annotation. The semi-automatic update method was meant to let users choose when they needed to see the updates. However, it appears that this update notification was not obvious to the lab investigators as they asked the research team how they would know if there were updates repeatedly. A more prominent update notification is therefore needed.
- **Visibility of updates.** Newly created sub-annotations were not apparent to both users of mobile and Web clients when replying to annotations. There was no indicator on the root annotations stating that there were new replies available. Instead, the investigators had to navigate through all the annotations in the client to check for replies. As a result, some annotations were missed and were not replied to. This occasionally led to miscommunication between the field and lab investigators as the missed annotations contained pertinent findings for their activity. At other times, they were frustrated as they had to browse through the annotations to find that no updates had been received. Similarly, there were no markers or labels for updated/new annotations in the Web client because adding more than three levels of a hierarchy would increase the number of horizontal scrollbars in the left column of the annotation window (Figure 2), compromising the aesthetics and balance of the annotation window. Further, the mobile client did not have any markers or labels to indicate newly created annotations on the map due to the limitations of the J2ME Map API which did not support additional labels to be overlaid on the map markers. Thus, only notifications of updates could be made via status messages.

4.3 Creating Annotations

- **Uninformative titles and tags.** Titles and tags created were sometimes not informative. One of the root annotations had the title, “Hi limestone?” which was not appropriate. Likewise for tags, some annotations were assigned terms such as “its hot here”, “dissolved?” and even ‘.’! This was probably due to unfamiliarity with social tagging as ten of the students reported to have not often used social bookmarking sites. Annotations with such titles and tags were more prevalent in those created by the field investigators. One possible reason was that the investigators could be new to such fieldwork and the instructor was not supposed to provide guidance. Another reason could be due to the limited input entry on the mobile phone, which will be explained in a subsequent section.
- **Notification of upload progress.** Due to slow network connections, the mobile client was not very responsive at times, especially during uploading. Here, a notification with the status of the upload (success or failure) is the only response the field investigator will see. The lag time between the selection to upload and the appearance of this notification caused some confusion as they did not know if any activity was taking place. For instance, the field investigators often had to be reassured by either the onsite researcher or their professor that the annotations that they had created were in the process of being uploaded.
- **Poor image quality.** Some images taken by the field investigators were not clear, which led to frustration of the lab investigators. With the poor image quality, the lab investigators were not able to compare with the rock samples in the laboratory. At the same time, they were not able to discern the characteristics of the rocks from the images. The camera function in the mobile client had basic functions but had no sophisticated features such as zooming.

4.4 Unfamiliarity with Mobile Phones

- **Multiple special keys.** As the mobile phone was equipped with multiple special keys for various features such as navigation, the field investigators often unintentionally pressed keys leading to unexpected results. For example, some investigators inadvertently pressed the mobile phone’s “application” key causing MobiTOP to be hidden in the background. In general, it was difficult for the field investigators learn the purposes of the various keys in such a short time. This often led to errors that required the onsite researchers to intervene.
- **Poor affordances for data entry.** The field investigators were told to avoid using typical SMS short forms. This led to some frustration as they had to spend time keying in the full text of the annotations. Another factor that contributed to this observation was that none of the investigators had prior experience in using this particular mobile phone model (which is a high-end model at the time of this writing). Their unfamiliarity with the mobile phone made the students feel the keypad was restrictive, hence hindering their efficiency.

5 Usability Evaluation

We adopted the heuristic evaluation approach to ascertain MobiTOP’s usability. Heuristic evaluation identifies usability problems by allowing the evaluators to examine

the interface and then proceeding to make judgments to its compliance to the heuristics. In the questionnaire survey, investigators were asked questions based on Nielsen's 10 usability heuristics [13]. Additionally, the students were also asked to evaluate the system in terms of effectiveness in contributing to team collaboration.

Table 1 shows the results of the usability evaluation and Table 2 shows the results of the team collaboration evaluation. Both tables show the mean and the standard deviation of the values obtained for both Mobile and Web clients. In Table 1, the students found the mobile application to be relatively usable in general. For instance, the terms used in the mobile client were familiar to frequent mobile phone users. They thus had a sense of recognition of the features of the client, and thus mentally map these features to their expectations easily. Likewise, the consistency of the user interface helped in the learnability of the system for some of the students. Similarly, the lab investigators found that language used and the layout in the application to be consistent. For example, the menu labels (e.g. "File", "View") were consistent with the conventions of a typical web application. They also felt that the labels used were understandable. Additionally, they felt that the map-based Web client was quite intuitive. This could be due to their familiarity with Google Maps as three of the lab investigators used Web based mapping applications somewhat frequently.

Table 1. Usability evaluation results (1 = strongly disagree; 5 = strongly agree)

No.	Heuristic	Mobile client		Web client	
		Mean	SD	Mean	SD
1.	Visibility of system status	1.75	0.71	2.50	1.29
2.	Match between system and the real world	3.00	1.29	3.25	1.50
3.	User control and freedom	2.50	1.07	2.50	1.00
4.	Consistency and standards	3.50	1.16	3.88	0.90
5.	Error prevention	2.71	0.95	2.00	1.15
6.	Recognition rather than recall	2.25	1.04	3.00	1.15
7.	Flexibility and efficiency of use	1.88	0.83	1.25	0.50
8.	Aesthetic and minimalist design	2.38	0.74	3.00	1.15
9.	Help users recognize, diagnose and recover from errors	2.75	0.71	3.00	1.15
10.	Help and documentation	2.75	1.28	3.00	0.82

In terms of usefulness for team collaboration, students felt that it was somewhat easy to view and create annotations. Perhaps factors like the phone's affordances and difficulty in the keying in of input contributed to the lower scores. However, they reported that that the mobile client allowed them to take photographs easily. One issue that emerged was that the field investigators were not able to communicate with the lab investigators easily. As highlighted, this was probably attributed to difficulty in finding the new annotations. On the other hand, the lab investigators were able to find the annotations that they needed easily. The list of root annotations in the right panel (Figure 2) probably helped them locate the newly uploaded root annotations quickly. They felt that notifications of new annotations could be improved by supporting automatic updates.

The students were also asked how they felt about the task and the applications. The open-ended questions elicited their opinions about the potential of the MobiTOP system as a learning tool. Additionally, they were also asked about useful features and those which could be improved.

Table 2. Usefulness for team collaboration results (1= strongly disagree; 5 = strongly agree)

No.	Actions	Mobile client		Web client	
		Mean	SD	Mean	SD
1.	Create annotations easily	2.63	1.06	2.88	0.66
2.	View annotations easily	2.75	0.89	2.25	0.96
3.	Take photographs easily	3.88	0.83	2.25	1.26
4.	Find annotations that you need easily	2.14	0.38	3.50	0.58
5.	Notifies new annotations created by other investigators	1.88	0.83	2.00	0.82
6.	Communicate with the lab investigators easily	1.75	0.71	2.25	0.96

Mobile Client. In general, the students appreciated the real life aspects of the field-work activity as it gave them the opportunity to discover for themselves facts which are not found from textbooks. The students were divided in their opinions in terms of the potential of the application as a learning tool. Half saw the potential while the other half did not feel the same way. From the perspective of one student who concurred, the application was helpful for fieldwork as the mobile phone was equipped with useful functions (“considering the availability of GPS tools and Internet on the mobile phone, it can be very helpful”). Other views included that MobiTOP was a useful for them to share and collaborate with other users. In contrast, those who did not agree felt that the problems encountered during the usage of the application marred its potential as a learning tool (“There are too many problems and errors with the tool”). These problems include the usability of the hierarchical annotations (“Only if the annotations appeared in a user friendly manner”) and the latency of retrieving annotations (“(only) if the time taken to be reflected (on the map) is shorter”). The fact the MobiTOP system is still in the prototype stage can account for many of these issues.

One of the features that students found favorable was MobiTOP’s built-in camera feature for annotations. Students were able to take images without changing the orientation of the mobile phone and by pressing a button on the keypad instead of the camera shutter button. This helped them capture images easily as their other hand might be occupied with notebooks or tools. Another feature that students found useful was the ability to view annotations on the map. This helped them better understand natural rock formations as they were able to pinpoint the exact location of the sites and observe that which were quite close had different types of rocks.

However, the constraints of the mobile phone did pose problems. One student felt that the mobile phone’s screen was too small to locate annotations easily. This sentiment was echoed by several other students (“the phone is too small an interface (to display a map)”). As the annotations required the user to press a button to update the annotations, one student felt that the “annotations should be refreshed by itself ... (for example) appear(ing) like a text message.” Another student suggested that “a more sophisticated platform will be better, maybe a Blackberry.”

Web Client. The lab investigators felt that they could only slightly appreciate the context of the work. One of the reasons for this outcome was that they were not able to see the rocks in their natural formations making their learning somewhat ineffective. Another contributing factor could be due to the lag time between receiving the notifications of new annotations. They spent their time discussing about the findings of their classmates while waiting for the outcome of their instructions.

When asked about the application as a potential learning tool, again, opinions were divided. Some felt that it had such potential (“technology is highly used in school(s) and such tool could be used for field work like what was done”). The rest felt that the application was “too tedious and time consuming”. This was primarily attributed to the lack built-in communication facilities to enable lab investigators to track the status of the field investigators. For example, the lab investigators kept checking for updates while the field investigators were making their way to a different site. This led one of the lab investigators to remark that she had already lost interest in the activity.

Despite the problems, students found that being able to create annotations and placing them on the map was a useful feature. This demonstrates that the students were able to appreciate the information creation and sharing aspect of the MobiTOP system. Another positive feature noted by the investigators was the ability to associate images with annotations, which they found useful for learning. Perhaps they understood that augmenting the annotations with photos would enrich the user experience. Others commented that they felt the application enabled them to communicate with the field investigators in an almost synchronous manner. Finally, some of the features in the Web client that students would like to see improvements on are automatic updates of annotations (“easier notification of new annotations”) as well as easier navigation between root annotations. Another crucial point made is that they would like to see an improvement in the ability to communicate with the field investigators easily (“communication with the field group should be made easier”).

6 Discussion and Conclusion

This paper introduces the MobiTOP system and describes the outcomes of its use in a geography study. From our investigations, three main findings emerge. First, training is essential when introducing a new technology. In our study, some of the field investigators were not familiar with the terms used and the mobile phone model. Although all students were experienced mobile phones users, they needed some time to get themselves acquainted with the phone’s keys and functions, mainly because of the usability problems found in smart phones [8]. This usability issue is often the product of the increasing number of features at the expense of usability. As the students did not own this particular phone model, problems during the fieldwork activity emerged. Perhaps the familiarization activity could be a take home exercise where students experimented with the phone over a longer period of time.

Secondly, the affordances of the mobile device should be taken into consideration when developing applications [14]. In our case, textual input should be minimized to counteract uninformative titles and tags. Both titles and tags are intended to give an overview of the annotations so that other users would be able to quickly understand the contents of the annotation. For example, titles of the root annotation could be automatically reused by the sub-annotations so that there is a flow in the thread. Further, the client could suggest tags based on an analysis of other annotations in the same thread, or on annotations nearby. However, users should have the ability to override these default values to meet their specific needs.

Thirdly, based on our observations of the investigators and their feedback in the survey, the following design lessons can be drawn from this study:

- Provide timely and informative updates in dynamic, information rich environments. In MobiTOP, indicating new contributions to the system would enable users to make serendipitous discoveries of information which meets their needs. Further, indicating new contributions in a thread of annotations is equally as important. This is because users interested about the thread's topic would be able gain new perspectives.
- Provide adequate and informative feedback to users' actions in interactive systems for tasks that require time to complete. Users expect a timely response to their actions in such systems so as to ensure that they would know the next step to take in order to fulfill their goals.
- Additional communication channels may be helpful. Apart from the asynchronous communication support afforded by the hierarchical annotations in MobiTOP, some users in the geography study commented that synchronous communication modes would be useful as well.

References

1. Bergin, D.A., Anderson, A.H., Molnar, T., Baumgartner, R., Mitchell, S., Korper, S., Curley, A., Rottmann, J.: Providing remote accessible field trips (RAFT): an evaluation study. *Comput. Hum. Behav.* 23, 192–219 (2007)
2. Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G., Pazzani, M.: Improving mobile internet usability. In: *Proceedings of the 10th international conference on World Wide Web*, pp. 673–680. ACM, New York (2001)
3. Chang, C.H., Ooi, G.L.: Role of Fieldwork in Humanities and Social Studies Education. In: Tan, O.S., McInerney, D.M., Liem, A.D., Tan, A.G. (eds.) *What the West can learn from the East. Asian Perspectives on the Psychology of Learning and Motivation. Research in Multicultural Education and International Perspectives Series*, vol. 7, pp. 295–312. Information Age Publishing, Charlotte (2008)
4. Costabile, M.F., Angeli, A.D., Lanzilotti, R., Ardito, C., Buono, P., Pedersen, T.: Explore! Possibilities and challenges of mobile learning. In: *Proceedings of the 26th annual SIGCHI conference on Human factors in computing systems*, pp. 145–154. ACM, New York (2008)
5. Danesh, A., Inkpen, K., Lau, F., Shu, K., Booth, K.: Designing a collaborative activity for palmTM handheld computer. In: *Proceedings of CHI Conference on Human Factors in Computing Systems*, pp. 388–395. ACM, New York (2001)
6. Dillenbourg, P.: What do you mean by collaborative learning? In: *Collaborative-learning: Cognitive and Computational Approaches*, pp. 1–19. Elsevier, Oxford (1999)
7. Haapala, O., Sääskilathi, K., Luimula, M., Yli-Hemminki, J., Partala, T.: Parallel Learning between the Classroom and the Field using Location-Based Communication Techniques. In: Montgomerie, C., Seale, J. (eds.) *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007*, pp. 668–676. AACE, Vancouver (2007)
8. Keijzers, J., Ouden, E.d., Lu, Y.: Usability benchmark study of commercially available smart phones: cell phone type platform, PDA type platform and PC type platform. In: *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pp. 265–272. ACM, New York (2008)
9. Kim, T.N.Q., Razikin, K., Goh, D.H.-L., Theng, Y.L., Nguyen, Q.M., Lim, E.-P., Sun, A., Chang, C.H., Chatterjea, K.: Exploring Hierarchically Organized Georeferenced Multimedia Annotations in the MobiTOP System. In: *Proceedings of the 6th International Conference on Information Technology: New Generations*, pp. 1355–1360. IEEE, California (2009)

10. Lim, E.-P., Goh, D.H.-L., Liu, Z., Ng, W.-K., Khoo, C.S.-G., Higgins, S.E.: G-Portal: a map-based digital library for distributed geospatial and georeferenced resources. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, pp. 351–358. ACM, New York (2002)
11. Lim, E.-P., Sun, A., Liu, Z., Hedberg, J., Chang, C.H., Teh, T.-S., Goh, D.H.-L., Theng, Y.L.: Supporting field study with personalized project spaces in a geographical digital library. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E., Lim, E.-p. (eds.) ICADL 2004. LNCS, vol. 3334, pp. 553–562. Springer, Heidelberg (2004)
12. Liu, Z., Yu, H., Lim, E.-P., Ming, Y., Goh, D.H.-L., Theng, Y.L., Ng, W.-K.: A Java-based digital library portal for geography education. *Sci. Comput. Program.* 53, 87–105 (2004)
13. Nielsen, J.: Finding usability problems through heuristic evaluation. In: Proceedings of the Conference on Human Factors in Computing System, pp. 373–380. ACM, New York (1992)
14. Patten, B., Sanchez, I.A., Tangney, B.: Designing collaborative, constructionist and contextual applications for handheld devices. *Comput. Educ.* 46, 294–308 (2006)
15. Razikin, K., Goh, D.H.-L., Chua, A.Y., Lee, C.S.: Can Social Tags Help You Find What You Want? In: Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries, pp. 50–61. Springer, Berlin (2008)
16. Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith, H., Randell, C., Muller, H., O' Malley, C., Stanton, D., Thompson, M., Weal, M.: Ambient wood: designing new forms of digital augmentation for learning outdoors. In: Proceedings of the conference on Interaction design and children: building a community, pp. 3–10. ACM, New York (2004)
17. Roschelle, J.: Unlocking the learning value of wireless mobile devices. *J. Computer Assisted Learning* 19, 260–272 (2003)
18. Sharples, M., Corlett, D., Westmancott, O.: The Design and Implementation of a Mobile Learning Resource. *Personal and Ubiquitous Computing* 6, 220–234 (2002)
19. Theng, Y.L., Tan, K.-L., Lim, E.-P., Zhang, J., Goh, D.H.-L., Chatterjea, K., Chang, C.H., Sun, A., Yu, H., Dang, N.H., Li, Y., Vo, M.C.: Mobile G-Portal supporting collaborative sharing and learning in geography fieldwork: an empirical study. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 462–471. ACM, New York (2007)
20. Yeh, R.B., Chunyuan, L., Klemmer, S., Guimbretière, F., Lee, B., Kakaradov, B., Stamberger, J., Paepcke, A.: ButterflyNet: a mobile capture and access system for field biology research. In: Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 571–580. ACM, New York (2006)
21. Zurita, G., Nussbaum, M.: Computer supported collaborative learning using wirelessly interconnected handheld computers. *Comput. Educ.* 42, 289–314 (2004)

Understanding Perceived Gratifications for Mobile Content Sharing and Retrieval in a Game-Based Environment*

Chei Sian Lee¹, Dion Hoe-Lian Goh¹, Alton Y.K. Chua¹, and Rebecca P. Ang²

¹ Wee Kim Wee School of Communication and Information

² School of Humanities and Social Sciences

Nanyang Technological University

{leecs, ashlgoh, altonchua, rpang}@ntu.edu.sg

Abstract. The confluence of mobile content sharing and pervasive gaming yields new opportunities for developing novel applications on mobile devices. Yet, studies on users' attitudes and behaviors related to mobile gaming, content sharing and retrieval activities have been lacking. For this reason, the objectives of this paper are two-fold. One, it proposes MARGE, a game which incorporates multiplayer, role-playing pervasive gaming elements into mobile content sharing activities. Two, it seeks to uncover the motivations for content sharing and content retrieval within a game-based environment. Informed by the uses and gratifications paradigm, a survey was designed and administered to 163 graduate students from two large universities. The findings revealed that perceived gratifications factors related to self, personal status and relationship maintenance were significant predictors for content sharing while information quality was a significant predictor for content retrieval. This paper concludes by presenting the implications, limitations and future research directions.

1 Introduction

The increasing popularity of mobile devices and their wireless networking capabilities offer new opportunities for social computing applications to be deployed on them, fostering interaction, content generation and participation amongst communities of users. The portability of mobile devices add a new dimension to user-generated content in which users can now co-create, seek and share information anytime, anywhere. Twitter, for example, is a micro-blogging and social networking service that allows content in the form of brief text messages to be created via a mobile device and posted on a user's Twitter page. Beyond ubiquitous content creation, context-aware, location-based information services are also now possible in mobile content sharing, allowing users to associate digital content with physical objects and locations in the real world, as well as receive content tailored to their specific needs [8]. Examples include ZoneTag [1] for photo sharing and Magitti [3], a context-aware mobile tour guide.

* This work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research Grant NRF NRF2008IDM-IDM004-012.

Mobile content sharing applications allow users to co-create, seek and share multimedia content such as text, audio and video, socialize anytime, anywhere, and do these in new ways not possible with desktop applications. Despite these benefits, a potential drawback could limit a more widespread acceptance of their use. For example, the motivations for creating and sharing content are mostly intrinsic to users, and may include both social (e.g. getting attention) and personal (e.g. future retrieval) reasons [1], utilitarian or opportunistic behaviors [7], altruism and social exchange norms [17]. Likewise, the motivations for retrieving and consuming user-generated may include trust and reputation [4, 20], characteristics of the message [10], and characteristics of the creator, consumer and community [5, 13]. Put differently, current content sharing applications only provide limited extrinsic motivational mechanisms, and are typically confined to viewership counts, content/user ratings and discussion facilities.

Mobile devices also add a new aspect to play. Mobile games have evolved from casual games such as *Snake* to sophisticated multiplayer location-based ones in which players either compete and/or cooperate to achieve the games' objectives within a geographic area set in the real world. Also known as pervasive games, examples include early commercial successes such as *Botfighters* in which players take on the role of robots and search and destroy other robots within the vicinity, and *Undercover 2* where players visit real cities around the world, complete missions, and look for friends, enemies, and landmarks in real streets.

The confluence of social computing, mobile content sharing, and pervasive gaming yields new opportunities for developing novel, engaging applications for content sharing on mobile devices that can address the lack of extrinsic motivational mechanisms identified above. In particular, a central theme of these new applications is that content is created and shared as a byproduct of gameplay, and the gaming experience becomes an extrinsic motivator for content sharing activities. In addition, many of these games are social in nature, requiring multiple players to achieve the game's objectives. One example of a mobile game for data collection is the Gopher Game [6]. Gophers are agents that represent missions to be completed, and are carriers of information between players. As players move about their physical surroundings, they pick up gophers and help them complete their missions by supplying them with camera phone images and textual content. By helping gophers complete their missions, content sharing among players is achieved since other users may pick up these gophers and view the images and text associated with them. Such games are inspired by the success of Web-based casual games such as the ESP Game [22] and Google Image Labeler that use humans to label images to facilitate future retrieval.

While such games for mobile content sharing have their advantages, three possible shortcomings may be identified. First, the reward systems in existing games tend to be basic, and are often similar to existing mechanisms found in current mobile content sharing systems, including reviews and ratings [e.g. 6]. Next, current mobile content sharing games may be characterized as casual games that have simple rules, allowing gameplay to occur in short bursts [2]. The lack of complex game mechanics and depth of play may eventually cause the novelty of the game to lose its luster. Third, there is very little work done in determining if and how these games motivate users to share content, and if users are willing to retrieve and consume such content generated from gameplay. We argue that a better understanding of users' attitudes and behaviors are

necessary to implement systems supporting mobile content sharing and retrieval activities more effectively.

The objectives of this research are two-fold. The first is to address the shortcomings of current mobile content sharing games by proposing MARGE (Mobile Alternate Reality Gaming Engine). MARGE incorporates multiplayer, role-playing pervasive gaming elements into mobile content sharing activities, allowing users to literally play with their content. The second is to uncover the motivations for content sharing and content retrieval within a game-based environment afforded by MARGE. Here, we employ the uses and gratifications paradigm [11], which essentially examines how and why people select specific media to meet their needs or to obtain specific gratifications.

The remaining sections of this paper are structured as follows. Section 2 provides an overview of the related work. Section 3 describes our mobile content sharing game. Section 4 presents the methodology of the study while Section 5 discusses our findings and analyses. Finally, Section 6 discusses the implications of this work as well as opportunities for future research.

2 Related Work

Web-based applications that blend content sharing and gaming elements have emerged recently. Also known as Games With A Purpose [23], content is created and shared through gameplay. The ESP Game [22] is such an example in which two unrelated players are tasked to create matching keywords to randomly presented images within a given time limit. Points are earned based on specificity of the keywords, and coupled with a countdown timer, these elements add excitement and hence motivation for players. While players have fun with the game, the matching keywords (content) can be used as tags for these images, and if sufficient data is collected, these tags can be used improve the performance of image search engines. Other examples include Google Image Labeler (<http://images.google.com/imagelabeler/>) which is a variant of the ESP Game, and a collection that can be found at the Games With A Purpose site (<http://www.gwap.com>).

Similar ideas that blend content sharing and gaming can also be seen in mobile applications. One such example is the Gopher Game [6]. As mentioned, gophers represent missions to be completed, and are carriers of information between players. The game is location-based and players collect gophers as they move about their physical surroundings. A player helps a gopher complete its mission by supplying it with camera phone images and textual content based on a task description. This information is submitted to a community of judges, and players earn points depending on the quality of the content submitted. Using these points, players can create new gophers and participate in other in-game activities. Through the process of helping gophers complete their missions, content sharing among players is facilitated because other users may collect these gophers and view the images and text associated with them. In *MobiMissions* [9], content sharing is accomplished through the completion of missions, which are defined by sequences of digital photographs and text annotations associated with specific locations. Players create missions for others to undertake, search locations for available missions, and create responses to missions created by others. To complete a mission, a player has to capture up to five photographs and add up to five text annotations. This content can then be shared with other players.

Research reported in the literature concerning these games suggests that users do find them entertaining and some useful content can be generated through gameplay (e.g. [9, 6]). Despite these encouraging results, the underlying dynamics that explain why users find gaming and content sharing appealing have not been well explored. Clearly, a theoretically-informed perspective ensures that applications adopting this genre of gaming provide facilities that sustain players' motivations both in creating and retrieving content. Otherwise, such applications are unlikely to succeed. Hence in our research, the uses and gratifications model [11] is employed to study content sharing and retrieval in a game-based environment. This model was originally developed to examine how and why individuals use and adopt mass media in their everyday lives [11]. Subsequent studies reveal that mass media are used for the purposes of both entertainment and utility [7], and that individuals seek gratifications in mass media use based on their needs and motivations [16]. Recently, the scope of such research has been extended to technologies, software and services, although this model has yet to be employed in a context related to ours. For example, [14] studied mobile phone usage through a uses and gratifications perspective with the goal of understanding how people use mobile telephony technology. From an analysis of 417 respondents, major gratification factors included affection and sociability, immediate access, entertainment, reassurance (safety) and fashion and status, and these had a strong influence on how the mobile phone was used. The findings revealed that talking to immediate family members provides affection gratifications while using mobile phones in cars, buses or in malls and restaurants offers immediate access gratifications. Next, [21] employed the model to examine why people play video games and the types of gratifications they would experience. A survey of 550 respondents yielded six uses and gratifications dimensions including competition, challenge, social interaction, diversion, fantasy and arousal. Different patterns of playing behavior were also found, and these were associated with the different uses and gratification dimensions found.

3 Introducing MARGE

MARGE (Mobile Alternate Reality Gaming Engine) is a system that realizes our goal of combining gaming elements in a mobile content sharing application. Play is intertwined with the collaborative creation, seeking and sharing of information such that these activities become the mechanics of gameplay. Unlike the casual games reviewed, we create a persistent layer of alternate reality [12] over digital information associated with the real world by weaving a storyline into mobile content sharing. Put differently, we introduce multiplayer role-playing pervasive gaming elements into mobile content sharing activities.

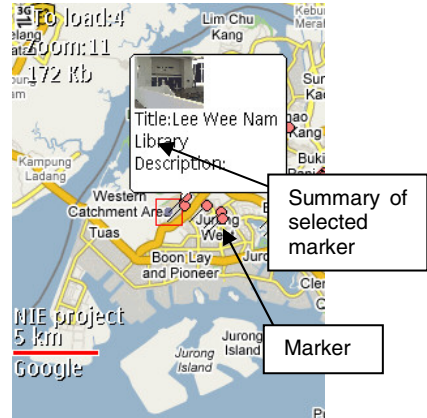
MARGE comprises the facilities for mobile content management, retrieval and discovery, and importantly the game engine that supports our approach to gameplay. Due to space constraints and the objectives of this paper, we will only highlight the major aspects of the user interface and game mechanics of the system.

3.1 Content Sharing Features

Using MARGE, players have an interface for creating, seeking and sharing content on their mobile devices, similar to the facilities offered by existing applications (e.g. [8]).

Here, content refers to location-based annotations, each comprising attributes such as title, tags, textual information, multimedia content (e.g. images) and users' ratings for that annotation (see Figure 1a). Other implicit attributes are also captured such as contributor name, location (latitude and longitude), and date. Annotations are displayed as markers on a map-based interface (see Figure 1b). Here, the map offers standard navigation features such as pan and zoom.

(a) Creating an annotation



(b) Map-based visualization

Fig. 1. Creating and viewing annotations

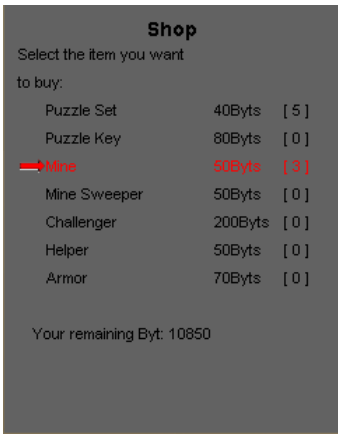
3.2 Gaming Features

At the same time, MARGE gives users the opportunity to concurrently play with their content, by taking on a role and progressing through a virtual world by interacting with other players and accumulating points, similar to a role-playing game. MARGE players start at the lowest level of a role's hierarchy and work their way up by earning points. Points are earned or spent through a player's actions and the actions of others, which may include the following.

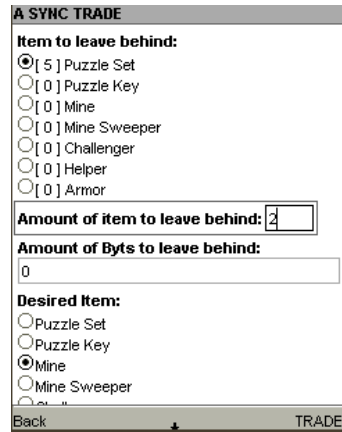
- **Collaborating.** Players earn currency (called *byts*) by contributing location-specific content or give ratings to existing content. These points can then be spent to acquire game tools and access other game-based features.
- **Competing.** Players may acquire traps or puzzles and lay them at various locations to prevent content from being accessed by others, or to inflict damage to other players. For example, a player encountering a puzzle (e.g. a slide puzzle) will have to solve it before being able to access content. A player stumbling upon a trap may cause him/her to lose byts and the trap setter to gain them.
- **Guiding.** This involves creating missions which comprise a list of locations for other players to visit. Creation costs byts to prevent a proliferation of frivolous missions, while embarking on a mission is free. Before embarking on one, a player can see a description of the mission as well as comments given by other players. Upon successful completion of mission (physically visiting all locations), the

player and the creator earns byts. The player is also given the opportunity to rate and comment on the mission.

- **Transacting.** MARGE provides a variety of game items that include traps, puzzles, items that negate traps and puzzles, and so on. These may be purchased from a virtual shop (Figure 2a). Here, players simply specify the item and quantity desired. Items may also be traded with other players by specifying the desired items for trading. Trading may be synchronous, in which communication is done between two parties in real-time, or asynchronous (Figure 2b), in which items are left at a specific location and an interested party may accept the trade if they encounter it.



(a) Buying items from the shop



(b) Trading items with other players

Fig. 2. Buying and trading game items

- **Socializing.** MARGE supports the creation of guilds in which players with similar interests may join, socialize and forge alliances (Figure 3a). Here, players may exchange messages that are accessible only to guild members, trade items that may not be accessible by non-members, and participate in events and gatherings. Players may request to join a guild or be recruited by existing members. Any player with sufficient byts can create a guild, and to further prevent proliferation of guilds in the system, a newly created guild remains in a pending state until at least three members join (Figure 3b).

In MARGE, gameplay is open and undirected, allowing a player to explore information anchored in the physical world, while interacting with other players through the game mechanics layered as a virtual world on top of this information. The game is persistent, meaning that the game state and items remain active independent of whether any particular user is logged in, and players have the flexibility to decide how much involvement with the game he or she desires. Here, we aim to cater to a wide spectrum of users, by allowing MARGE users to vary the levels of usage of the gaming features. For hardcore gamers whose interest can be sustained primarily through gameplay, they

can explore the storylines, roles, in-game quests and items in greater depth. For casual gamers, MARGE offers a rich gaming environment with a multi-faceted reward structure that keeps them engaged when their schedules permit. For non-gamers, gameplay can be dislodged from content sharing activities completely if desired. Further, by exploiting the duality of the “player-as-user” and “user-as-player” dynamics [2], users lacking the propensity to share information may be motivated to do so.

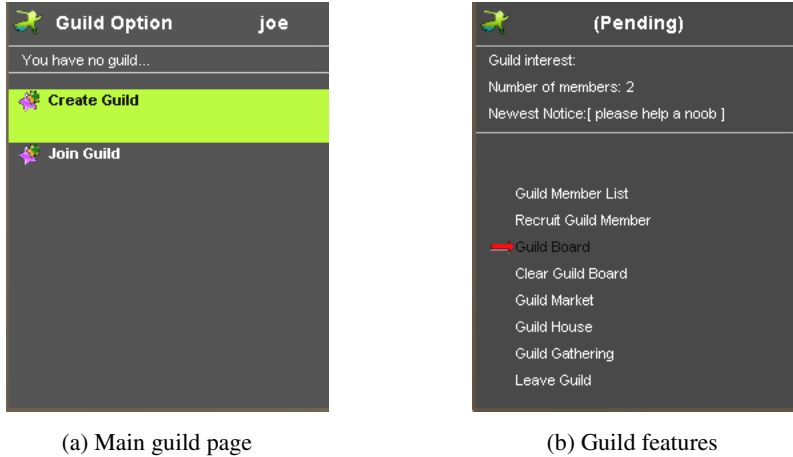


Fig. 3. Socializing through guilds

4 Methodology

To measure perceived gratifications for content sharing and retrieval from using MARGE, we developed a survey instrument based on constructs from past uses and gratification studies [e.g. 7, 14, 15, 18]. Specifically, according to the findings reported in past studies, several factors (i.e. time management, trust, sociability, leisure, relationship maintenance, personal status, reassurance) were identified as potential motives or perceived gratifications behind usage of MARGE, and were incorporated into the survey. For example, perceived gratifications questions to measure relationship maintenance include “I use the application to share information to help others”, “I use the application to share information to thank others”, and “I use the application because I am concerned about others”. A total of 20 question items were used to assess perceived gratifications from the use of MARGE for content sharing and another 20 items were used to assess perceived gratifications of MARGE for content retrieval. The actual questionnaire cannot be included in this paper due to space constraints. Nonetheless, it can be obtained from the authors upon request.

Graduate students from two large universities in Singapore were invited to a presentation on the use of location-based content sharing using mobile devices. The concept of MARGE and its gameplay features were also introduced during the presentation. To help the students understand how MARGE is played, four possible usage scenarios were presented. These scenarios included using MARGE for: (1) creating content and planting puzzles; (2) retrieving and rating content; (3) trading game items

with other players; (4) embarking on missions; (5) joining a guild. Thereafter, the survey, which sought to determine the perceived gratifications in retrieving and sharing content using MARGE, was administered. Participation was voluntary and anonymous. A total of 163 students participated in the written survey of which approximately 51% were male, 45% female and the rest not indicated. The majority of the respondents (70%) were between 18 to 29 years of age, and most (85%) were from the computer science and engineering disciplines, with the remainder from business, arts and the social sciences.

5 Data Analyses and Results

Principal component factor analysis with varimax rotation was run to determine the potential groupings of the perceived gratifications items. Varimax rotation was used to better account for expected correlations among potential factors. Accordingly, several factors emerged with eigenvalues greater than 1.0. A total of 4 items under content sharing and 6 items from content retrieval were dropped during the analysis due to high cross loadings to multiple constructs.

Eight distinct perceived gratifications factors emerged from the factor analysis with 4 perceived gratifications factors supporting content sharing and another 4 perceived gratification factors supporting content retrieval. The 4 factors for sharing are discussed next. The first factor consisted of 8 items which were related to satisfying needs of the individuals which we labeled as SELF. Examples of items in this factor include “need to control”, “need to interact” and “need to experiment with my identity”. The second factor consisted of 4 items which were related to personal status (e.g. to make me look good) which we labeled as “PERSTATUS”. The third perceived factor consisted of 2 items was related to leisure purposes and so was labeled as LEI. Examples of items in this factor include “helps me to combat boredom” and “helps me to pass time”. The last factor consisted of 2 items which centered around issues related to relationship maintenance which we labeled as RELMAINT. Items included “to help others” and “to thank others”.

The 4 perceived gratifications factors for content retrieval are described as follows. The first factor consisted of 5 items which were related to the quality of the information and so we labeled as INFOQUAL. Examples of items in this factor are “I trust the information”, “I know the information will be accurate”, and “I can get the latest news and updates”. The second factor was labeled as SOCBOND and it was related to social bonding issues. This factor consisted of 4 items (e.g. “to keep in touch with people”). The third factor consisted of 4 items and which were related to the process of searching for people, location and information (i.e. “easy to get information I need”, “helps me to find locations”, “I can look for people to interact”). This factor was labeled as SEARCH. The last factor for content retrieval consisted of 2 items which were related to time management and was labeled as TIME (e.g. “immediate access to information anywhere anytime”).

The reliability constructs for the 8 gratification factors were assessed using Cronbach’s Alpha. The results exhibited acceptable alpha values (i.e. ranged from 0.6 to 0.8) for the given sample size.

Statistical analyses were carried out using least squares regression on the 2 models. The first model (Model 1) examined the associations between the extracted 4 perceived gratifications factors from the factor analysis and the intention to use MARGE for content sharing. The second model (Model 2) examined the associations between the 4 extracted perceived gratification factors from the factor analysis and the intention to use MARGE for content retrieval. Table 1 below shows the regression models used.

Table 1. Regression models

Model 1: Usage Intention <i>Sharing</i> = f (SELF, PERSTATUS, LEI, RELMAINT)
Model 2: Usage Intention <i>Retrieval</i> = f (INFOQUAL, SOCBOND, SEARCH, TIME)

Note: SELF: Self, PERSTATUS: Personal Status, LEI: Leisure, RELMAINT: Relationship Maintenance, INFOQUAL: Information Quality, , SOC: Social Bonding Needs, SEARCH: Searching process for locations, information and people , TIME: Time Management

Table 2. Regression Results

Model 1				Model 2			
	Coeffs	t-value	Sig.		Coeffs	t-value	Sig.
SELF	0.22	2.61	0.01**	INFOQUAL	0.38	4.03	0.00**
PERSTATUS	0.17	1.92	0.06+	SOCBOND	0.00	0.03	0.97
LEI	-0.02	-0.27	0.79	SEARCH	0.16	1.32	0.19
RELMAINT	0.22	2.80	0.01**	TIME	0.15	1.62	0.11
Adjusted R ²	0.18				0.29		
F-Stats	9.90**				16.00**		

Note: * $p < 0.05$; ** $p < 0.01$; + $p < 0.10$

Note: SELF: Self, PERSTATUS: Personal Status, LEI: Leisure, RELMAINT: Relationship Maintenance, INFOQUAL: Information Quality, , SOC: Social Bonding Needs, SEARCH: Searching process for locations, information and people , TIME: Time Management

Results from the analyses are shown in Table 2. The regression results of Model 1 show that three factors (i.e. SELF, PERSTATUS and RELMAINT) are significant in predicting the intention to use MARGE for content sharing. Interestingly, our results show that the association between intention to use MARGE for content sharing and for leisure purposes (e.g. combat freedom or to pass time) is not significant. The regression results of Model 2 show that only one factor (i.e. INFOQUAL) ($p < 0.01$) is significant in predicting the intention to use MARGE for content retrieval.

6 Discussion and Conclusion

The objectives of the present paper are two-fold. We first present MARGE, a mobile application that blends content sharing activities with gameplay. Here, gaming features

serve as an extrinsic motivator for content sharing and retrieval, and differs from standard mobile content sharing applications where motivations for use are mostly intrinsic. Next, we conducted an exploratory study that seeks to understand the basic motivational factors influencing people's intention to use MARGE (and in general, its game genre) for content sharing and retrieval. In our work, we employed the uses and gratifications paradigm. Our study highlights two important points. First, our results indicate that the motivations to share content are different from the motivations to retrieve content. Second, through the lens of the uses and gratifications paradigm, our results show that content sharing is a much more complex process than content retrieval. Specifically, the motivations behind content sharing arise from multiple sources – self (satisfying personal needs), relationship maintenance, (i.e. enjoy helping others), and personal status (i.e. to look good). In contrast, the motivation behind content retrieval was centered mainly around the quality of information such as trustworthiness and accuracy.

Our work has yielded the following implications. The different perceived gratifications for mobile content sharing and retrieval would indicate the need for different mechanisms that would motivate such activities. Whereas current mobile content sharing applications have tended to implicitly consider sharing and retrieval as a single construct, we argue that by understanding the differences in gratifications between these two activities, developers would be better informed to implement applications supporting them more effectively. From an application design perspective, it appears that more features are required to better support content sharing activities. In particular, such features would have to address the gratifications factors of self, personal status and relationship maintenance. In this regard, the incorporation of gaming mechanisms in MARGE would be helpful. For example, the ability to play with content such as planting puzzles or creating missions would help to sustain one's interest in content creation. Further, the accumulation of points and ranking up within a role serve to enhance a player's personal status. As well, relationship maintenance can be accomplished through trading of items, joining guilds and creating content that meets other users' needs.

Next, the gratifications associated with content retrieval appear to be much simpler, that of obtaining quality information to meet one's needs. Here, features should be put in place to help users make such decisions. In MARGE, support includes content ratings, commenting on contributions, and viewing of users' profiles. Finally, it was interesting to note that the other perceived gratifications factors that emerged from the factor analysis (i.e. SOCBOND, SEARCH, and TIME) were not statistically significant in predicting the intention to use MARGE for content retrieval. This is consistent with the task-oriented nature of information retrieval as suggested in past studies [19]. The implication is that an application designed for this purpose should provide efficient and effective access to information. In the case of MARGE and similar games, users should ideally have immediate access to content and bypass gaming features if necessary. Hence, we have deliberately designed MARGE such that players have the liberty to determine the level of involvement with the game. This can range from enjoying a full-featured gaming experience to focusing only on content sharing and retrieval without gameplay.

Caution, however, should be exercised when interpreting our results because the nature of this study may reduce the generalizability of its findings. Specifically, the

majority of the respondents was from the IT and engineering backgrounds, and was between 18 and 29 years old. Replication of this study in other contexts (e.g. other age-groups, different backgrounds) or in a specific domain (e.g. tourists, students) would be useful to understand the basic motivational factors influencing people's intention to use MARGE or similar applications. Further, expanding the study to compare MARGE with other mobile content sharing applications would allow researchers to investigate the impact of gameplay and its influence on why people share and retrieve content via mobile communication devices.

References

- [1] Ames, M., Naaman, M.: Why we tag: Motivations for annotation in mobile and online media. In: Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems, pp. 971–980 (2007)
- [2] Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S., Capra, M., Hampshire, A.: Interweaving mobile games with everyday life. In: Proceedings of the 2006 Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 417–426 (2006)
- [3] Bellotti, V., Begole, B., Chi, E.H., Ducheneaut, N., Fang, J., Isaacs, E., King, T., Newman, M.W., Partridge, K., Price, B., Rasmussen, P., Roberts, M., Schiano, D.J., Walendowski, A.: Activity-based serendipitous recommendations with the Magitti mobile leisure guide. In: Proceedings of the 2008 Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 1157–1166 (2008)
- [4] Bolton, G.E., Katok, E.: How effective are electronic reputation mechanisms? An experimental investigation. *Management Science* 50(11), 1587–1602 (2004)
- [5] Brown, H.G., Poole, M.S., Rodgers, T.L.: Interpersonal traits, complementarity, and trust in virtual collaboration. *Journal of Management Information Systems* 20(4), 115–137 (2004)
- [6] Casey, S., Kirman, B., Rowland, D.: The gopher game: A social, mobile, locative game with user generated content and peer review. In: Proceedings of the 2007 International Conference on Advances in Computer Entertainment Technology, pp. 9–16 (2007)
- [7] Flanagin, A.J., Metzger, M.J.: Internet use in the contemporary media environment. *Human Communication Research* 27(1), 153–181 (2001)
- [8] Goh, D.H., Sepoetro, L.L., Qi, M., Ramakrishnan, R., Theng, Y.L., Puspitasari, F., Lim, E.P.: Mobile tagging and accessibility information sharing using a geospatial digital library. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 287–296. Springer, Heidelberg (2007)
- [9] Grant, L., Daanen, H., Benford, S., Hampshire, A., Drozd, A., Greenhalgh, C.: MobiMissions: The game of missions for mobile phones. In: Proceedings of the International Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2007 Educators Program (2007), <http://doi.acm.org/10.1145/1282040.1282053>
- [10] Hong, T.: The influence of structural and message features on web site credibility. *Journal of the American Society for Information Science and Technology* 57(1), 114–127 (2006)
- [11] Katz, E., Blumler, J.G.: *The Uses of Mass Communications: Current Perspectives on Gratifications Research*. Sage, Beverly Hills (1974)
- [12] Kim, J.Y., Allen, J.P., Lee, E.: Alternate reality gaming. *Communications of the ACM* 51(2), 36–42 (2008)

- [13] Lee, C.S., Goh, D.H., Razikin, K., Chua, A.: Tagging, sharing and the influence of personal experience. *Journal of Digital Information* 10(1) (2009), <http://journals.tdl.org/jodi/article/view/275/275>
- [14] Leung, L., Wei, R.: More than just talk on the move: A use-and-gratification study of the cellular phone. *Journalism & Mass Communication Quarterly* 77(2), 308–320 (2000)
- [15] Leung, L., Wei, R.: The gratifications of pager use: Sociability, information seeking, entertainment, utility, and fashion and status. *Telematics and Informatics* (15), 253–264 (1998)
- [16] Lin, C.A.: Looking back: The contribution of Blumler and Katz's uses of mass communication to communication research. *Journal of Broadcasting & Electronic Media* 40(4), 574–582 (1996)
- [17] Lui, S., Lang, K., Kwok, S.: Participation incentive mechanisms in peer-to-peer subscription systems. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, p. 302b (2002)
- [18] Payne, G.J.H.S., Dozier, D.: Uses and gratifications motives as indicators of magazine readership. *Journalism and Mass Communication Quarterly* (Winter), 909–913 (1988)
- [19] Reid, J.: A Task-oriented non-interactive evaluation methodology for information retrieval systems. *Information Retrieval* 2(1), 115–129 (2000)
- [20] Ridings, C.M., Gefen, D., Arinze, B.: Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems* 11, 271–295
- [21] Sherry, J.L., Lucas, K., Greenberg, B.S., Lachlan, K.: Video game uses and gratifications as predictors of use and game preference. In: Vorderer, P., Bryant, J. (eds.) *Playing Video Games: Motives, Responses and Consequences*, pp. 213–224. Lawrence Erlbaum Associates, Mahwah (2006)
- [22] von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: *Proceedings of the 2004 Annual SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326 (2004)
- [23] von Ahn, L., Dabbish, L.: Designing games with a purpose. *Communication of the ACM* 51(8), 58–67 (2008)

Why We Share: A Study of Motivations for Mobile Media Sharing*

Dion Hoe-Lian Goh¹, Rebecca P. Ang², Alton Y.K. Chua¹, and Chei Sian Lee¹

¹ Wee Kim Wee School of Communication and Information
² School of Humanities and Social Sciences
Nanyang Technological University
{ashlgoh, rpang, altonchua, leecs}@ntu.edu.sg

Abstract. Mobile phones equipped with cameras have become popular among consumers, and this has fuelled an increase in mobile media sharing. The present research investigates the sharing of mobile media by conducting a diary study to specifically understand the type of media captured and shared, and the motivations behind these activities. Participants maintained a month-long diary, documenting their media sharing activities. Post-study interviews were also conducted to elicit additional information not captured in the diary. Results suggest a range of motivational factors, and that social and emotional influences played an important role in media sharing behavior. Participants were also more inclined to share photos than any other media due to cost and transmission time considerations. Implications of our work are also discussed.

Keywords: Mobile media sharing, diary study, user-generated content, motivations, mobile phone.

1 Introduction

Mobile phones equipped with cameras have become popular among consumers, and this in turn, has fuelled an increase in interest in digital photography and mobile media sharing due to the ease of use of these devices. In addition to the camera, many of these mobile phones also offer wireless networking functionality through WIFI, 3G, Bluetooth, MMS, and so on. When combined, these mobile phones become an important and easy to use platform for media capturing and sharing. For example, it is now common for someone to take a photo, video or audio recording and send it directly to a friend via MMS or Bluetooth, or even upload it to an online service such as Flickr or Facebook where it can be shared to a community of users.

Beyond hardware availability and infrastructure support, people capture and share media for a variety of purposes. At an individual level, this activity can be personally satisfying and even meet emotional needs [14]. People may also share media to maintain and strengthen social relationships, or to express oneself to the public [19].

* This work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research Grant NRF NRF2008IDM-IDM004-012. The authors also wish to thank Keng Hong Low, Wai Man Yau and Riduan Hassim for their assistance.

Conversely, people may choose to capture media for personal use only, perhaps for personal recollection or self-reflection [13]. Here no sharing takes place because of distrust, the fear of harming oneself, or the perceived lack of interest by others. Instead, the captured media may be stored locally in the phone, or uploaded to an online service but made restricted to personal access only.

Clearly, there are myriad reasons for capturing and sharing photos and other media, spawning a wealth of research examining this phenomenon from multiple perspectives. For example, such work includes investigating personal and group sharing (e.g. [5, 13]), designing systems for supporting these activities (e.g. [16]), as well as studying annotations created for captured media (e.g. [1]). By understanding how and why people capture and share media, designers will be better able to implement systems supporting such activities that meet users' needs more effectively.

The present research extends and complements research in this area by conducting a diary study to understand the type of media captured and shared, and the motivations behind these activities. Previous methods have involved direct observations which may be intrusive and may not fully capture user intentions when observers are not present [5]. Interviews have also been employed but responses are elicited after the activities are completed, potentially leading to information being filtered out by interviewees [13]. Transaction log analyses have been used as well but the actual intentions of the activities of users may be difficult to ascertain [16]. Diary studies on the other hand, are useful in capturing everyday experiences and require participants to complete a diary that details their activities [20]. They are non-intrusive as no observers are present, and diary entries are recorded *in situ*, as activities occur, and therefore are not retrospective like interviews. In addition to methodology, we aim to add to the body of knowledge in media capturing sharing by identifying additional motivational factors not covered in prior research such as the influence of emotions.

The remaining sections of this paper are structured as follows. Section 2 provides an overview of the related work. Section 3 describes the methodology of the diary study while Section 4 presents our findings and analyses. Finally, Section 5 discusses the implications of this work as well as opportunities for future research.

2 Related Work

Here, we review research investigating mobile media capturing and sharing behavior. In [6], the use of images taken by mobile phones from 34 volunteers as a means of communication among people was studied. From the 295 images analyzed, reasons for image capture and sharing were grouped into two dimensions, affective versus functional, and social versus individual. Affective images can be used to enrich mutual experiences between co-present people or to communicate with absent family or friends. Functional images on the other hand, are used to support both mutual tasks among co-present people, and remote tasks. Interestingly, almost half of all images taken were not shared, that is, used for individual purposes. Such images were used for personal reflection or to support a future personal task.

In similar work, [19] studied 60 graduate students and faculty to understand how they used digital images captured using the MMM2 system. Improving on the taxonomy of reasons by [6] and the researchers' own earlier work, five major uses of images

were determined: (1) creating and maintaining social relationships; (2) as a record and reminder of personal and collective experiences; (3) as a means of self-expression to voice one's views; (4) to influence others' view of oneself through self-presentation; and, (5) as a means to support both personal and group tasks. Later work has mostly concurred with these earlier reasons for mobile media capturing and sharing, but with the provision of more nuanced categories. For example, [9] organized reasons for co-present media sharing into storytelling, identity presentation, social information sharing and serendipitous discovery.

Work has also been done on motivations for annotating captured media. For example, [1] examined the motivations for annotation and tagging Flickr images using the ZoneTag cameraphone application. Through interviews with a select group of ZoneTag users, motivations for annotation were cast along two dimensions of sociality and function. The sociality dimension relates to whether an annotation was meant for personal use or for others, while the function dimension refers to an annotation's intended use – organization or communication. Similarly, [10] employed a statistical approach to determine tagging motivations based on the social dimension in the taxonomy developed by [1]. Using survey and log data obtained from Flickr users, statistically significant sociality motivators included individual use and public sharing as these were positively correlated with a high number of tags.

Related to motivations, other research has focused on the type of content that people capture and share. For example, studies have shown that the images taken on mobile phones tend to be personal, short-lived and ephemeral, in other words, used more for personal record-keeping and archiving [4, 12]. Likewise, [7] concluded that mobile phones tended to participate in the “aesthetics of banality” in which the images captured were mainly focused on the mundane, trivial aspects of everyday life which are shared with friends and acquaintances [15]. Here, [17] further suggested three categories of media that could be shared: performative (used to generate an act, e.g. an image of a place to which the recipient should go as soon as possible), informative (used to convey information, e.g. letting friends know you have new significant other through a posed picture), and problem-solving (taken to address an issue, e.g. taking picture evidence of a road accident for an insurance company's damage appraisal).

3 Methodology

As discussed, the goal of our study was to gain an understanding of the types of media users typically captured and shared, and the motivations for doing so. To accomplish this, we adopted the diary study methodology to gather data from our participants. This was supplemented with an interview at the end of the diary study which gave the researchers an opportunity to shed light on unclear diary entries as well as elicit additional information not captured in the diary.

Participants in the study were 18 graduate students who were also working full-time jobs. They were between 25 to 40 years and were taking a course on mobile applications development. Of the 18, 16 were male and 2 were female. All of them owned mobile phones that were capable of capturing media such as photos, video and audio. Participants had varied backgrounds, coming from industries such as banking, security, education, telecommunications, information technology, and logistics. All participants had some prior experience in media sharing activities, either through the

phone's networking capabilities (e.g. 3G, WIFI, MMS) or through direct transfer such as Bluetooth and infrared.

Participants maintained a diary over four weeks. The diary was an electronic template in the form of a spreadsheet. Participants were asked to record an entry in a diary whenever they shared media using their mobile phones. The information recorded included: the time and location of media capture, the intended recipient, the reason for capturing and sharing, and the emotions experienced at the time of capture.

Since participants may not be carrying the template during a diary incident, we asked participants to use their mobile phones to record "snippets" of their diary entries, similar to [18] except that the snippets were saved on the mobile phone rather than posted online. At the end of each day, participants were requested to complete their diaries based on the snippets they created. Although some information may be lost using this two-step approach, the advantage of this technique is that it reduced the burden of data collection since logging complete diary entries at the point of the activity may be difficult for some participants and/or situations.

At the end of the four week period, participants uploaded their captured media and diary entries to a designated Web site where they were then analyzed by the researchers. Participants were also requested to attend a post-study interview in which they further elaborated on their motivations and clarify unclear diary entries. On average, each interview lasted 20 minutes.

4 Results and Analyses

The diary study generated a total of 375 entries, with an average of 20.8 entries per person (minimum=7, maximum=30). In the following, we elaborate on the findings from the analysis of these diary entries.

4.1 What and for Whom: Media Content and Recipients

Of the 375 entries each of which represented one sharing activity, an overwhelming majority were images (n=370; 98.76%). The remaining were video (n=4; 1.07%) and audio (n=1; 0.27%). These skewed numbers were obtained despite the fact that participants were encouraged to capture and share any type of media, and that they all owned feature-rich mobile phones that could capture images, audio and video.

In the interviews, several reasons were uncovered. First, many of the mobile phones had limited storage capacities and participants did not want to capture media types with large file sizes for fear of rapidly filling up the phone's storage space as there were other uses for that space such as music files, contacts, ring tones, applications and so on. Large file sizes also meant longer transmission times needed for sharing, and if GPRS/3G networks were used, this also meant higher telecommunications charges incurred. Finally, many of the participants felt that images already contained a lot of information that the sender wanted to convey to the recipient, and therefore, alternative media types such as video were not necessary.

Figure 1 shows the categories of media captured for sharing. These categories were deduced from a manual inspection by the researchers. Here, the sharing of media associated with people represented the most frequently occurring activity, accounting for approximately 15% (n=57) of all diary entries. This was followed closely by 14% for

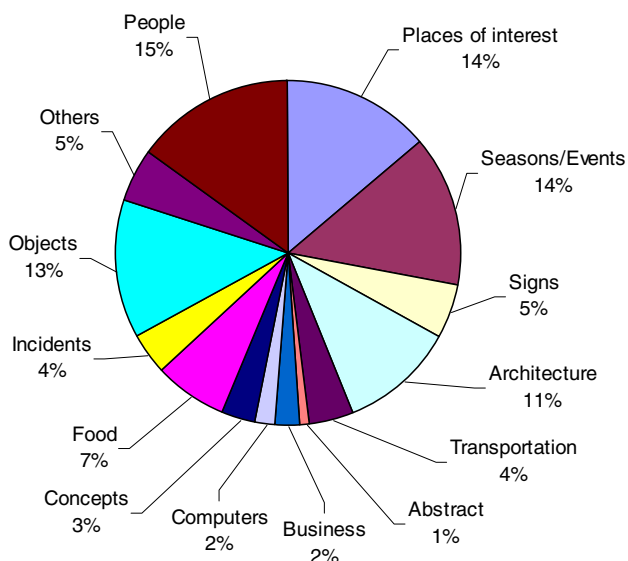


Fig. 1. Media categories captured and shared

the places, objects and seasons/events categories. The remainder of the categories was low in frequency, ranging between 2%-11%.

Next, the diary entries also revealed a diversity of recipients. Here, friends represented the majority of recipients at 30% of all media sharing activities. This was followed by family (24%) and colleagues (18%). The remainder (28%) consisted of acquaintances and others, such as friends of friends. The latter finding is interesting as it suggests that media sharing even occurs among people the sender is not close to.

4.2 Why: Motivations for Media Sharing

As discussed, there has been active research conducted in identifying motivations for media sharing. In our analysis, we sought to identify similar patterns of motivational factors from related work as well as reasons that have received little attention in the literature thus far. For the former, motivational factors were culled from the work of [1, 6, 19]. These were grouped into the following categories: (1) creation/maintenance of social relationships; (2) reminding of individual and collective experiences; (3) self-presentation; (4) self-expression; and (5) task performance. We manually inspected the motivations provided by the participants and attempted to include them into these predefined categories. For those not classifiable, we then constructed new categories of motivational factors by identifying similarities across diary entries and coding them into logical groupings. The following subsections discuss the motivations that emerged from our analysis.

4.2.1 Creating and Maintaining Social Relationships

This refers to forging and sustaining relationships between the sender and recipients through the shared media. Such media may contain representations of the people involved in the sharing activity or content that is of interest to one or more of the

parties. This motivation was the most frequent and occurred in more than half (52.80%) of all diary entries. Within this category, four subthemes emerged.

- *Sharing common ground.* Both sender and recipient shared similar interests, and sharing was conducted because both parties would appreciate the media object, thus reinforcing the common bond. One participant took a photo of himself having a good time singing at a karaoke session. When interviewed, the participant mentioned that he was motivated to share this moment with another person who also enjoyed this activity.
- *Interest of recipient.* Sharing occurs because the sender knows that the recipient would be interested in the media object. The media object could be captured deliberately or serendipitously, and done either at the request of the recipient or at the sender's own initiative. One participant happened to chance upon a book in a bookstore she knew a friend was looking for. She took a photo of the cover and sent it to her friend to inform her that it was available for sale.
- *Telling stories.* Just as people communicate verbally with narratives of various types to maintain a relationship, sharing media also fulfills this purpose. When sharing, participants would include textual messages to supplement the story being told by the media object. Figure 2a shows a photo of an automobile accident witnessed and taken by a participant. He sent the image via MMS to his friend with a short textual description. His reason was simply to tell his friend of the incident just as one would do so over a meal.
- *Connecting with loved ones.* Sharing typically occurs among close family members or people involved in relationships. Again, the mobile phone functions as a communicative device [19] employing media instead of voice. The media shared may either take the place of actual conversation or to supplement it. One participant took a picture of himself and sent it to his family currently living in another country. He mentioned that they had been apart for some time, and sharing this picture was important for both him and his family.

4.2.2 Reminder of Individual and Collective Experiences

This motivational factor involves sharing media as a record and reminder of individual and collective experiences and may include key moments, everyday activities or even mundane content related to oneself, others within a social circle or even the public. It occurred in 41.6% of all motivational factors indicated by our participants. The key differences between this motivation factor and the previous are: (1) the media captured may be for personal consumption (no sharing) or for public consumption, rather than for a select social group; (2) record keeping, archiving and narrating is the main purpose and whereas the strengthening of relationship bonds is secondary at best. Within this motivational factor, three sharing activities emerged:

- *Personal memory.* Several diary entries contained media that participants said were for personal use only. These included pictures of loved ones, pets, hobbies, sports, nature and so on.
- *Sharing key moments.* The content of the media is meant to interest groups of people or the public. The goal was to report rather than socialize. Typically, media would be uploaded to blogs, photo sharing sites (e.g. Flickr) or personal Web sites. Examples were varied and included nature scenes, sports, and events.

- *Sharing daily activities.* This is a more mundane version of reporting captured moments as content relates to everyday activities. Again, the destination of these media would be on publicly accessible online services (e.g. blogs). Examples observed in our diary submissions included food, an office gathering and a nice window view from a participant's home.

4.2.3 Self-expression

This was the least frequently occurring category and accounted for 21.1% of all submissions. Here, we adopt the definition by [19] in which self-expression refers to the sharing of media to express one's view of the world. However, we elaborate on this definition by identifying more nuanced themes. Our analysis suggests that participants expressed three major types of views through their shared media:

- *Environmental.* Media were related to environmental issues. One participant was appalled by a stack of garbage bags piled under a tree in a park. He took a photo (Figure 2b) and sent it to the authorities and his friends to express his concern.
- *Social.* Media captured expressed views and incidents about people and society. A participant took a photo of a food court with dirty trays stacked on several tables. The sender wanted to complain about the lack of courtesy of patrons at food courts, and used this photo as an example to express his point.
- *News reporting.* This theme reflects the rise in popularity of citizen or participatory journalism in which news items are contributed from the perspective of the man-on-the-street, often through the media capturing features of the mobile phone. Examples include a video snippet of a Formula One night race, and an accident along a busy highway

4.2.4 Task Performance

Media shared in this motivational factor have a functional value and are meant to assist the sender and/or recipient to complete a task [6, 19]. This factor occurred in 22.93% of all submitted diary entries. A more fine-grained set of tasks that emerged from our analysis includes the following:

- *Reminding.* Shared media were used to remind oneself or others about a task. For example, a participant sent a photo of the entrance to a shopping to her group of friends to indicate the location and time of their meeting.
- *Informing and decision making.* Media is also shared among people to provide information for decision making. An example was a participant who took a photo of a map of the university and sent it to a friend who needed driving directions. Likewise, an earlier example of a participant who took a photo of a book cover would fall in this category as well.
- *Replacement for writing.* With the increased resolution of cameras on mobile phones, photos can be used as a replacement for writing and note taking. For example, a participant captured photos of hand drawings of user interface designs on large pieces of paper rather than photocopying them. During the interviews, a few participants also mentioned (but did not submit) that they would use their mobile phones to capture video snippets of lectures or presentations to share with their fellow students or colleagues who were not present.



(a) An example of story telling



(b) Expressing an environmental view



(c) A photo associated with happiness



(d) A photo associated with loneliness

Fig. 2. Sample media shared by participants

4.3 The Influence of Emotions and Others

Apart from the motivational factors highlighted above, we also sought to determine the influence of emotions and other people in the sharing of media. To the best of our knowledge, these have not been extensively studied but are relevant as humans are emotional and social in nature, and these attributes are likely to affect sharing.

4.3.1 Emotional Influences

Participants were requested to record their emotions associated with each sharing episode in the diary, if any. At the final tally, 44.27% of all entries contained expressions of emotions. From our analysis, a total of nine major types were identified. This ranged from happiness (23.91% of all entries), excitement (20.65%), surprise (12.23%) at the top frequently occurring end, to fear (3.26%), loneliness (3.80%) and anger (4.08%) at the bottom end. Note that each diary entry may have multiple emotions and hence the percentages do not sum to 100%.

An example of a photo associated with the positive emotions of happiness and excitement is shown in Figure 2c. During the interview, the participant shared that this was a picture of his sleeping child in his home country. When he received this image, he felt happy and excited, and was compelled to share it with his close friends. Figure 2d is a photo of a tent in the middle of a large field. The participant was alone when he captured this photo and associated it with loneliness when he shared it.

Interestingly, our data suggests that positive emotions have a greater influence on media sharing behavior rather than negative ones. Specifically, the top three emotions of happiness, excitement and to a certain extent, surprise, accounted for more than

55% of all emotions reported by our participants. This is supported by research which suggests that positive emotions create a sense of openness and cooperation, making a person more inclined to share. On the other hand, negative emotions may cause a person to be more reticent and hence less sharing occurs [3].

4.3.2 Social Influences

Media sharing takes place within a social setting with its own norms, values, language, culture and incentives. In addition, the people around the sender at the time of media capture, and the intended recipients may also exact influences on the sender. We therefore sought to determine whether any of these factors played a role in shaping sharing behavior by asking questions specific to these areas in the diary.

Our analysis of the diary entries showed that 42.12% of submissions were influenced by other people and social norms. Four major categories may be identified:

- *People the sender knew.* This accounted for 14.13% of all diary entries and refers to people at the time of sharing or the recipients of the media. Put differently a sender's companions, like his/her recipient, are likely to influence sharing behavior. For example, a participant took a photo of a manatee during a visit to an aquarium because his sister wanted a picture of this sea creature.
- *Trends/culture.* The influence of current affairs and issues in society as well as cultural values were found in 8.15% of all diary entries. In other words, we found that senders were inclined to share media on the latest trends and issues as well as what is culturally acceptable. For example, the period of data collection coincided with a Formula One night race (a first for the competition) and several participants took the opportunity to share media related to it. In our interviews, many participants also expressed reservations about sharing "sensitive" images that may offend others or those that may land them in trouble with the law.
- *People the sender did not know.* A total of 6.79% of diary entries reported this influence. The media object typically captured the actions or behaviors of people which somehow piqued the sender's interest. This interest could arise from a positive encounter, such as a photo of children playing at a beach, but our analysis shows that this was more likely to be something negative. One participant took an image of badly parked car that took up two parking lots and sent it to a local citizen journalism Web site, while another shared a photo of the poor state of his rented apartment left by the previous tenant.
- *Incentives or rewards.* This influence accounted for 6.25% of all diary entries. Here, media objects were shared to gain praise, attention or other incentives from the recipients, similar to that found by [1]. The recipient of the shared media could be friends/family, or even the public, in which case they are uploaded to online to blogs, photo sharing sites and so on. One participant mentioned that he regularly sent photos to a local citizen journalism Web site as he often looked forward to the comments received by others.

5 Discussion and Conclusion

Our analysis of the diary entries yielded seven major motivational factors for media sharing. In decreasing order of frequency, these are creating and maintaining social

relationships (52.80%), emotional influences (44.27%), social influences (42.12%), reminder of individual and collective experiences (41.60%), self-presentation (37.30%), task performance (22.93%), and self-expression (21.10%). Note that entries could contain multiple factors and hence percentages do not sum to 100%.

Findings show that emotional and social influences play a major role in media sharing behavior. To the best of our knowledge, these areas have received little attention but are crucial because research has demonstrated that in general, emotions [2] and the social context [16] play significant roles in shaping behavior. Further, we drilled down each influence to uncover more specific factors and found that positive emotions are associated with higher incidences of media sharing than negative ones. In addition, we identified four components within the social context that have an effect on media sharing behavior. We also extend the existing work of [1, 6, 19] and others by providing a more nuanced analysis of each motivational factor for media sharing. Specifically, we uncovered distinct patterns of behavior within each factor, and these yielded multiple sub-constructs per factor. Through a better understanding of why people share media, improved support for such activities can be designed and implemented.

To summarize, our key findings and their implications include:

- *Photos as the primary means of sharing.* Participants seemed unwilling to use other media despite their mobile phones supporting video and audio capture. This was mainly due to the higher costs of data transmission and longer transmission times. Clearly for a full range of mobile media sharing to occur, service providers would have to address these issues. In our interviews, participants mentioned that transmission charges should ideally be in the range of 10 to 20 cents per media object, closer to the cost of sending an SMS message locally.
- *Refining content organization techniques.* By understanding the types of content that people share and their motivational factors, automated techniques could be refined to help senders and recipients annotate and organize their media objects. For example, knowing that people and places are the top two types of content shared, content analysis techniques could be refined to improve their image recognition accuracy rates to aid access of related media. Likewise, annotations could be suggested based on the frequently encountered motivational factors.
- *Leveraging positive emotions.* Although software developers cannot control a user's emotions at the point of media capture and sharing, they would do well to ensure a more than satisfying usage experience for their software. As suggested by our findings, frustrated or unhappy users are less inclined to share and therefore use applications designed for this purpose. Further, helping users create and access media based on emotive terms may be an area worth investigating.
- *Leveraging the social context.* Mobile media sharing applications should use the social context to help a user capture, annotate and access shared media. For example, suggestions of annotations of captured media could take into account the intended recipients or the sender's social circle to establish a sense of common ground. Similarly, accessing shared media could employ social network analysis to determine relevant content a user might be interested in.

In future work, we plan to extend the study to include participants from a larger age range. Although our sample of 25-40 year old participants represent a key demographic

in terms of high mobile phone usage, other age groups are increasingly using these devices and they may exhibit different media sharing behaviors than what we have observed. Gender differences could be studied as well because research has suggested males and females are motivated differently [11]. Further, the effect of other personal characteristics such as personality, experience with sharing, and other relevant attitudes could also be investigated [8]. Finally, our research has shown the importance of emotions in media sharing. These findings are preliminary and future work could delve deeper by examining the strength of the association it has on media sharing as well as its association with type of content shared.

References

- [1] Ames, M., Naaman, M.: Why we tag: Motivations for annotation in mobile and online media. In: *Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems*, pp. 971–980 (2007)
- [2] Baumeister, R.F., Vohs, K.D., DeWall, C.N., Zhang, L.: How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review* 11(2), 167–203 (2007)
- [3] Bryant, B.K.: Context of success, affective arousal, and generosity: The neglected role of negative affect in success experience. *American Educational Research Journal* 20(4), 553–562 (1983)
- [4] Gye, L.: Picture this: The impact of mobile camera phones on personal photographic practices. *Journal of Media and Cultural Studies* 21(2), 279–288 (2007)
- [5] Jacucci, G., Oulasvirta, A., Salovaara, A., Sarvas, R.: Supporting the shared experience of spectators through mobile group media. In: *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, pp. 207–216 (2005)
- [6] Kindberg, T., Spasojevic, M., Fleck, R., Sellen, A.: I saw this and thought of you: Some social uses of camera phones. In: *CHI 2005 Extended Abstracts on Human Factors in Computing Systems*, pp. 1545–1548 (2005)
- [7] Koskinen, I.: Seeing with mobile images: Towards perpetual visual contact. In: Nyíri, K. (ed.) *A Sense of Place: The Global and the Local in Mobile Communication*. Passagen Verlag, Vienna (2005)
- [8] Lee, C.S., Goh, D.H., Razikin, K., Chua, A.: Tagging, sharing and the influence of personal experience. *Journal of Digital Information* 10(1) (2009), <http://journals.tdl.org/jodi/article/view/275/275>
- [9] Naaman, M., Nair, R., Kaplun, V.: Photos on the go: A mobile application case study. In: *Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems*, pp. 1739–1748 (2008)
- [10] Nov, O., Naaman, M., Ye, C.: What drives content tagging: The case of photos on Flickr. In: *Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems*, pp. 1097–1100 (2008)
- [11] Nysveen, H., Pedersen, P.E., Thorbjørnsen, H.: Explaining intention to use mobile chat services: Moderating effects of gender. *Journal of Consumer Marketing* 22(5), 247–256 (2005)
- [12] Okabe, D., Ito, M.: Everyday contexts of camera phone use: Steps toward technosocial ethnographic frameworks. In: Hoflich, J., Hartmann, M. (eds.) *Mobile Communication in Everyday Life: An Ethnographic View*, pp. 79–102. Frank & Timme, Berlin (2006)

- [13] Olsson, T., Soronen, H., Väänänen-Vainio-Mattila, K.: User needs and design guidelines for mobile services for sharing digital life memories. In: Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices, pp. 273–282 (2008)
- [14] Pering, T., Nguyen, D.H., Light, J., Want, R.: Face-to-face media sharing using wireless mobile devices. In: Proceedings of the Seventh IEEE International Symposium on Multimedia, pp. 269–276 (2005)
- [15] Rodden, K., Wood, K.R.: How do people manage their digital photographs? In: Proceedings of the 2003 SIGCHI Conference on Human Factors in Computing Systems, pp. 409–416 (2003)
- [16] Salovaara, A., Jacucci, G., Oulasvirta, A., Saari, T., Kanerva, P., Kurvinen, E., Tiitta, S.: Collective creation and sense-making of mobile media. In: Proceedings of the 2006 SIGCHI Conference on Human Factors in Computing Systems, pp. 1211–1220 (2006)
- [17] Scifo, B.: Domestication of camera phone and MMS communications: The Italian youth experiences. In: Nyíri, K. (ed.) *A Sense of Place: The Global and the Local in Mobile Communication*, pp. 363–373. Passagen Verlag, Vienna (2005)
- [18] Sohn, T., Li, K.A., Griswold, W.G., Hollan, J.D.: A diary study of mobile information needs. In: Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems, pp. 433–442 (2008)
- [19] van House, N., Davis, M., Ames, M., Finn, M., Viswanathan, V.: The uses of personal networked digital imaging: An empirical study of cameraphone photos and sharing. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1853–1856 (2005)
- [20] Wheeler, L., Reis, H.T.: Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality* 59(3), 339–354 (1991)

The Layout of Web Pages: A Study on the Relation between Information Forms and Locations Using Eye-Tracking

Mi Li^{1,2}, Yangyang Song¹, Shengfu Lu¹, and Ning Zhong^{1,3}

¹ International WIC Institute, Beijing University of Technology
Beijing 100022, P.R. China
lusf@bjut.edu.cn

² Liaoning ShiHua University, Liaoning, 113001, P.R. China
limi.666@emails.bjut.edu.cn

³ Dept. of Life Science and Informatics, Maebashi Institute of Technology
Maebashi-City 371-0816, Japan
zhong@maebashi-it.ac.jp

Abstract. Web pages are an important human-computer interface, and the layout of Web pages has an effect on visual search. This study focuses on using eye-tracking to explore the relation between information forms and locations, as well as the impact of floating advertisements (ads for short) for visual search on Web pages. The results show that for the information presented as text, the fixation duration in quadrant I (right upper), II (left upper) and III (left lower) are shorter than in quadrant IV(right lower) and the center area; for the picture, the fixation duration in quadrant II is the shortest and in quadrant III is the longest. No matter what location the information is placed on a Web page, there is no significant difference between the Web page with floating ads and without floating ads on the fixation duration, which indicates that floating ads have no significant effect on the layout of Web pages. The results suggest that the important information whether text or a picture is suitable to place in quadrant II; however, it is the worst for the layout of Web pages to arrange text information in quadrant IV and the center area or picture information in quadrant III.

1 Introduction

The Internet is playing an important role in people's daily life, and becoming one of the main channels from which people can get their desired information. Web pages are the carrier of Internet information and the important human-computer interface. The users' visual search efficiency is directly related to the quality of Web page design, thus it is hard to find the desired information quickly on an irrational Web page design. Users' visual search on Web pages is impacted by many factors, such as colors [1,2], font size [3], information forms [4,5] and the scan path [6]. And the information layout of Web pages is an important aspect in Web page design.

The users' visual behavior on Web pages includes two types which are visual browsing without a target and visual search with a target. For the study of browsing on Web pages, the users' scan paths usually present a rough and general pattern as an "F-shape" [7]. For the visual search with a target, the users' scan paths usually present the peripheral characteristic of visual search [8,9]. Animation prolonged the visual search time [10], and floating ads didn't impact the mode of visual search [9], however, it would just make users feel bored [11]. The studies about the effect of information locations on visual search demonstrated that, the search time was the longest when the information was located on the right side or lower side of the Web page [12,13], the upper information was found first, and the lower information was always neglected [14]. The previous studies about visual search focused on the information forms or locations, however, the study on the relation between them for the effect on users' visual search has not been reported yet.

Web pages are an important human-computer interface between people and the Internet. The studies about the human-computer interface have been gradually transformed from the machine-oriented to people-oriented. Thus, the philosophy for the Web page design has gradually been changed from that users have to adapt to Web pages to that the Web pages are designed according to users' habit. That is because the Web page design based on the habit of users' browsing can attract much more users' attention.

Eye-tracker can capture and record the detailed data of a user's eye movement, such as fixation duration and fixation count, which can reflect the human mental activities effectively. So far, eye-tracking has become an effective method to investigate the users' visual search behavior on Web pages [8 - 14].

This study focuses on exploring the relation between information forms and locations, as well as the floating ads for the impact of a Web page layout on visual search using eye-tracking. The results may be helpful for the Web page designers to find out users' interested part of Web pages and arrange the layout of different information forms on a Web page.

2 Experiments

2.1 Experiment 1: Relation between Information Forms and Locations

Participants. The participants were 50 undergraduates and postgraduates with the age range of 21 ~ 25 years old ($M = 23.0$, $SD = 1.3$), and half of them were female. All participants were right-handed, had normal or corrected-to-normal vision, often surf on the Internet and were skilled users of the mouse. None of them had the experience on eye movement experiment.

Apparatus. In the study, eye movements were recorded using the Tobii T120 eye-tracker made in Sweden at the rate of 120 HZ. The software of Tobii presented the Web pages automatically and recorded the data during participants' visual search on Web pages. The Web pages were displayed on the screen which

is a 19 LCD monitor with resolution set to 1024×768 pixels and at the refresh rate of 60 HZ. The distance between participant and screen was about 60 cm.

Experimental Material and Design. Studies about the layout of a Web page usually divide the Web page into several parts and the divisions were different. In some researches, the Web page was divided into upper and lower or left and right roughly; some divisions were in according with the four quadrants; and some were divided into 9 locations, etc. Actually, the layout of a Web page usually likes a tortoise shell that the center area is added on the base of 4 quadrants and each location is about 20 percent of the Web page, as shown in Fig. 1. In this study, this method is used for the layout of Web pages. The forms of the information are text and picture which are commonly used on the Web page. Text is the Chinese phrase consisted of 3 ~ 6 Chinese words, and picture is a well-known logo picture, such as Sony or Motorola.

In order to investigate the effect of the layout of 5 different locations (center and 4 quadrants) and the forms of information (text and picture) on users' visual search efficiency, we designed two Web pages on every location, thus the total number of Web pages was 10. The topics of the 10 Web pages were common to people involving mobile, diet, clothing and so on. In the two Web pages of each location, the information form on one page was text, and that on the other page was a logo picture. Each search target was presented only once on every Web page. In our study, all participants didn't know the target forms, and the target was described by the text on the pre-page. Moreover, all the participants were randomly divided into 10 groups.



Fig. 1. Five locations defined on a Web page. The information form on each location is text or a logo picture, and participants don't know the information form before the experiment.

Procedure. At the beginning of the experiment, participants were given the target description on the pre-page (whether the target was text or picture, the target was described as the text, and all participants didn't know the information form); then participants should find out the target on the search-page and if they find it, they would click the left mouse button on the target. After that, participants could have a rest and then they could click into the next task. The eye-tracker recorded the whole process of participants' visual search.

2.2 Experiment 2: Impact of the Floating Ads on Visual Search

Participants. The participants were 50 undergraduates and postgraduates with the age range of 18 ~ 27 years old ($M = 23.0$, $SD = 2.0$), and 26 were female. All participants were right-handed, had normal or corrected-to-normal vision, often surf on the Internet and were skilled users of the mouse. None of them had the experience on eye movement experiment.

Apparatus. The apparatus were the same as Experiment 1.

Experimental Material and Design. Floating ads were added on the base of the experimental material in Experiment 1. The size and floating speed of the floating ads were the same as the real situation, the pathway of the floating ads were different on each Web page and the floating ads appeared on random position. And others of the design were the same as Experiment 1.

Procedure. The procedure was the same as Experiment 1.

3 Results

In the study, we analyzed fixation duration during participants searching text or logo picture which was arranged on each location including the center area, quadrant I, II, III and IV on Web pages.

The fixation duration is the sum of all the fixations length while completing a visual search task, and the shorter the fixation duration is, the higher the visual search efficiency would be. Eye-movement researchers have established 100 ms as the minimum amount of time necessary for a pause to be considered a fixation [15] - [17]. Therefore, this study based on the measures that the fixation duration with a minimum threshold of 100 ms.

3.1 Relation between Information Forms and Locations

By analyzing the fixation duration, there was a significant main effect on information forms [$F(1,490) = 76.64$, $P < 0.001$], and also a significant main effect on locations [$F(4,490) = 8.34$, $P < 0.001$]. There was also a significant interaction between information forms and locations [$F(4,490) = 11.16$, $P < 0.001$].

Fixation Duration in Text vs. Logo Picture on Each Location. As shown in Fig. 2, there was a significant difference between text and logo picture on the fixation duration when the target was on each location. Except the quadrant III, the fixation duration in text was significantly longer than logo picture, which indicates that picture has superiority to text on all the locations except on quadrant III.

Fixation Duration in Text on Each Location. Figure 3 shows the fixation duration in text on each location. The fixation duration in the center area or quadrant IV was significantly longer than in other quadrants. However, there was no significant difference between the center area and quadrant IV. And there was also no significant difference between every two among quadrant I, II and III. Therefore, the results show that it is the worst to arrange text in the center area or quadrant IV.

Fixation Duration in Logo Picture on Each Location. Figure 4 shows the fixation duration in logo picture on each location. The fixation duration in quadrant II was significantly shorter than in other locations. However, the quadrant III was significantly longer than other locations. The results indicate that it is the worst to arrange the logo picture in quadrant III.

3.2 Effect of Floating Ads on Locations

To investigate the effect of floating ads on locations, we analyzed the fixation duration between the Web page with floating ads and without floating ads on each location. No matter what location the information is placed on the Web page, there is no significant difference between with floating ads and without floating ads on the fixation duration, which indicates that floating ads has no significant effect on the locations (see Table 1).

4 Discussion

4.1 Relation between Information Forms and Locations

Glaser WR. and Glaser MO. reported that picture had superiority effect [18], which had been demonstrated in early research [19]. When users viewing picture and text which have a certain relation in content, they usually have a glance at picture first, then read text, and view picture again. This pattern has been reported for viewers looking at magazine advertisements containing text and picture [20]. In our study, the fixation duration on logo picture is significantly shorter than text on locations but quadrant III (see Fig. 2). The result also indicates that picture has superiority effect to text.

Previous studies showed that visual search behavior sometimes would be guided by the physics location in display [21]. Though picture has superiority effect to text, Fig. 2 shows clearly that when information was arranged in

quadrant III, fixation duration in logo picture was significantly longer than text, which was opposite to other locations. Furthermore, Fig. 3 and Fig. 4 show that for the text, the fixation duration in quadrant IV and the center area was the longest; in contrast, for the picture, that is in quadrant III.

Early research indicated that eyes need time to extract information, and long fixation duration was correlated with extracting more information [22]. Moreover, the difficulty of extracting information also may have an effect on fixation duration. Therefore, hardly identify or dense information display need long fixation duration [23]. In the study, we considered that there was a correlation between the fixation duration and the difficulty of users' visual search. The fixation duration would be shorter if it is easier to find out the information, otherwise the fixation duration would be longer. A website usually contains a series of Web pages, which were linked by Web navigation, to place different information. The result shows that the fixation duration in quadrant II is shorter whether the information is text or picture, which suggests that important links, such as navigation, should be placed in quadrant II. It is easier and more convenient for users to find out other related Web pages and view their interested information. The results also show that the fixation duration in text placed in quadrant IV and the center area are longer, therefore it is best not to place important text information in quadrant IV and the center area, so as not to increase difficulty for users' visual search. In contrast to picture, that is the quadrant III, so it is best not to place important picture information in quadrant III, which would be advantage for users' visual search.

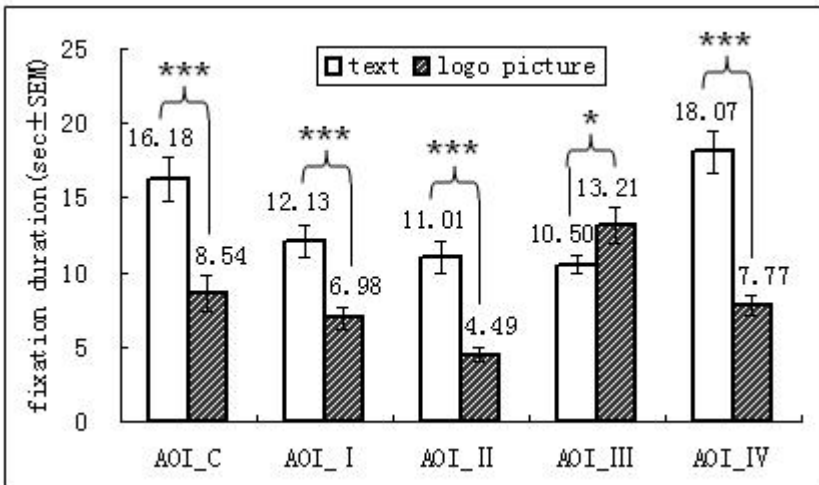


Fig. 2. The fixation duration when the text and logo picture information were on 5 locations. The fixation duration in logo picture was significantly shorter than text on all the locations except on quadrant III (*: the significant level of F-check < 0.05; *** : < 0.001).

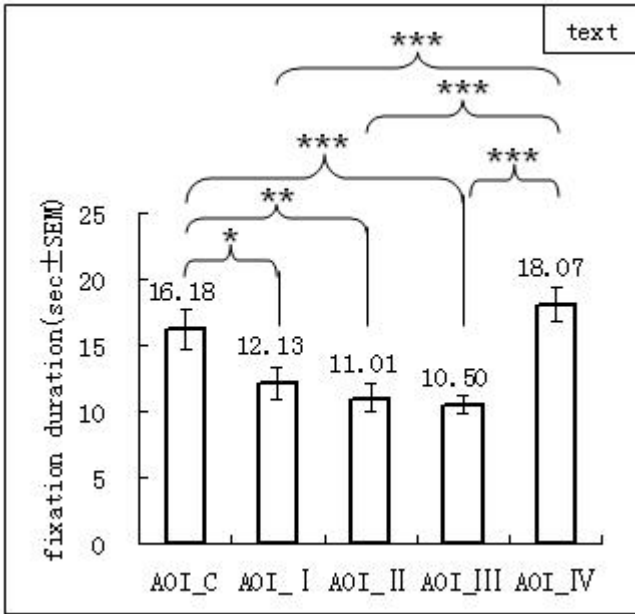


Fig. 3. The fixation duration in text on each location. The fixation duration in the center area or quadrant IV is significantly longer than in quadrant I, II and III (*: the significant level of F-check < 0.05; **: < 0.01; *** : < 0.001).

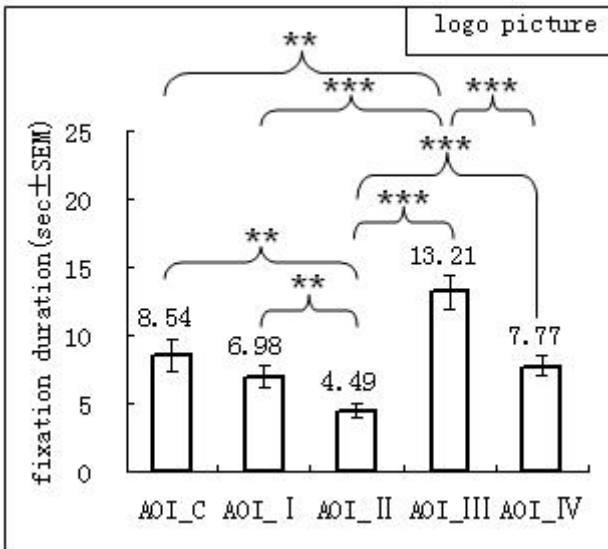


Fig. 4. The fixation duration in logo picture on each location. The fixation duration in quadrant II is the shortest and in quadrant III is the longest (**: the significant level of F-check < 0.01; *** : < 0.001).

Table 1. The fixation duration of each location on the Web page without floating ads and with floating ads ($M \pm SEM$), there was no significant difference between with floating ads and without floating ads

Location	Without floating ads(s)	With floating ads(s)	Significant level
AOI_C	12.01 \pm 1.70	11.57 \pm 1.64	$F(1,98) = 0.10, P = 0.75$
AOI_I	9.80 \pm 1.39	9.70 \pm 1.37	$F(1,98) = 0.01, P = 0.92$
AOI_II	7.82 \pm 1.11	7.95 \pm 1.12	$F(1,98) = 0.02, P = 0.89$
AOI_III	11.76 \pm 1.66	14.11 \pm 2.00	$F(1,98) = 3.87, P = 0.05$
AOI_IV	12.94 \pm 1.83	11.40 \pm 1.61	$F(1,98) = 1.70, P = 0.20$

Through analyzing the first fixation, we found that 90% of the participants' first fixation was in the center area. The result shows that the central area is the first location that participants run their eyes over the Web page. According to the inhibition of return (IOR) theory [24], participants' attention returned to the center area was inhibited. Therefore, participants' visual search in this area needs to spend more time, and it is not easy to find out the target quickly. The study suggests that when text is placed in the center area, it is best to use larger size or bright-colored title which can distinguish the main text to help users to search their desired information quickly.

4.2 Impact of Floating Ads on Each Location

As shown in Table 1, there is no significant difference on fixation duration between without floating ads and with floating ads when the target was on each location, which also demonstrated the recent studies [9][11] that users often consider that the floating ads have no relation to their desired information, and don't pay more attention to them, thus the floating ads have no significant impact on visual search efficiency.

5 Conclusion

This study investigated the relation between information forms and locations, as well as the floating ads for the impact of a Web page layout on visual search. Through analyzing the fixation duration, the results show that for the text, the quadrant IV (right lower) and the center area are longer than other locations; for the picture, quadrant III (left lower) is the longest. Whether the target is text or picture, quadrant II (left upper) is shorter. And no matter what location the information is placed on the Web page, there is no significant difference between the page with floating ads and without floating ads. The results indicate that the important information whether text or picture is suitable to place in quadrant II; however, it is the worst to arrange text information in quadrant IV and the center area or picture information in quadrant III on the layout of Web pages.

Acknowledgements

This work is partially supported by the National Science Foundation of China (No. 60775039 and No. 60673015) and the grant-in-aid for scientific research (No. 18300053) from the Japanese Ministry of Education, Culture, Sport, Science and Technology, and the Open Foundation of Key Laboratory of Multimedia and Intelligent Software Technology (Beijing University of Technology) Beijing.

References

1. Shih, H.M., Goonetilleke, R.S.: Effectiveness of Menu Orientation in Chinese. *Human Factors* 40, 569–576 (1998)
2. Byrne, M.D., Anderson, J.R., Douglass, S., et al.: Eye Tracking the Visual Search of Click-down Menus. In: *Conference on Human Factors in Computing Systems-Proceedings*, pp. 402–409 (1999)
3. Halverson, T., Hornof, A.J.: Local Density Guides Visual Search: Sparse Groups are First and Faster. In: *Proceedings of the 48th Meeting of the Human Factors and Ergonomics Society*. HFES Press, New Orleans (2004)
4. Jay, C., Stevens, R., Glencross, M., Chalmers, A., Yang, C.: How People Use Presentation to Search for a Link: Expanding the Understanding of Accessibility on the Web. *Universal Access in the Information Society* 6, 307–320 (2007)
5. Djeraba, C., Stanislas, L., Dan, S., Sylvain, M., Nacim, I.: Eye/gaze Tracking in Web, Image and Video Documents. In: *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 481–482 (2006)
6. Sheree, J., Michael, E.H.: Visual Attention to Repeated Internet Images: Testing the Scanpath Theory on the World Wide Web. In: *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA 2002)*, pp. 43–48 (2002)
7. Jakob, N.: F-Shaped Pattern For Reading Web Content. Jakob Nielsen's Alertbox (2006), http://www.useit.com/alertbox/reading_pattern.html
8. Li, M., Zhong, N., Lu, S.F.: Exploring Visual Search and Browsing Strategies on Web Pages Using the Eye-tracking. *Journal of Peking Polytechnic University* (in press, 2009)
9. Li, M., Zhong, N., Lu, S.F.: A Study About the Characteristics of Visual Search on Web Pages. *Journal of Frontiers of Computer Science & Technology* (in press, 2009)
10. Zhang, P.: The Effects of Animation on Information Seeking Performance on the World Wide Web: Securing Attention or Interfering with Primary Tasks? *Journal of the Association for Information Systems* 1, 1–28 (2000)
11. Li, M., Yin, J.J., Lu, S.F., Zhong, N.: The Effect of Information Forms and Floating Advertisements for Visual Search on Web page: An Eye-Tracking Study. In: *International Conference on Brain Informatics (BI 2009)* (in press, 2009)
12. Russell, M.C.: Hotspots and Hyperlinks: Using Eye-tracking to Supplement Usability Testing. *Usability News* 72 - Russell 7, 1–11 (2005)
13. Guan, Z.W., Cutrell, E.: An Eye Tracking Study of the Effect of Target Rank on Web Search. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 417–420 (2007)
14. Joachims, T., Granka, L., Bing, P., Hembrooke, H., Gay, G.: Accurately Interpreting Clickthrough Data as Implicit Feedback. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161 (2005)

15. Baron, L.: Interaction Between Television and Child-related Characteristics as Demonstrated by Eye Movement Research. *Education, Communication and Technology: A Journal of Theorg Research and Development* 28, 267–281 (1980)
16. Fischer, P.M., Richards, J.W., Berman, E.F., Krugman, D.M.: Recall and Eye Tracking Study of Adolescents Viewing Tobacco Ads. *Journal of the American Medical Association* 261, 90–94 (1989)
17. Stark, L.W.: Sequences of Fixations and Saccades in Reading. In: *Eye Movements in Reading*, pp. 151–163. Elsevier Science Inc., Amsterdam (1994)
18. Glaser, W.R., Glaser, M.O.: Context Effects in Stroop-like Word and Picture Processing. *Journal of Experimental Psychology: General* 118, 13–42 (1989)
19. Houwer, D.J., Fias, W., d'Ydewalle, G.: Picture-word Stroop-like Effects: A Test of the Glaser and Glaser (1989) Model. *Psychol Res.* 56, 293–300 (1994)
20. Rayner, K., Rotello, C.M., Stewart, A.J., Keir, J., Duffy, S.A.: Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements. *Journal of Experimental Psychology* 7, 219–226 (2001)
21. Megaw, E.D., Richardson, J.: Target Uncertainty and Visual Scanning Strategies. *Human Factors* 21, 303–316 (1979)
22. Bellenkes, A.H., Wickens, C.D., Kramer, A.F.: Visual Scanning and Pilot Expertise: The Role of Attentional Flexibility and Mental Model Development. *Aviation, Space, and Environmental Medicine* 68, 569–579 (1997)
23. Machworth, N.H.: Ways of Recording Line of Sight. In: *Eye Movements and Psychological Processing*, pp. 173–178. Erlbaum, Hillsdale (1976)
24. Posner, M.I., Cohen, Y.: Components of Visual Orienting. In: *Attention & Performance X*, pp. 531–556. Erlbaum, Mahwah (1984)

Human Characteristics on Length Perception with Three Fingers for Tactile Intelligent Interfaces

Haibo Wang¹, Jinglong Wu^{1,2}, and Satoshi Takahashi¹

¹ Graduate School of Natural Science and Technology, Okayama University,
3-1-1 Tsushima-naka, Okayama, Japan

² The International WIC Institute, Beijing University of Technology, China
wu@mech.okayama-u.ac.jp

Abstract. When an object is touched, the properties like shape, temperature, rigidity, texture and weight should be perceived. As a basic research of human tactile, it is most important to know how humans could perceive the shape of the object with fingers. Thus length and contacting curvature of an object are thought to provide important perception information. Most of the length perception studies are about two fingers using the thumb and the index finger. There are scarce perception studies using multi-fingers grasping. In present study, to investigate the human tactile length perception mechanism with multi-fingers, we develop a tactile length interface device for one and two lengths with three fingers and conduct two experiments of one and two lengths perception with three fingers grasping. The results showed that within the length range of 40~100mm, the one length perception was better than two lengths length perception and in two length perception, better perception took place when shorter lengths were presented to the thumb and the index finger compared to being presented to the thumb and the middle finger. As a result of force sensor, grip forces show no significant difference in the range of stimulating length.

1 Introduction

When the shape of object is perceived by fingers with human tactile, length and contacting curvature of object are thought to provide important perception information. Here we investigate the psychophysics of object length perception with the three fingertips.

At a first approximation, the length of an object grasped by the hand depends solely on proprioceptive information from joint, muscle spindle, and skin afferents (Burke et al. 1988) [4] signaling the distance between the digits; for large objects the fingers are spread apart widely and for small objects the fingers are close together. However, cutaneous information from mechanoreceptors must also play a role because object size perception requires contact between the hand and the object.

Tactile object recognition involves the perception of an object's shape and texture. The local two-dimensional form and texture of the object surface are conveyed to the Central Nervous System by cutaneous mechanoreceptive afferents that innervate the skin (Johnson and Hsiao 1992) [2].

The experiment that perceived the length with gripping the aluminum cylindrical of length within the range of 25-100mm by using the thumb and the index-finger was

conducted (H.F.Gaydos 1958) [5]. Aluminum cylindrical which standards of 10, 30 and 50mm gripped with thumb and index-finger to measure the length perception threshold (A. G .DIETZE 1961) [6]. In order to investigate the length perception mechanism, thumb and other digits were selected for length perception experiment. (Lu 2004) [7]. However, the information of the weight is also included when gripping stimuli in general length perception experiment. In present study, we adopted a new device to exclude the effect of object's weight.

The other factors which may influence the length perception are concerned. When perceiving length with gripping, the object surface is touched, the information of hardness and texture also have an influence on length perception. Although Gepshtein and Banks (2003) showed that the perception of length is independent of the way that the object is oriented in space, it is unclear how or whether object length is affected by changes in contact area or force. L.J.Berryman (2006) [9] conducted an experiment by gripping length stimuli with thumb and index finger in the range of 50~62mm length and varied contact area. They found that objects exhibit size constancy such that perception of object length using haptics does not change with changes in contact area.

Therefore, the researches on length perception with thumb and index finger gripping were much reported. However, there are scarce length perception studies using multi-fingers grasping. Here we adopted new tactile interface device that presented one and two lengths stimuli with three fingers and investigated the length perception mechanism with three fingers.

2 Method of Behavioral Experiments

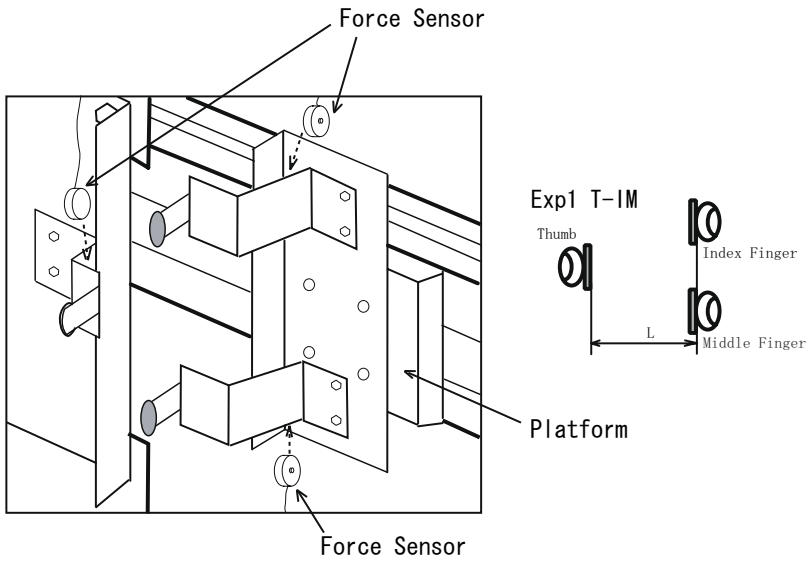
2.1 Subjects

Ten males, ranging in age from 22~30 years, participated in two psychophysical experiments. All subject reported that they had normal sensations from hands. All subjects are the students of Kagawa University.

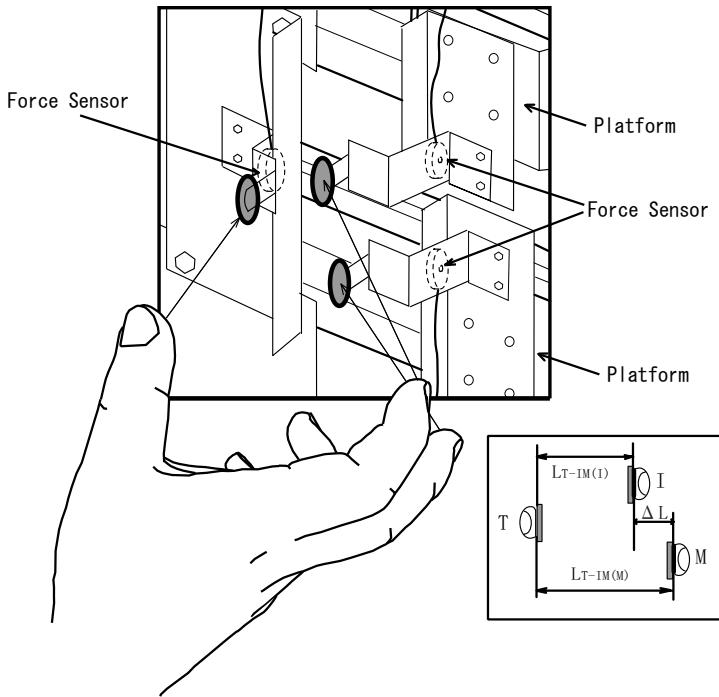
2.2 Length Presentation Device

In this study, to present the one and two lengths, first we constructed a new device that presented one length stimuli with an electric slider shown in Fig. 1(a). The electric slider was set on the board which is vertically placed, and the size of the wooden board is 90×20cm. The controller of the electric slider was also set on the board. We constructed another device that presented two lengths stimuli with two electric sliders. The hand posture of subjects and configuration of two lengths in our present device system is shown in Fig. 1(b).

As shown in Fig. 1(b), the subject gripped the stimuli with thumb, forefinger, and middle finger. There was an acrylic disc contact with the fingertips respectively and separately. Three acrylic discs which were set up in a certain distance could present two lengths for three fingers. Aluminum frameworks could support the acrylic discs. The force sensors which measured the grip force of the fingertips were set under the aluminum frameworks. The grip forces of fingertips were recorded in PC. The aluminum frameworks were set on electric sliders' moving platform, so we could adjust the distances between three acrylics continually to get optional two lengths stimuli by moving the two platforms.



(a) Device for one length perception experiment

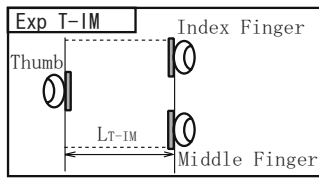


(b) The subjects hand posture of two lengths perception

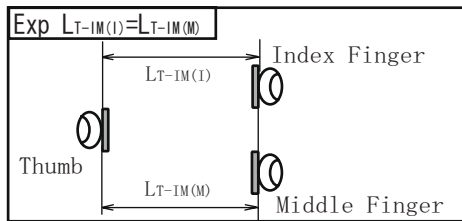
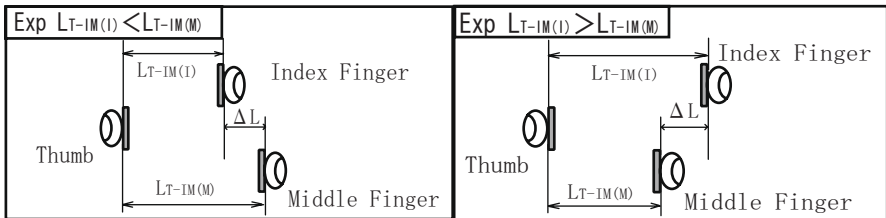
Fig. 1. An illustration for our experimental device for one and two lengths perception with three fingers

2.3 Two Kinds of Length Experiment

As shown in Fig. 2(a), one length was presented. The length presented to thumb index finger and middle finger was defined as LT_{-IM} . With the larger lengths ranging 10~100mm at intervals of 5mm, each length gripping with three fingers was test five times randomly. For entire five blocks, each one subject took total $19 \times 5 = 95$ trails. As shown in Fig. 2(b), two difference lengths were presented. The length presented to index finger and the thumb was defined as $LT_{-IM(I)}$. The length presented to the middle finger and the thumb was as $LT_{-IM(M)}$. Difference of two lengths was defined as ΔL . There were two conditions of two lengths experiment which were $LT_{-IM(I)} < LT_{-IM(M)}$, $LT_{-IM(I)} > LT_{-IM(M)}$ and $LT_{-IM(I)} = LT_{-IM(M)}$. With the larger lengths ranging 10~100mm at intervals of 10mm, the according smaller lengths were 0, 2, 4, 6, 8, 10, 12mm less (only 0, 2, 4mm less when the larger length was 10mm). Each pair of two lengths gripping with three fingers in this experiment was tested three times randomly. For entire three blocks, each one subject took total $132 \times 3 = 396$ trials.



(a) One length perception experiment



(b) Two lengths perception experiment

Fig. 2. Two kinds of length perception with three fingers grasping

2.4 Procedure and Experiment Environment

Procedure of this experiment is as followed. There were several practice trials before starting two experiments. In one length perception experiment, before taking the

length stimuli, the subject opened the thumb, index finger and middle finger, with the other fingers in natural state. The experimenter adjusted the length LT-IM with PC and indicated the beginning of experiment. The subject gripped stimuli with thumb, index finger and middle, perceived the length stimuli, and then opened fingers, told the length LT-IM which was perceived in order an mm unit. In the two lengths perception experiment, The experimenter adjusted two lengths LT-IM(I) and LT-IM(M) with PC and indicated the beginning of experiments. The subject gripped stimuli with thumb, index finger and middle finger, perceived two lengths stimuli, then told the lengths LT-IM(I) and LT-IM(M) which were perceived in order at mm unit. The experimenter recorded the number which subject answered. The grip force of each trial was also recorded by PC.

Experiments went at the sufficiently quiet laboratory. During this experiment, a curtain was set between the subject and the experiment device to cut off visual information. The subject wore eye masks during gripping. The subjects sat comfortably on chair with a slight abduction of the right upper arm (the height of the chair could be adjusted). The wrist of the subject was fixed by the magic band, and thumb, index finger and middle finger could still move freely. Thus, posture is somewhat different depending on the individual, because the subjects were told to be most relaxed during the experiment.

3 Results

Fig. 3 shows the results of one length perception experiment by three fingers. The horizontal axis is the actual length which was presented to the subject; vertical axis is the perceptual length of subject. As Fig. 3 showed, perceptual lengths of subject became larger when two actual lengths became larger. The length perception results were under 45 degree lines which proved that perceptual lengths were smaller than actual lengths, that is, the subjects' perception showed underestimation.

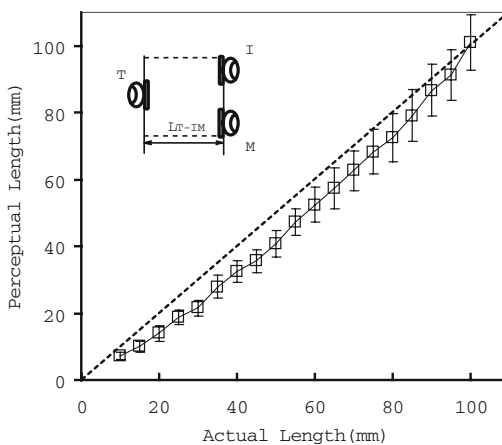


Fig. 3. Average of one length perception with thumb, index finger and middle finger

Fig. 4 shows the results of two lengths perception experiments by three these fingers when $L_{T-IM(I)} < L_{T-IM(M)}$, Fig. 4(a) is the result of $L_{T-IM(I)}$ which perceived with thumb and index finger. Fig. 4(b) is the result of $L_{T-IM(M)}$ which perceived with thumb and middle finger. The horizontal axis is the actual length which was presented to the subject; vertical axis is the perceptual error of subject which defines as $PE = PL - AL$. which PL was perceptual length and AL was actual length. The length perception results were negatives which proved that perceptual lengths were smaller than actual lengths, that is, the subjects' perception showed underestimation. As Fig.4 showed, perceptual error of subject became larger when two actual lengths became larger.

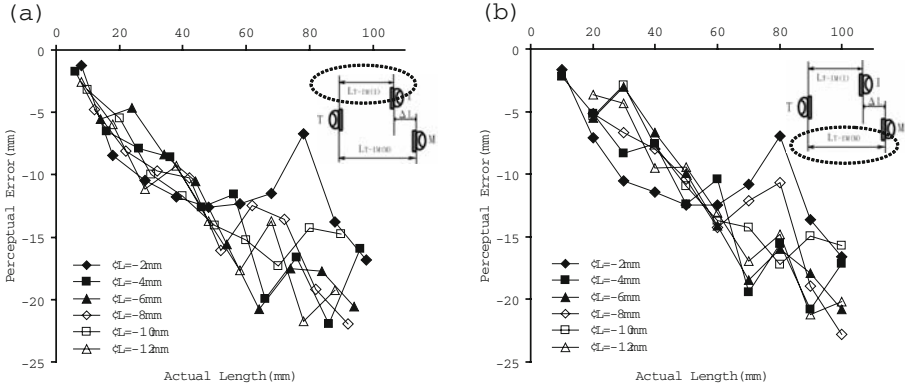


Fig. 4. Results of two lengths perception error when $L_{T-IM(I)} < L_{T-IM(M)}$

Fig. 5 shows the results of two lengths perception experiments by three these fingers when $L_{T-IM(I)} > L_{T-IM(M)}$, Fig. 5(a) is the result of $L_{T-IM(I)}$ which was perceived with thumb and index finger. Fig. 5(b) is the result of $L_{T-IM(M)}$ which was perceived with thumb and middle finger. The horizontal axis is the actual length which was presented to the subjects; vertical axis is the perceptual error of subject.

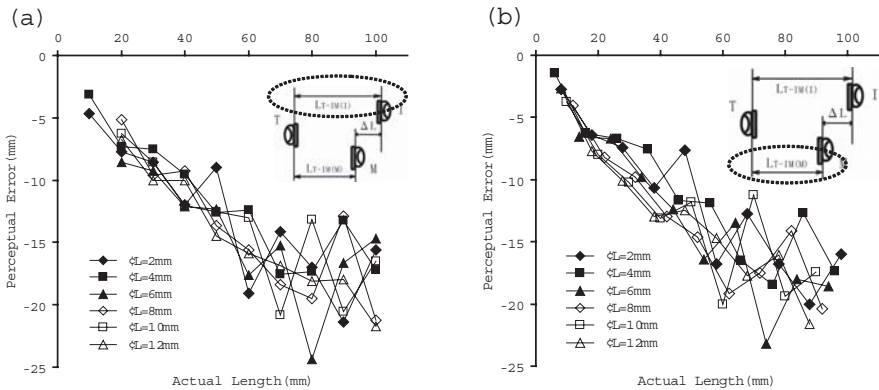


Fig. 5. Two lengths perceptual error for condition $L_{T-IM(I)} > L_{T-IM(M)}$

As shown in Fig. 5, the subjects' perception also showed underestimation, as perceptual lengths were smaller than actual lengths with the length perception results were negatives.

Fig. 6 shows the difference of one and two lengths which were perceived by subjects in length perception experiment with three fingers. The horizontal axis is the actual length which was presented to subjects. Vertical axis is the perceptual error of subject. Solid diamond mark was the perceptual error of one length perception with three fingers, hollow square mark and hollow ring mark was the two lengths perceptual with three fingers when two lengths were same. As Fig. 6 showed, within the range of 40~100mm, the tendency of two lengths perceptual errors, was enlarged when the tendency of one length perceptual error was lessened.

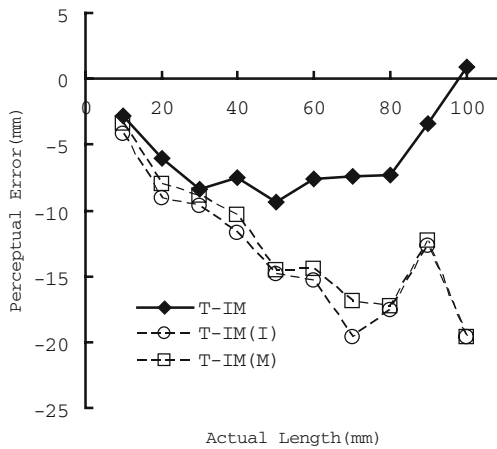


Fig. 6. Comparison of perceptual error of one and two lengths perception

4 Discussion

In our study, grip force was measured during length judgment. No effects were found on length judgment from 10~100mm by grip force. This accorded to L.J. Berryman's study which had similar results in length judgment between 50~62mm by different grip forces.

We set two conditions in two lengths perception: $LT-IM(I) < LT-IM(M)$ and $LT-IM(I) > LT-IM(M)$. From the results, it is shown that more precise judgment occurs under the condition of $LT-IM(I) < LT-IM(M)$. This may be owed to the physiological fact that the distance between thumb and index finger is less than that between thumb and middle finger.

Our previous study has shown that the perception error was enlarged by the stimulation length within the length range of 10~50mm and lessened within the range of 60~100mm during one length perception with three fingers [14]. We hypothesized that a similar tendency would happen during two lengths perception with three fingers. As a result, a similar enlarging tendency was found within 10~30mm, but a

continuous enlarging tendency was shown within 40~100mm which were out of our hypothesis. We try to make some explanations from both physiological and psychological aspects.

Physiologically, lengths can be judged and determined by perception. Westling, G. et al. [12] have pointed out that four categories of human skin receptors serve different functions during object gripping. Fast adapting with small receptive fields (FAI) acts at first object contacting and slows down at static state, while fast adapting with large receptive fields (FAII) acts only during object gripping. Slowly adapting with small fields (SAI) acts with the progressing of skin deformation and becomes null at static state, while slowly adapting with large fields (SAII) acts only at static state. It was generally considered by L.J. Berryman et al. that the perception of object size depends on the combinative information of SA1 and FA1 inputs and proprioceptive afferents. In our study, two lengths judgment within 10~30mm showed the similar enlarging tendency, as shown in previous studies, was a possible result of dominant physiological control during small length judgment. However, one more length judgment during 40~100mm in the present study was different from our previous results involving only one length judgment during 40~100mm [14]. There is no great physiological difference between one length and two lengths judgment, so we owe this discrepancy to an added psychological burden when two lengths were adopted. That is, judgment accuracy became decreased, possibly due to a dominance of psychological effects within the range of 40~100mm.

In our present study, we adopted a tactile interface device for length perception, by which two lengths judgment can be performed meanwhile and perceptive properties of two lengths judgment were got accordingly. However, our explanations, especially from psychological aspects, are far beyond convincing, and further studies like fMRI, are still needed to validate our explanations.

Acknowledgment

A part of this study was financially supported by JSPS AA Science Platform Program and Grant-in-Aid for Scientific Research (B) 21404002 from the Japanese Society for the Promotion of Science and Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Loomis, J.M., Lederman, S.J.: Tactual perception. In: Boff, K.R., Kaufman, L., Thomas, J.R. (eds.) *Handbook of Perception and Human Performance*, vol. 2, pp. 1–41. Wiley, New York (1986)
2. Johnson, K.O., Hsiao, S.S.: Neural mechanisms of tactual form and texture perceptio. *Annu. Rev. Neurosci.* 15, 227–250 (1992)
3. Gepshtein, S., Banks, M.S.: Viewing geometry determines how vision and haptics combine in size perception. *Curr. Bio.* 13, 483–488 (2003)
4. Burke, D., Gandevia, S.C., Macefield, G.: Responses to passive movement of receptors in joint, skin and muscle of the human hand. *J. Physiol.* 402, 347–361 (1988)

5. Gaydos, H.F.: Sensitivity in the judgement of size by fingerspan. *Am. J. Psychol.* 71, 557–562 (1958)
6. Dietze, A.G.: Kinaesthetic discrimination: the difference limen for finger span. *J. Psychol.* 51, 165–168 (1961)
7. Lu, S., Sakai, Y., Wu, J.-l., Yamamoto, M.: Measurement and Analysis of the Object Length Perceptive Characteristics with Visual and Tactile Information for Proposal of Three-Dimension Tactile Shape Display. *Transactions of the Japan Society of Mechanical Engineers C* 70(697), 2699–2706 (2004)
8. Vallbo, A.B., Hagburth, K.E.: Impulses recorded with microelectrodes in human muscle nerves during stimulation of mechanoreceptors and voluntary contractions. *EEG Clin. Neurophysiol.* 23, 392 (1967)
9. Berryman, L.J., Yau, J.M.: Representation of Object Size in the Somatosensory System. *J. Neurophysiol.* 96, 27–39 (2006)
10. Napier, J.: The prehensile movements of the human hand. *Bone Joint Surg.*, 902–913 (1956)
11. Westling, G., Johansson, R.S.: Responses in glabrous skin mechanoreceptors during precision grip in humans. *Exp. Brain Res.*, 128–140 (1987)
12. Goodwin, A.W.: Tactile discrimination of thickness. *Exp. Brain Res.*, 62–68 (1989)
13. Weiss, E.J.: Muscular and Postural Synergies of the human hand. *J. Neurophysiol.*, 523–535 (2004)
14. Wang, H., Wu, J.: Difference of Human Tactile Length Perception between Two and Three Fingers Grasping. In: *The Second International Symposium on Complex Medical Engineering 2008*, pp. 129–132 (2008)

A Model and Environment for Improving Multimedia Intensive Reading Practices


Thomas Bottini, Pierre Morizet-Mahoudeaux, and Bruno Bachimont

UMR CNRS 6599 HEUDIASYC
Université de Technologie de Compiègne
B.P. 20529, 60206 Compiègne-CEDEX, France

Abstract. The evolution of multimedia document production and diffusion technologies have led to a significant spread of knowledge in form of pictures and recordings. However, scholarly reading tasks are still principally performed on textual contents. We argue that this is due to a lack of semantic and structured tools: (1) to handle the wide spectrum of interpretive operations involved by the polymorphous intensive reading process; (2) to perform these operations on a heterogeneous multimedia corpus. This firstly calls for identifying fundamental document requirements for such reading practices. Then, we present a flexible model and a software environment which enable the reader to structure, annotate, link, fragment, compare, freely organize and spatially lay out documents, and to prepare the writing of their critical comment. We eventually discuss experiments with humanities scholars, and explore new academic reading practices which take advantage of document engineering principles such as multimedia document structuration, publication or sharing.

1 Introduction

Amongst all the reconfigurations of human practices which have emerged with the spread of digital technologies, documents critical appropriation tasks take up a particular place. Indeed, crucial aspects of our intellectual life rely on the ability to perform an efficient work with documents. In this perspective, digital document technologies have brought decisive innovations, which have forever changed the way we *read* and *write* — and consequently, *think*. We can especially mention some fundamental document engineering principles, which daily feed our "document digital life": document logical structuration, semantic annotations, hypertext dynamic maps, separation of content from presentation. These concepts are pivotal in academic context, where *scholar practices* (learning, critical reading and writing) constantly call for innovative document tools. The aforementioned operations have now reach maturity for textual contents, notably thanks to the development of the Web as a "global document environment", where writing, annotation and sharing operations are strongly articulated. Concomitantly, digital cognitive technologies have also dramatically changed reading practices by bringing readers face to face with non-textual contents.

Since Vannevar Bush formalized the basics of what would become one of the founding utopias of multimedia work personal environments — the *Memex* 

—, scholars still endure a lack of efficient personal software environment to read, to analyze and to elaborate non-textual documents. Yet, knowledge is nowadays frequently embodied in lecture recordings, pictures, music scores, etc. Such documents cannot be ignored due to the fact that they do not "fit to print" [2] and to its technical tradition [3]. Due to their technical nature, typical interpretive and thorough appropriation operations — such as annotation, (re)structuration, fragmentation so as to pick quotes for a future writing, linking and spatialization to clarify content — which implies an intensive reading are not straightforward. The problem becomes even more complex if we consider the necessity of gathering all of these operations within a single software environment that enables the reader — who is also a *writer* — to carry out an intensive reading of *heterogeneous documents* aiming at *producing a new critical document* (memo, course material, dissertation, etc.).

In this paper we propose a document model which handles the aforementioned multimedia intensive reading operations, as well as an experimental software tool for its use in an academic context. In the next section, we present a quick overview of the state of the art of the "multimedia scholarly reading" problem. In section 3, we exhibit three requirements which are mandatory for an efficient model to handle multimedia intensive reading operations. The model is described in section 4, and the experimental tool that enables us to validate our theoretical approach is lastly presented in section 5.

2 Related Work

The complex "multimedia scholarly reading" problem has been studied through many approaches, but rarely for its own sake *and* in the perspective to design and implement a complete personal environment.

According to [4], scholars are poorly equipped when it comes to perform active reading tasks on hypermedia documents, despite they are nowadays part of their daily work. The authors propose a scholar audiovisual information system for linking and personal annotation of video documents, which rely on an unconstrained metadata schema. Bulterman [5] tackles the multimedia annotation problem in a document enrichment perspective. However, he only focuses on the annotation of multimedia presentations. Concerning ancient digitized manuscripts, [6] propose a model and Computer Human Interface principles to create and exploit annotations, while [7] emphasise the need to structure pictures so as to *identify* and *articulate* pertinent zones for active and scholarly reading tasks.

Researches in the field of "Spatial hypertexts" have underlined the founding role of documents spatial layout and relations in complex reading and writing activities [8,9]. In [8], Marshall and Shipman introduce Spatial hypertexts as "*most appropriate when there is no distinction between readers and writers*", while "*more prescriptive methods might hamper exploratory structuring*". It so happens that scholarly reading operations first and foremost rely on the ability to annotate, comment, cut out and freely organize contents (the scholarly reader

is a kind of writer, who "writes his/her reading"). Moreover, interpretation — as every creative process — is neither immediate nor univocal, and thus needs to be supported by a flexible reading space so as to emerge. Our theoretical analysis of scholarly reading practices as creative activity is in keeping with Nakakoji and Yamamoto's approach [10,11]. The authors address intensive reading and scholarly writing crucial aspects, and especially the space, as a mean to make structures emerge among elements. They view scholarly writing as "a process of articulation" [10].

Interviews aiming at making explicit reading practices of scholars from different communities (mainly musicologists and humanities and social sciences teachers and students) drew out the need of an environment that combines the aforementioned document operations [3].

3 Requirements for a Scholarly Reading Environment

We now expound three document requirements to formalize the need by which the previous section has been concluded.

3.1 A Semantic Approach to Intensive Reading

Requirement 1 — *The model must support a semantic and structured approach of the intensive reading process, from the thorough appropriation to the elaboration of a new critical document.*

These concepts, which are the very basis of *What You See Is What You Mean* (WYSIWYM) document *elaboration* tools, are here exploited to design more efficient *reading* tools. Interviews of researchers, students and musicologists who have to analyze recordings or pictures have shown that they generally use non-semantic tools, such as sound or graphics editing softwares. These softwares appear not to be suited for conducting intensive reading and scholarly writing processes. Indeed, these are complex cognitive processes, which imply rich interpretive operations such as structuration, gloss, critical comparison, fragmentation and recombination. Handling these complex processes calls for a model which can articulate them in a single environment. As a result of this WYSIWYM-based approach, the model must be exploitable by a publishing tool to create the resulting critical documents (an example is given in section 5). Moreover, it will then satisfy the separation of content from presentation principle, which is the keystone of multi-media publishing tools (for the benefits of such a paradigm, see [12]).

3.2 Performing Unified Interpretive Operations on Heterogeneous Documents

Requirement 2 — *The model must make possible the whole spectrum of interpretive operations that a thorough active reading requires for pictures as well as recordings and textual contents.*

The reader must be able to perform the following operations on documents or fragments, regardless of their nature: 1. localize, qualify and structure (the "three steps of indexing", [13]); 2. annotate/enrich and link (the very basis of gloss).

Interviews have shown that scholars mainly rely on a common word processor to perform most of these operations. Therefore they often neglect pictures or recordings because they are not able to "have a grip" on them.

3.3 Structured Reading and Writing Spaces

Requirement 3 — *The model must support a structured reading and writing space to organize and articulate documents parts, annotations and ideas that emerge while the corpus is glanced through.*

The reader has to freely compare, categorize and reorganize the documents fragments to make new ideas or unexplored interpretive trails emerge. An active and intensive reading relies on a *content appropriation* process along with a *content creation* process. Since the shape of the final document is unstable throughout the reading, the model must provide a flexible and convenient way to organize, articulate and structure content fragments at will. Observations about the founding role of space in complex cognitive work [8,10] are greatly corroborated by the observations we made on scholars practices. For example, musicologists need to spatialize score fragments so as to bring to light thematic variations, which are undetectable in the linearity of the score¹. The model must then provide *constrained structures* (charts, lists, trees) as well as free-form spaces to organize fragments.

A critical appropriation can be a long process involving a great number of documents and interpretive trails, the complexity of which can not entirely be represented by a single resulting critical document. Thus, the model must keep track of the whole reading material — such leftover annotations and fragments or anything that could be summoned for a future reading — to share not only this final document, but the reading environment of its genesis. The model has then to support a broader acceptance of the notion of "document": from a monolithic object to a flexible, structured and semantically organized fragments-made corpus.

4 Designing a Conceptual Model

This section describes the model we have developed according to the above stated requirements. They can be achieved through combining the logical entities presented hereunder. In that purpose we use a UML class diagram, the visual expressiveness of which helps for a clearer understanding of the nature of the elements and their relations (cf. figure [1]).

¹ For an anthropological approach to the triad mind/space/writing, see [14].

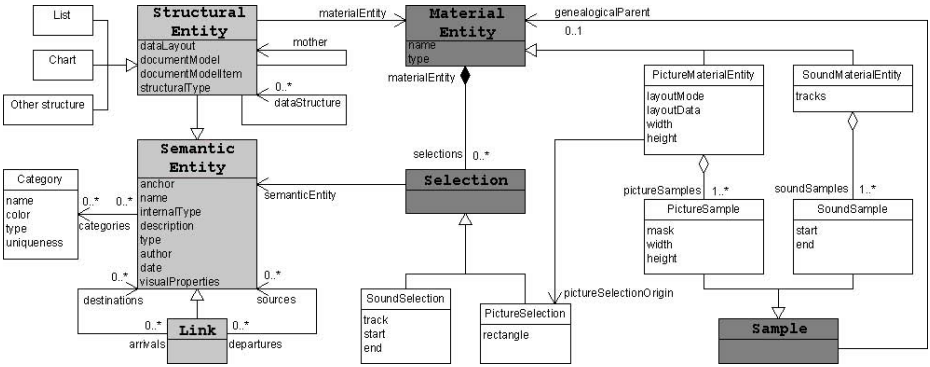


Fig. 1. A flexible model to handle interpretive operations on a multimedia corpus

4.1 Managing Heterogeneous Raw Resources

The very first operations a scholar reader may want to perform is to pick and to recompose the documents parts to study, and to reject useless ones. Working on digitized paintings or musical recordings may require, for example, to "materially" extract faces or to concatenate motives respectively. Fragments of resources files — such as a selection in a JPEG file or a segment of a MP3 or ASCII file — are stored into what we call **Samples**. **Samples** are clustered into **MaterialEntities** to handle selected fragments as a single entity. Subclasses of **MaterialEntity** and **Sample** are defined to handle the specific natures of picture, sound and textual contents.

4.2 Qualification, Annotation and Linking

Intensive reading relies on the ability to identify pertinent portions within documents. This is handled by the **Selection** class, which delineates a fragment of a **MaterialEntity**. Then, the reader may want to *describe* them with text content and metadata, to *label* them with a given category or to *link* them to other documents. These "annotation" and "gloss" aspects are handled by **SemanticEntities**, which can be associated to **Selections**, and interconnected by **Links**. A link may be glossed, consequently, **Link** is a subtype of **SemanticEntity**.

4.3 Document and Reading Space Structuration

To materialize, for example, the logical structure of a lesson, or to make emerge the pages/systems/staves material structure of a printed score, we define what we call **StructuralEntities**. They are subtype of **SemanticEntity**, and can be associated to a **Selection** to build the document hierarchical structure of unstructured **MaterialEntity**. Every interpretive unit identified in this way automatically creates its own **MaterialEntity** (which is a part of the main **MaterialEntity**). The same **MaterialEntity** may be encompassed in several

StructuralEntity, reflecting the need to process a given content in different interpretive contexts.

StructuralEntity may also be created from scratch and freely combined, with no reference to a **MaterialEntity**, to store and organize gloss or notes. Moreover, **StructuralEntity** may be used to build different kinds of interpretive structures (lists, trees, charts, etc.), according to the users' needs. In addition, any **StructuralEntity** can be viewed as a bidimensionnal space, within which sub-entities are freely laid out and scaled (the `dataLayout` field holds a list of **StructuralEntities** references associated `tdr2o` layout properties).

4.4 Implementation Issues

Data are stored in a single XML file and organized in a database way, which can be directly understandable and editable by the user. Despite the multimedia facet of the reading practices we deal with, we don't rely on SMIL (contrary to [5]). Effectively, its presentation-oriented semantics does not fit well for a work on a heterogeneous fragments-made corpus.

5 Experimentations with the Environment

Critical work supposes the possibility to vary the documents visual presentation for their better understanding and articulation. Consequently, we strongly rely

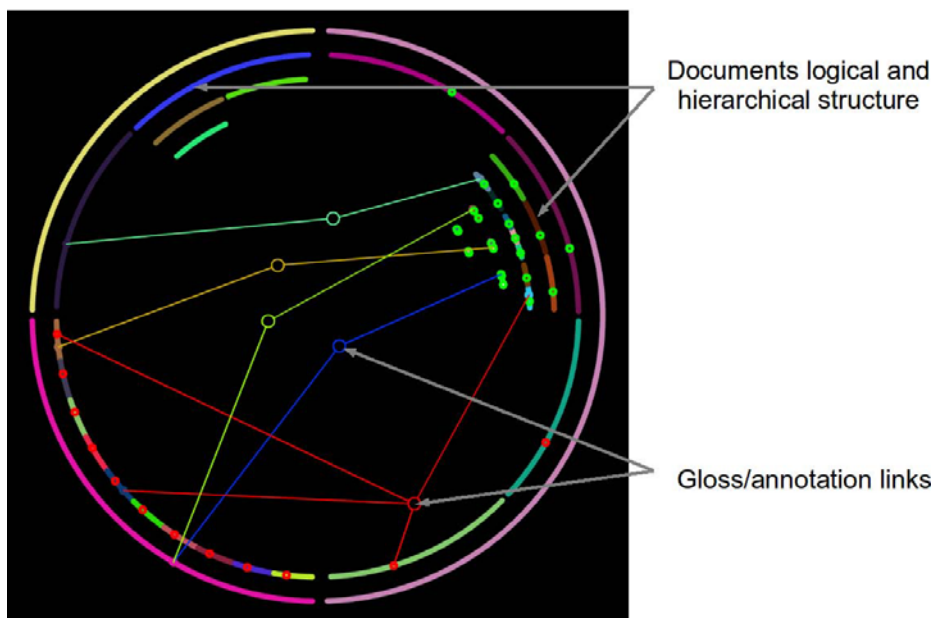


Fig. 2. Circular map: visualizing *structural* and *hypertext* relations

on the Model-View-Controller design meta-pattern to take advantage of the model. We especially focus on synoptic structures visualization. The "free" space illustrates the gain of flexibility when handling picture and text fragments (cf. figure 5). Structural (StructuralEntities) and gloss (Links) relations can be concomitantly visualized on a synthetic circular map (cf. figure 2).

As a complement to these generic instruments, we developed dedicated tools for specific reading communities needs.

5.1 Lessons Appropriation and Publication

We have developed a specific audio tool for a group of humanities students who had to produce a critical analysis of philosophy lessons recordings (cf. figure 3A). It firstly enables to build the hierarchical structure of the lessons (B). In addition, the students can freely set transversal thematic annotations (C). The Structural Entities and Selections identified in this way can then be commented (D).

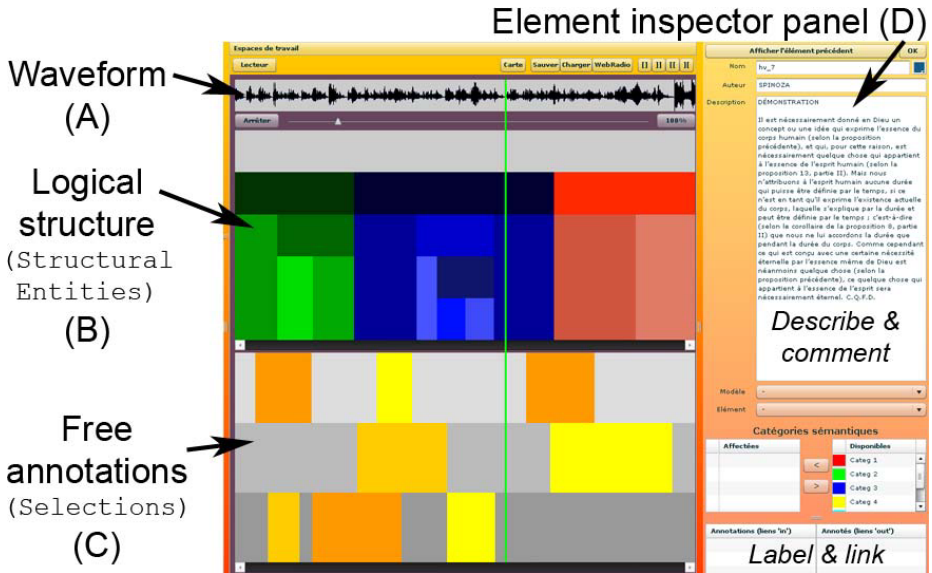


Fig. 3. Audio structuration and annotation tool

Critical appropriation of lecture recordings can here be seen as a kind of multimedia authoring activity. Indeed, the students structure and enrich the source material with text (summarizations of debates that took place during the lessons, concepts explanations, references) or picture contents so as to provide a future reader with a more complete reading and argumentative experience. In this perspective, we used the WebRadio publishing chain² to enable students to

² <http://scenari-platform.org/projects/webradio/fr/pres/>

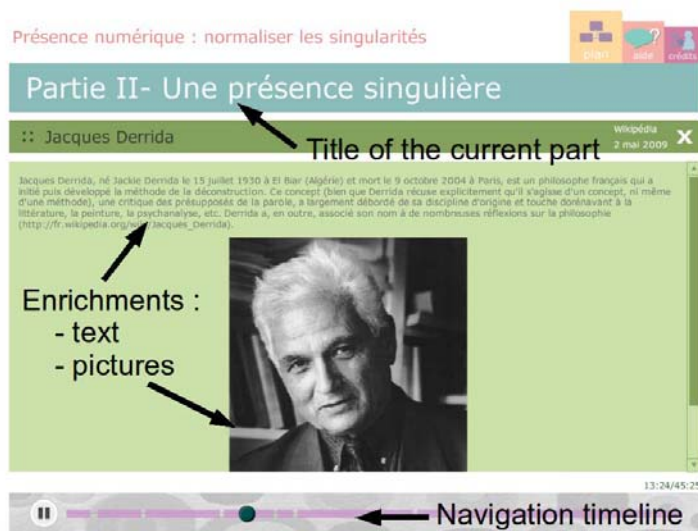


Fig. 4. An interactive flash publication of an enriched and structured audio content

export pertinent material of their reading environment as an interactive Flash online presentation (cf. figure 4).

These innovative academic practices have raised some interesting methodological issues. Beyond the satisfaction to directly deal with multimedia contents, students have to take on a double role. As students, they share their scholarly comments with teachers who are also equipped with the environment, and can then import, annotate and "review their readings". The synoptic circular representation (cf. figure 2) then offers an efficient way to perceive the gloss and corpus organization. But the students are also "publishers", who produce a multimedia online publication for a wider public, and consequently have to select and reword complex philosophical arguments. These approaches can be concomitantly conducted thanks to the flexibility of our model.

Other academic experiments are in progress with semiotics students who analyze press pictures, and computer music students who produce a multimedia comment of a piece, which articulates scores, recordings, composition softwares screenshots and textual contents.

5.2 Multimedia Musicological Analysis

We also collaborate in a participatory-design way with musicologists who annotate, cut out and produce analytical charts of musical pieces (cf. figure 5). The generic entities of the model can be derived to fit musicologists specific scholarly operations. For example, we use subtypes of `Links` to synchronize pertinent points between different scores and recordings which embody the same musical piece. The resulting "synchronized score" can then be fragmented and annotated as simply as a single-media document. `SemanticEntity` and `StructuralEntity`

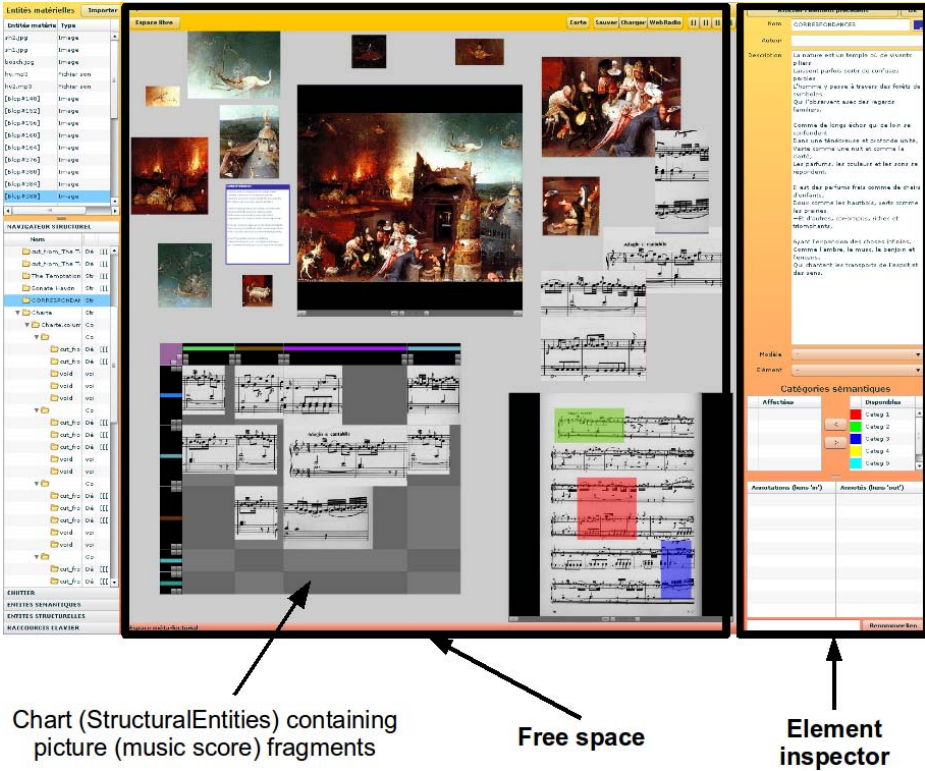


Fig. 5. A bi-dimensional workspace: free spatial positioning of elements

are used in conjunction with `SynchronizedMaterialEntities`, which articulate picture and sound samples.

Figure 5 depicts the workspace where fragments can be listened to, studied, qualified, compared and classified. The use of free bi-dimensional space enable the musicologist to freely explore and make emerge semantic relations (which is the main cognitive benefit of the spatial hypertext approach in creative work, see 2). Furthermore, more formalized spatial structure like charts (see figure 5) or lists can be used to build new analytical documents which highlights one analytical criteria or another (14). These spatial structures does offer a visual organization of the documents, but are also kind of "sound maps", since every synchronized score fragment can also be listened to.

Our environment gives rise to an epistemological shift in the century-old "score chart-making" musicological practice. Manipulating synchronized scores enables the analyst to categorize fragments by *hearing* whereas traditional practices are only based on the visual appearance of the score. Moreover, the possibility to share the analysis environment (cf. 3.3) and not only a static chart enables the analysts to interact with the whole material and to materially involve it during the critical discussions between peers.

5.3 Benefits of the Model

Our experimentations bring to light the constituent role of cognitive technologies (here, computers) in intellectual practices. Indeed, providing the readers with new scholarly operations can lead to significant evolutions in such practices. Since ideas and concepts are embodied in documents, knowledge appropriation and diffusion tasks have to rely on concrete documents manipulation in order to be performed. The model we have presented has been designed to allow the materialization of the specific structures and connections a scholar may need to establish and build to organize documents during his/her reading, according to his/her interpretive methodology and to the "interpretive conventions" of his/her scholarly community. From this point of view, we can classify the benefits of our model in two main categories : (1) the development of new scholarly reading operations and their strong articulation with more traditional ones, (2) the elaboration of hypermedia documents which are new semiotic forms that cannot be manipulated without computers.

Widening the scholarly operations spectrum. In the experimentations with philosophy students (see 5.1), we have expounded the fundamentals of what we could call an "hypermedia reading chain"³. Our generic model then acts as a "cement" in such a chain. Indeed, the generic objects (**Material, Semantic and Structural Entities, Selections, Links**) it provides build up a "milieu" inside which a wide spectrum of scholarly operations can be articulated, regardless of the material nature of the involved contents. "Traditional" scholarly operations (structuration, annotation, spatialization, description) can be finely articulated, which provides a complete interpretive experience. Moreover, the concepts of the model offer a coherent and convenient contents organization from which publication tools can transform the corpus (or part of it) into different document formats, according to the needs: interactive presentations for future readers, L^AT_EX or OpenDocument files for written article preparation, minimal software environments to share analytical results amongst peers. The experimentations with musicologists have especially shown the scientific interest of our model which enables to augment traditional score analysis operations and to contribute to the elaboration of new ones. Thus, by emphasizing the need of an open and generic model of reference, our approach favours the emergence of new knowledge manipulation and diffusion practices. Beyond the multimedia scholarly reading practices, our model intends establishing principles which can be useful for more generic active reading practices. We can remark that after twenty years of personal graphical desktop environments, operating systems still lack from unified concepts for accessing documents in a more efficient way than what common file management systems propose. The interoperability of documents and softwares could greatly benefit from a deep reflection on the essential document cognitive operations (again, annotation, structuration, fragments delimitation and access, description, categorization, hyperconnection).

³ The word "chain" connotes the idea of a reading "path", from the first appropriation of the corpus to the publication of a structured comment.

Building and interpreting new semiotic forms. The experimentations with musicologists presented in section 5.2 illustrate the basis of our approach of the hypermedia document elaboration and appropriation problem. Computers make possible the construction of new "species" of documents, such as hypermedia documents (multimedia documents which are hyper-linked or synchronized). Our model enables a reader to easily build such documents by connecting (with `Links`) relevant instants in audio documents (`SoundSelections`) and relevant zones in pictures (`PictureSelections`). The resulting documents are also semiotic forms carrying concepts and ideas, which are thereby susceptible to be interpreted. The generic scholarly concepts of the model (annotations structuration, categorization...) can then be put into practice, regardless of the nature of the contents (see section 3.2).

6 Conclusion and Future Works

In this paper, we have presented a model and a complete personal environment for supporting intensive reading operations on heterogeneous multimedia documents. They enable to better understand, comment and contextualize such a reading material, and to prepare its sharing and critical discussion among a reading community. Our approach consists in putting document *authoring* traditional concepts into practice. These concepts — publication, resources files editing and structuration, free creation/articulation/spatial layout of elements — enable to design more efficient *reading* tools. New academic document practices can now be explored, thanks to the genericity and the flexibility of our model. Moreover, it makes the critical analysis of hypermedia documents possible.

The next step will be to provide the users with a built-in multimedia publication tool, in order to make them able to produce and to parametrize by themselves structured and enriched presentations based on SMIL⁴ or Adobe Flex formats. Such a device could enable the "digital scholar" to have a complete control on the whole reading process, from the appropriation to the critical comment writing, publication and discussion.

Acknowledgments

This work has been developed in the scope of the Poliesc project, funded by the Region Picardie and the European Social Funding Program.

References

1. Bush, V.: As we may think. *The Atlantic Monthly* 176(1), 101–108 (1945)
2. Ingraham, B.D.: Scholarly rhetoric in digital media. *Journal of Interactive Media in Education*, JIME (September 2000)

⁴ Synchronized Multimedia Integration Language, <http://www.w3.org/AudioVideo/>

3. Bottini, T., Morizet-Mahoudeaux, P., Bachimont, B.: Instrumenter la lecture savante de documents multimédia temporels. In: Holzem, M., Trupin, E. (eds.) 11ème Colloque International sur le Document Electronique, Paris, Europia productions (September 2008)
4. Aubert, O., Prié, Y.: Advene: Active reading through hypervideo. In: HYPERTEXT 2005: Proceedings of the Sixteenth ACM Conference on Hypertext and hypermedia, pp. 235–244. ACM Press, New York (2005)
5. Bulterman, D.C.A.: Using smil to encode interactive, peer-level multimedia annotations. In: DocEng 2003: Proceedings of the 2003 ACM Symposium on Document engineering, pp. 32–41. ACM Press, New York (2003)
6. Doumat, R., Egyed-Zsigmond, E., Pinon, J.M., Csiszàr, E.: Online ancient documents: Armarius. In: Bulterman, D.C.A., Soares, L.F.G., da Graça, C., Pimentel, M. (eds.) DocEng 2008: Proceedings of the 2008 ACM Symposium on Document engineering, pp. 127–130. ACM Press, New York (2008)
7. Faure, C., Vincent, N.: Document image analysis for active reading. In: SADPI 2007: Proceedings of the 2007 international workshop on Semantically aware document processing and indexing, pp. 7–14. ACM Press, New York (2007)
8. Marshall, C.C., Shipman III, F.M.: Spatial hypertext: designing for change. *Communications of the ACM* 38(8), 88–97 (1995)
9. Buchanan, G., Blandford, A.: Spatial hypertext as a reader tool in digital libraries. In: Visual Interfaces to Digital Libraries [JCDL 2002 Workshop], pp. 13–24. Springer, Heidelberg (2002)
10. Nakakoji, K., Yamamoto, Y., Akaishi, M., Hori, K.: Interaction design for scholarly writing: Hypertext representations as a means for creative knowledge work. *New Review of Hypermedia and Multimedia* 11(1), 39–67 (2005)
11. Yamamoto, Y., Nakakoji, K., Nishinaka, Y., Asada, M., Matsuda, R.: What is the space for?: the role of space in authoring hypertext representations. In: HYPERTEXT 2005: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, pp. 117–125. ACM Press, New York (2005)
12. Mikáč, J., Roisin, C., Le Duc, B.: An export architecture for a multimedia authoring environment. In: Bulterman, D.C.A., Soares, L.F.G., da Graça, C., Pimentel, M. (eds.) DocEng 2008: Proceedings of the 2008 ACM Symposium on Document engineering, pp. 28–31. ACM Press, New York (2008)
13. Morizet-Mahoudeaux, P., Bachimont, B.: Indexing and mining audiovisual data. In: Tsumoto, S., Yamaguchi, T., Numao, M., Motoda, H. (eds.) AM 2003. LNCS (LNAI), vol. 3430, pp. 34–58. Springer, Heidelberg (2005)
14. Goody, J.: *The Domestication of the Savage Mind*. In: Themes in the Social Sciences, November. Cambridge University Press, Cambridge (1977)

Study on Adaptive Computer-Assisted Instruction for In-Service Training

Yu-Teng Chang^{1,2}, Chih-Yao Lo^{1,2}, and Ping-Chang Chen²

¹ School of management, Huazhong University of Science & Technology

² Department of Information Management, Yu Da University, Miaoli County,
Taiwan 361, R.O.C.

{cyt, jacklo, pcchen}@ydu.edu.tw

Abstract. E-learning is not perfect and it is difficult to construct adaptive e-learning websites. The effect of e-learning depends on different people; there is the lack of learning motives and the evaluation on individual learning effectiveness is incomplete. Thus, the inappropriate use of e-learning will negatively affect learning. Knowledge acquisition is connected with the people's personality and situation and significantly influenced by their personality traits, activities in the situation, environment and culture. Thus, with the result of workplace aptitude test, this study provides adaptive learning path for the employees under the influence of personality and situation, expects the employees in the firms will enhance individual adaptive learning upon adaptive mechanism, reduces complicated learning path and constructs complete adaptive e-learning environment.

Keywords: e-Learning, In-Service Training, Adaptive Computer-Assisted Instruction.

1 Introduction

In recent years, with the prevalence of internet technology and multimedia, more and more firms apply them to in-service training and intend to enhance the employees' interest and learning effectiveness by mutual communication, vivid display and sound effect. Rapid development of e-learning and the progress of learning approaches aim to make learning simpler, more interesting and more effective. The learners should not feel satisfied with the classroom learning model upon textbooks or other written media and they must demand for professional courses, development of multiple learning activities and integration of varied media (flash cards, pictures, recorders, video players, computers, projectors, etc.). The original materials and training approaches should be broken through and employee-oriented adaptive courses should be developed to result in multiple adaptive learning and the learning model with multiple, vivid and personalized language without the limitation of space.

There are many studies on personal adaptive learning and varied obstacles to construct an adaptive learning website. For instance, learners tend to be lost in the internet during the learning process and excess connections will also become their burdens. Thus, "to be lost" and "cognitive burden" are the major problems of hyper media studies. Besides, teaching materials of most of the instruction websites are not flexible

and diverse. Even though there are rich teaching materials, the websites cannot provide proper ones to the learners according to their learning style and knowledge. In other words, these websites cannot teach the learners in accordance with their aptitude and provide adaptive materials.

Thus, personal adaptive e-learning should still be improved. Likewise, the firms should also enhance e-learning introduced as in-service training for the employees to facilitate the employees' active learning to the optimized learning model and personal adaptive learning. Therefore, this study intends to propose a corporate support system of computer-assisted instruction to allow the employees to clearly realize that the e-learning aims to enhance the employees' work ability and reduce training cost by effective e-learning to ultimately accomplish the goal of long-term adaptive learning. This study attempts to construct a support system of computer-assisted instruction; thus, by saving the costs and time, the firms can recognize the learners' potential, characteristics and abilities by workplace aptitude test in order to study the employees' adaptive characteristics to provide proper learning and training when planning in-service training and thus allow the employees to develop the second profession according to their advantages. With regard to different learners' characteristics (such as age, gender and educational background), this study shows teaching materials by XML semantic web and combines the learning effects upon knowledge management and XML semantic web. This study intends to function as the criterion for the adaptive learning introduction and educational training for the firms by these techniques in order to reduce complicated learning path and construct complete adaptive e-learning.

2 Literature Review

2.1 Advantages of e-Learning Introduction in the Firms

Employee training by e-learning in business was the trend around the world. The advantages of e-learning for the firms included below: knowledge exchange among different regions, enhancement of knowledge spread of different areas with minimum time and costs; delivery of corporate task and new production information so that the salespersons and new employees could immediately obtain new production information; rapid knowledge training, employee upgrading, enhancement of the employees' professional fields and knowledge management training, rapid response to severely changeable environment and times; product training between upstream and downstream companies, etc. Besides, the benefits of e-learning for the firms involved the levels below: saving training cost was the basic motive of e-learning introduction for almost all firms; to reduce training time and accelerate the efficiency of employees training. In addition, knowledge management by e-learning could save and share tangible and intangible knowledge and even construct new knowledge and thus it could create more values for the firms. Once knowledge was properly used and managed, it would enhance the employees' work efficiency and product yield rate, reduce operational cost and even enhance product R&D knowledge, sales ability and business volume [1].

In the development of e-learning, learning pattern changed from "portal websites", "vertical websites" to "personal adaptive learning websites". Learning patterns in the firms should meet the demands of different learners and sometimes learning content

should be changed according to learners' situations, by experience learning instruction and learning, identification and use of learners' experience; thus, the start and end of the courses would be more flexible. The learners could obtain maximum learning benefits in the shortest time upon rich online resources and media. This study intended to probe into corporate e-learning according to the employees' personal needs, experience and situation and find if the firms could recognize the employees' adaptive characteristic when planning educational training by workplace test to further arrange proper learning and training for the employees. In the future, the employees' personal adaptive learning and effective training learning would significantly influence the operational effectiveness of the companies.

2.2 Correlation between Learning Style and Workplace Aptitude Test

2.2.1 Learning Style

Learning style referred to the learners' preference in the process of learning. The scholars suggested that it was the individual preference in learning. Some studies indicated that learning style was the important index to predict computer end-users' learning results. With regard to the learners with different styles, it was extremely important to select proper learning model as the training [2]. The scholars' definitions on learning style were described below:

Gregorc[3] indicated that the "style" of learning style referred to the individuals' preference to adapt to the environment through inborn coding system, surrounding culture and subjectivity. It was the common characteristic of behavior and would not be changed with time, places, targets or content. The style would be shown by the individuals according to their preference.

Huang [4] suggested that learning style was inherited and also affected by the environment. Some styles were stable and some would change with the environment.

Su [5] generalized the following concepts with regard to learning style:

- (1) Learning style was constructed upon the interaction among individuals, cognition, affection, society and environment.
- (2) Learning style demonstrated the learners' unique learning preference of inclination. There was no standard with regard to learning style which led to different evaluations in different time and places.
- (3) Learning style was the learners' strategy or preference when dealing with learning or problems. Different people had different styles. Thus, it was unique.
- (4) Learning style was developed in longer term and it could hardly be changed in short time. Thus, learning style was consistent and stable.

Wu [6] generalized several major concepts of learning style:

- (1) Learning style referred to the learners' special preference or inclination.
- (2) The preference could be inherited or influenced by the environment.
- (3) Each learner's learning style was unique.
- (4) It was consistent and stable in learning situation. In other words, in short time, it would not be changed in short time.
- (5) It was fixed for learning content. In other words, learning style was the common characteristic of the individuals and it would not change in short time because of different learning content.

Keefe [7] defined learning style below: stable reaction of the learners in the interaction with the learning environment. It tended to involve individual cognitive pattern, affective characteristics and physical habits.

Many studies suggested that different learning models and research targets would influence learning style differently. The variables of learning style were indicated by the scholars below:

Gregorc [3] suggested that when the teachers' instruction could match the learners' learning style, the learners' ability would significantly reveal and they would not resist learning. In the contrary situation, learning, the learners' ability shown would be different due to the difficulties and frustration.

Clarina and Smith [8] indicated that learning style could lead to the learners' learning types. The learners with different learning styles had different learning approaches. Thus, the learners with varied specialties could fulfill their talents in different fields.

Upon the research of Kolb on learning style theory and different learning training measures, Bostrom [2] indicated that learners with different learning styles had different learning effectiveness with regard to software such as trial balance and e-mail.

Garner indicated that the advantage to apply Kolb's learning style was that the learning process would become clear and the learners would recognize different learning approaches in different stages. The said researcher also suggested that learning style theory was not supported in psychology and thus, it was limited in different applications.

Based on Kolb's learning style theory, Wang [9] studied the learners' web page construction and demonstrated that the interaction between learning style and training significantly influenced learning performance.

The study of Chen [10] on the relationship among adult online learners' learning style, self-monitoring and learning effectiveness demonstrated below:

- (1) The performances of adult learners' learning style, self-monitoring and learning effectiveness were excellent on current e-learning.
- (2) The individual variables (such as gender, age, educational background and occupation) would influence adult online learners' learning style.
- (3) Gender and educational background would affect the adults' e-learning effectiveness.
- (4) Learning effectiveness of adult learners with "assimilating" learning style was the most significant and that of adult learners with "converging" learning style was the most insignificant.
- (5) Self-monitoring of "searching learning data" was the most influential for e-learning effectiveness. Adult online learners' learning style and self-monitoring could effectively predict e-learning effectiveness.

Based on the definitions, theories and research findings of domestic and foreign scholars, learning style referred to the learners' preference in the process of learning or the strategy when managing learning or other problems. Thus, it was unique. Many literatures also indicated that learners' learning style in specific experience, observation and reaction result in new experience in abstract concept and action.

2.2.2 Workplace Aptitude Test

Aptitude was an individuals' inborn ability which was their learning ability. It included two categories: normal ability and special ability. The test allowed the participants to predict their future learning effectiveness. Thus, for planning the career, people could join in related aptitude tests to recognize their abilities and develop their potential. The test (aptitude test) was the most common one among ability tests. Aptitude referred to an individual's potential ability before learning something. Aptitude test was used to measure the said ability. According to the scores, people's possibility and learning ability on certain fields could be predicted.

The scholars' studies on people's abilities demonstrated that the abilities among people were considerably different; besides, there was extremely significant inner difference in one person. Thus, without recognizing the advantages and disadvantages of the individuals' abilities, it would be difficult to further understand them. The psychological tests were continuously developed and it focused on not only people's intelligence, but also interpersonal relationship, motivation, emotional adaptation, interest, attitude, behavior and personality traits. They would enhance the recruitment, personnel arrangement, supervisor selection, career planning, educational training and employee consultation of the firms. Comparing with most of psychological tests, aptitude test was more suitable for the selection, arrangement, training, management diagnosis, employee guidance and career planning in the firms. Aptitude test aimed to evaluate the individuals' personality traits and it was the important criterion for the companies.

Based on above, the correlation between corporate workplace aptitude test and learning style would be different on different individuals. For instance, the extravert people with stable emotion would show significant e-learning effectiveness. Besides, the employees' personal variables such as gender, age, educational background and occupation would also influence their e-learning effectiveness.

2.3 Display of Corporate Adaptive e-Teaching Materials

XML was a kind of meta-language which aimed to describe the content of the documents. The designer could define the documents by the tags. For instance, the name of the course in teaching materials was <course>, teaching approach referred to <teach-type>, dependency among teaching materials was <dependency>. The tags were considered as metadata. Semantic internet theory suggested that human beings' memory and cognition was a network which included numerous nodes and relationships. A node was a concept or fragmental information. It could be names, characteristics or compete ideas. The nodes were connected by relationships which led to the paths. Thus, this study applied ontology to corporate educational training to fulfill the reuse by introducing ontology language in teaching material and base; besides, it also recorded the learners' learning process and represented the process by structural ontology language, XML which became the base of semantic web and thus this study could easily use it to probe into the learners' behavioral model and further provide personalized on-line learning.

3 Research Method

This study designed a corporate support system of computer-assisted instruction, selected one firm in Optoelectronics Industry upon the result of corporate workplace aptitude test and individual learning and learning style, provided adaptive learning upon individual situational learning process and further found the rules of the learning record. The research framework was shown in Figure 1.

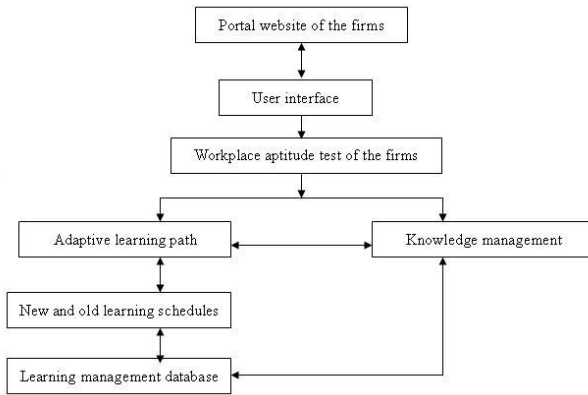


Fig. 1. Research framework

Step 1: The British psychologist Hans Eysenck suggested that personality included two stability dimensions: introvert & extravert and stable & unstable. He further defined introvert and extravert stability of personality by 12 personality traits.

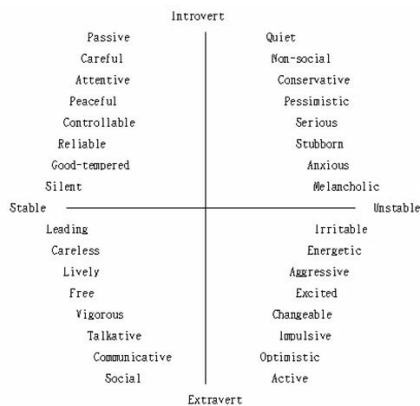


Fig. 2. Personality includes two stability dimensions [10]

Table 1. Dimension of traits[11]

Emotional stability / introvert and extravert	Items	Description	Emotional stability / introvert and extravert	Conclusion
Low	S: Social	Activeness and adaptability in interpersonal relationship; enjoying making friends, being attractive and social lives	Social	Active
Low	A: Active	Leading the opinions in the groups and active in terms of actions; feeling easy when speaking in front of others; willing to serve the group and hospitable	High	
Low	G: Generalship	Quick and dealing with things rapidly; finishing many things in short time; doing things energetically and happily	High	
Low	R: Relationship	Talkative, enjoying social lives and impulsive	Social, impulsive and active	Relationship
Enjoy thinking	T: Thinking	Cautious; enjoy thinking; guessing others' intention; pondering on difficult things; preferring planning things	Not good at thinking	
Low	Ag : Aggressive	Self-oriented and impulsive; fighting against the attack; unfriendly to others	High (self-oriented)	Social adaptability
Low	Co : Complain	Unsatisfied with reality; critical; feeling that people do not understand them	Complaining about reality	
More objective	O: Objective	Will not imagine things; practical, rational; accept the reality	prejudice	
Low	N: Nervous	Sensitive to others' comments; anxious when people are looking at them; changeable emotion; worry about trivial things	High	Emotional stability
Low	I: Inferior	Lack of confidence; anxious about inferiority	High	
Low	C: Changeable	Changeable emotion; their facial expression shows their feelings; emotionally impulsive	High	
Low	D: Depressive	Melancholic, pessimistic less energetic and anxious	High	

Table 2. Description of five types of trait dimensions [11]

Code	Types	Description	Description of personality
ES	Extrovert and Stable Emotion	Extravert and stable Emotion	Jobs with frequent interpersonal relationships and more pressure, such as salespersons, managers, customer service personnel, etc.
EU	Extrovert and Unstable Emotion	Extravert and Unstable Emotion	Jobs with frequent interpersonal relationships but with less pressure. When the participants are younger, their emotional stability should be enhanced.
M	In the middle of Diagram	Introvert and extravert, emotion is in the middle	Jobs with middle level of interpersonal relationships and pressure, such as administration.
IS	Introvert and Stable Emotion	Introvert and stable Emotion	Jobs with interpersonal relationships which requires patience; such as R&D.
IV	Introvert and Unstable Emotion	Introvert and Unstable Emotion	Jobs with less interpersonal relationships and pressure.

The advantages and functions of aptitude test for the firms are below:

- (1) Talent selection: the supervisors of personnel affairs can immediately acquire the information related to the employees' personality to reduce the unnecessary interview time and cost; properly selecting the talents by rapidly searching for the talents with personality matching different jobs.
- (2) Talent cultivation: The test result allows people with different aptitude advantages to develop the second profession; the employees can also analyze themselves and grow with this test; the test provides learning and training with adaptive teaching materials.
- (3) Talent use: It is the aptitude criterion for employee arrangement and selection; it considers the employees' different aptitude advantages for career planning.
- (4) Talent keeping: it can guide the employees.

This study probes into the cultivation of talents. For instance, "am I extravert, emotionally stable, extravert or emotionally unstable? What is my learning direction? Do I have the potential to be the salesperson (accountant)? Am I nervous or impulsive? Posing good questions is the first step to lead to useful results of psychological tests. In varied tests, when the participants select the tests matching their conditions (such as age and educational background) and answering their worries, they will finally realize their real situations.

Although there are many ways to recognize oneself and others, normalized psychological tests will demonstrate more objective information for the participants.

Step 2: This study classifies the information on the website and includes the tags which describe the content of each piece of information. For instance, the tag "business department of marketing office" is attached to "Liu H. L.". Thus, it is obvious that Liu H. L. is the employee of business department of marketing office. This study can also involve many descriptions on the attributes of "Liu H. L.", such as: <aptitude >, <age>,

<birthplace>, <learning process> and <performance >. This study shows the information related to educational training teaching materials, system and the employees. Thus, the information can be used and easily be searched by other users to accomplish the purpose of reuse.

In addition, currently, most of the information on internet is constructed in database. There are varied cases in database and many of them are repetitive for the learners. Likewise, when a person downloads one case for several times, it might be because that he forgets the saving. Thus, it should be considered as the same piece of information. Besides, when this study intend to find the specific rule of the learners' downloading in certain period of time (such as on the same day), it must transform the downloading time, consider the limitation of the same day and turn the original time into simple date for further analysis. For instance, this study can construct the employees' real experience and problem-solving process accumulated in the system in the documents in order to search for similar cases. When the solution of the case is replaced by better approach, the original solution will be renewed and it will be invalid. However, this study will not eliminate it to avoid losing the precious experience. This study then analyzes the employees' knowledge and experience to find the effective learning strategy in the shortest time. Before comparing the cases, the employees can input the weights according to the attribute indices they demand. Learning items which they are interested in can involve more weights. Thus, the cases searched will match their needs.

This study aims to provide the employees' personalized internet environment upon ontology. From the perspective of the employees, the display of user interface is upon "the employee guide system". In other words, each employee has aptitude test at user interface. Based on test result, the system shows different order of teaching materials, the level of learning questions and cases with similar attributes in database needed by the employees. Knowledge base of "employee guide system" refers to "teaching material model" and "the employee model". This study described the information in these two databases by XML semantic web; with regard to "the employee model", this study records "the employees' basic information base" and "the employees' learning process". The source of "the employees' learning process" is based on "the journal of the employees' learning process" sent back by the system. "Employee guide system" is based on the employees' information in "the employee model" and "teaching material model" for further analyzing the employees' learning and related learning characteristics and providing the employees' adaptive learning, as shown below.

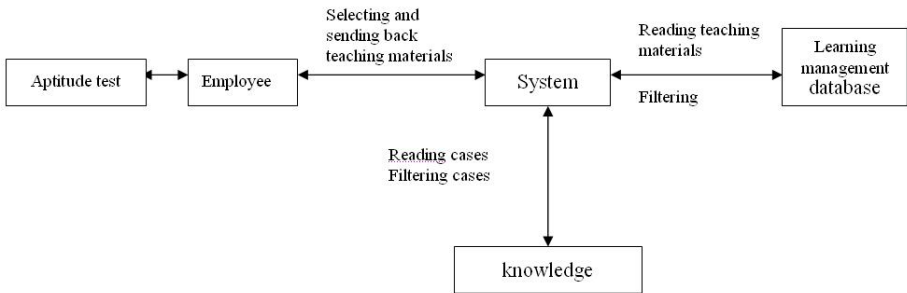


Fig. 3. Process needed by the learners for browsing teaching materials in this system

Table 3. The employees' online behavior recorded in the system

No.	Behavior	Descriptions
1	Aptitude test	Recording the advantage, disadvantage and characteristic of the employee's personality
2	Learning process	Recording the employee's background knowledge
3	Learning starting time.	Recording the time
4	Learning ending time	Recording the time
5	Time to browse certain lesson(web)	Accumulating the time
6	Frequency to browse certain lesson(web)	Accumulating the frequency
7	Order to browse each lesson (web)	Recording learning process
8	Total time to enter the system	Accumulating the time
9	Total frequency to enter the system	Accumulating the frequency
10	Testing	Questions
11	The employees' test scores	Recording learning test
12	Searching	Key words searched

Step 3: There are many cases in database and the indices for the description are based on many employees' "employee learning record". When this study decides the weights of similar cases, it can recognize more proper cases; after inputting the weight, this study calculates the similarity. For instance, the users select the attribute indices needed, such as "objective", "social", and "nervous" and adjust the weight of attributes by fuzzy. Thus, this study calculates the similarity of each piece of information and arranges the order to further show the top five pieces.

Table 4. Attribute index

Parameter	Formula
N (Number of indices)	$\frac{\sum_{i=1}^n w_i \times SIM(f_i^I, f_i^R)}{\sum_{i=1}^n w_i}$
W I (Weight of each index)	
f_i^I (Case input)	
f_i^R (Case in database)	
$SIM(f_i^I, f_i^R)$ (Attribute similarity of "case input" and "case in database" in the index)	

For instance, the user selects the attribute indices needed, such as "objective", "social", and "nervous" and adjusts the weights of attributes by fuzzy: 0.9(extremely care) 0.5(normal) 0.3(do not care). The calculation of similarity is below:

Case: ES (extravert and stable emotion) and EU (extravert and unstable emotion) are substituted in the formula,

$$SIM = \frac{0.9 * 0 + 0.5 * 0 + 0.3 * 0}{0.9 + 0.5 + 0.3} = 0 = 0\%$$

Thus, this study calculates the similarity of each piece of information and arranges the order to further show the top five pieces.

4 System Construction

In order to collect the users' learning record to analyze the learners' learning process in the system, it is necessary to identify the employees. After the employees register, they will enter the system by their personal accounts. When they input the accounts and secret codes upon the confirmation of the system, they will have the aptitude test. This system evaluates the employees' learning according to the result. It can not only rapidly and properly respond to the employees' learning, but also objectively reveal their learning to provide appropriate learning path to meet their needs. It means to enter the learning system. The system will immediately search for the employees' previous learning. When the employees finish the learning of the previous unit and will approach new learning, the system will guide them. After the employees finish the learning, they will enter the learning of the lesson. When the employees do not accomplish the original learning, the system will continue this lesson. In this learning system, the employees cannot randomly select all lessons on the website; instead, they should follow the order planned. It will prevent the employees from being lost on the websites. Besides, when the employees follow the learning schedule planned, their e-learning will match the traditional courses. With regard to the system, it can control the employees' learning quality and they will follow the adaptive learning. The employees initially learn the basic concepts to be able to approach more difficult units.

In this study, the system records the employees' learning on the website in detail, such as the frequency and staying time. It will allow the experts to analyze the employees' learning model and the system can construct the employee model (Student Model). The employees can also examine their situation and cognition through the learning process. The system records the employees' learning on the website and allows this study to analyze the influence of adaptive learning system on the employees' learning. This system record the employees' every connection with the teaching materials and the related learning paths to observe the employees' adaptive learning path. Besides, the employees' staying time at the teaching materials will also be recorded in the employees' personal file.

5 Conclusions

Currently, e-learning mostly is not based on adaptive characteristic and everyone acquires the same content. It will consume plenty of human resources to construct the proper content for individuals. Although it involves the characteristics of internet, it is simply the transformation of the past computer-assisted software. According to the employees' individual learning, this study constructs employee learning model by simple computer-assisted system upon adaptive navigation support to enhance learning efficiency.

This study aims to construct corporate support system of computer-assisted instruction by the original information devices of the firms. This study demonstrates the result of workplace aptitude test by XML. Upon the classification by XML, this study

can rapidly search for the related cases and data to share knowledge and pass the experience. Thus, the employees will immediately control the learning points and enhance learning ability and learning quality.

The contribution of this study is constructing computer-assisted instruction system of small and medium enterprises by adaptive theory and low-cost tool to enhance learning effectiveness; in the future researches, the firms can apply more effective and precise aptitude test to obtain more reliable adaptive conditions as better learning conditions and paths of system decision and they can integrate the techniques related to new web language to show more complete situational e-learning system.

References

- [1] PC School, <http://www.pcschool.tv/news/news.aspx?id=33&t=4>
- [2] Bostrom, R.P., Olfman, L., Sein, M.K.: The importance of learning style in end-user training. *MIS Quarterly* 14(1), 101–119 (1990)
- [3] Gregorc, A.F.: Learning teaching styles: Potent forces behind them. *Educational Leadership* 36(4), 234–236 (1979)
- [4] Huang, Y.C.: Learning Style of Gifted and Normal Students in Junior High Schools and Comparative Study on Their Adaptability to Schools. *Bulletin of Special Education* 9, 249–276 (1993)
- [5] Su, H.M.: Aboriginal Students' Learning Style and Educational Application (2004), <http://mail.nhu.edu.tw/~society/e-j/40/40-11.htm> (retrived December 24, 2006)
- [6] Wu, B.H.: Theoretical Study on Learning Style. *Guidance of Elementary Education* 37(5), 47–53 (1998)
- [7] Keefe, J.W.: Profiling and Utilizing Learning Style. NASSP, Reston (1988)
- [8] Clarina, R.B., Smith, L.: Learning style shifts in computer-assisted instructional setting (retrieved November 12, 1998), <http://eric.ed.gov/ERICWebPortal.html>; Garner, I.: Problems and inconsistencies with Kolbs learning styles. *Educational Psychology* 20(3), 341–348 (2000)
- [9] Chen, M.C.: Research on Relationship among Adult Online Learners' Learning Style, Self-Monitoring and Learning Effect. Master's thesis, Graduate Institute of Adult Education, National Kaohsiung University of Education, Kaohsiung City (2005) (unpublished)
- [10] Chen, Y.C.: Study on Learning Style and Learning Performance of the Employees' e-learning: Using Institute for Information Industry as an Example, Master's thesis, Department of Information and Communications, Shih Hsin University (2007)
- [11] Lin, H.N.: Study on Adaptive Learning Websites: Using Senior High School Math as an Example, Master's thesis, Graduate Institute of Computer Science, National Chiao Tung University (2002)

Research on Recreational Sports Instruction Using an Expert System

Chih-Yao Lo^{1,2}, Hsin-I Chang³, and Yu-Teng Chang^{1,2}

¹ School of management, Huazhong University of Science & Technology

² Department of Information Management, Yu Da University, Miaoli County

³ General Education Committee, Yu Da University, Miaoli County,
Taiwan 361, R.O.C.

{jacklo, chsini, cyt}@ydu.edu.tw

Abstract. With the society fast change, people are facing threat of various chronic ailments due to lack of effective sport. How to exercise in more effective way become important to the sports instruction, it's also the key factor that foster people's recreational sports become daily routine. Although the fundamental knowledge of sports instruction can be obtained from courses or books, expert's instruction and experience is hardly to obtain especially in particular practice environment. Thus, the beginners tend to be lack of the expertise and information in order to exercise effectively. This study combines IT techniques and traditional sports instruction knowledge to construct an expert system of sports instruction in the field of recreational sports to solve the problems of traditional instruction (time-consuming and uneven quality of instruction). First of all, upon popular exercises, diverse training goals and physical fitness levels, this paper constructs the knowledge base with 324 flexible training courses and further obtains the figures of BMI, physical fitness and aptitude through user interface. By the establishment of fuzzy inference mechanism, this study intends to effectively enhance the preciseness and accomplish the objective of interactive training courses. This paper will provide the new thinking of recreational sports instruction and the approach not only keeps the experts' experience and knowledge, but also solves the difficulty of the people's exercise to finally fulfill the health lives.

Keywords: Recreational Sports Instruction, Expert System, Fuzzy Inference Mechanism.

1 Introduction

1.1 Research Background and Motives

Since there are more diseases such as obesity, high blood pressure and diabetes, the decline of the people's physical fitness becomes significant. Most of the people cannot attain the professional knowledge of recreational sports instruction. Thus, they are not able to select the exercises to improve their physical fitness and they cannot effectively plan the training courses. Besides, when doing the sports, they tend to be injured because of the lack of information.

There are varied recreational sports. Since the system of this study aims to enhance the people's physical fitness and healthy live and the targets are common people in the society, the researcher will eliminate the recreational sports which do not result in exercise effect and enhancement of physical fitness and those with high risk, require professional skills or which are less popular.

Besides, when selecting the factors, the study will encounter the limitation of the uncontrollable factors, such as the weather, living environment and economy. After precise assessment, this study only analyzes and infers the objective figures such as "physical fitness" "BMI" and "individual aptitude" to lead to the most proper training course planning.

2 Literature Review

2.1 Physical Fitness and Sports Training

Physical fitness was called "Leistungsfähigkeit" by the German, "physical aptitude" by the French and "physical strength" by the Japanese. In Taiwan, people tended to call it "physical fitness" [4]. It meant the physical ability to adapt to the environment. It was the base of physical activities and health.

The test of physical fitness included the following performance: 1) endurance; 2) strength; 3) muscular endurance; 4) flexibility. The formula of modern Sports Training was based on three factors.

(1) Intensity

It was measured by the most proper individual heartbeat: $220 - \text{individual age} \times 60\% \sim 80\%$, according to training objectives and physical fitness level.

(2) Duration

Duration referred to total time of aerobics with the most proper heartbeat.

(3) Frequency

With regard to physical fitness, except for the specific training objectives, the experts suggested the exercises for three to five times every week and 20-50 minutes every time upon the idea of physical and mental health and the most proper heartbeat.

2.2 Expert System

Expert System was also called knowledge system or knowledge-based system. It was one of the most popular topics in the field of artificial intelligence. At present, Expert System could assist the human beings with many complicated problems upon the knowledge and experience in the books and from the experts.

Expert System could construct expert knowledge and experience in the specific fields. After the users input the needs and conditions, it could analyze, infer and lead to the expertise conclusion or suggestion. In fact, it was the computer system with the inference ability to solve the problems by the knowledge and experience saved.

2.3 Fuzzy Inference

Fuzzy inference meant to deal with imprecise knowledge by precise approach. In reality, many thoughts could not be simply "supported" or "not supported". Fortunately, Fuzzy proposed by L.A. Zadeh [5] could solve the problem.

There were fuzzy aspects in daily lives and fuzzy inference included four major processes: “Fuzzification”, “Fuzzy Inference”, “Fuzzy rule base” and “Defuzzification”.

3 Research Design

3.1 Design of Expert System Framework

The development of the system included three stages-system analysis, design and practices. Thus, the system could be perfect and accomplish the goals. After analysis, the design and practice of this system would be fulfilled by the following:

(1) Knowledge system

This study first acquired knowledge in the books and expert experience and then transformed the knowledge into knowledge base of recreational sports by knowledge engineer and sub-system of knowledge acquisition.

The design of recreational sport knowledge base was the core of knowledge system. There were varied recreational sports, strength training theories and sports training and only the training courses upon academic base could be the most effective.

(2) User interface

User interface was the most approachable part in the system. The users would be impressed on the clear design of the forms and logic operation. Thus, the design should be cautious.

(a) “Sub-system of user consultation”: after entering the system, the users filled in the forms according to the instruction in order to analyze and compare the figures. The interface of the forms would be based on clear design of Microsoft Visual Basic 6.0.

(b) “Sub-system of explanation”: it acquired the proper recreational sports and instruction for the users upon the rules of system inference and the connection with recreational sport knowledge base. The explanatory ability of the system should be enhanced to result in more precise inference. Thus, the users would be more willing to use this system.

(c) “Sub-system of knowledge acquisition” could rapidly and structurally construct the knowledge base for future modification. The sub-system was upon Microsoft Visual Basic 6.0 and Microsoft Office Access 2003.

(3) Fuzzy inference mechanism:

The innovation of this study on recreational sport instruction was the introduction of Fuzzy which could further enhance knowledge expression and inference of the system. When expert system of recreational sport instruction dealt with the data of physical fitness, for the preciseness of the output result, “yes” and “no” inference decision would not be proper; instead, the interaction among the factors should be involved.

Thus, only fuzzy inference which could transform semantic or semi-structure expert knowledge into the rules of fuzzy control for further fuzzy inference would thoroughly solve the problems of training objectives and course planning.

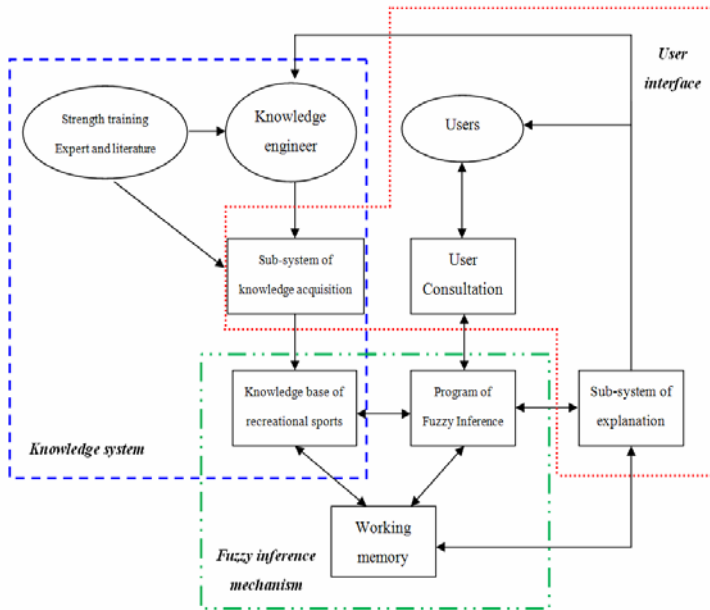


Fig. 1. Framework of Expert System of recreational sports instruction

3.1.1 The Design of Knowledge Base of Recreational Sports

The construction of knowledge base meant the knowledge engineer organized the expert knowledge and experience by proper knowledge expression skills. Continuous knowledge accumulation and reorganization could increase knowledge base and its value of reference, diagnosis, evaluation and suitability. Besides, complete knowledge framework planning was closely related to the preciseness of overall Expert System.

(1) Selection of recreational sports

Recreational sports in knowledge base should be common exercises which everyone could easily participate in. Thus, this knowledge base would be simple, useful and highly effective. Recreational sports in this study included three categories: 1) ball sports (6 items); 2) non-ball sports (7 items); 3) sports with specific strength training (7 items). There were totally 20 sports items.

(2) Planning of sports training courses according to training objectives and physical fitness level

Based on three principle categories of physical fitness (endurance, muscular endurance and strength) which were the most important and meaningful, this study divided the above 20 exercises into 7 training objectives according to the results by expert training experience: (1) exercises to enhance endurance (11 items); (2) exercises to enhance strength fitness (7 items); (3) exercises to enhance muscular endurance (10 items); (4) exercises to enhance endurance and strength fitness (5 items); (5) exercises to enhance endurance and muscular endurance (7 items); (6) exercises to enhance strength and muscular endurance (6 items); (7) exercises to enhance endurance, strength and muscular endurance (8 items). There were totally 54 items.

Training courses of 54 items were further designed according to the levels of the users’ physical fitness. For instance, with regard to the bicycle, upon different training approaches, there were 5 training objectives. With 3 levels of physical fitness, there would be 15 different training courses. With flexibility training courses, this knowledge base included 324 kinds of training course planning for the participants with different physical fitness.

Table 1. Recreational sports of 7 strength training objectives

endurance training	Strength training	Muscular endurance training	endurance and strength training	endurance and muscular endurance training	Strength and muscular endurance training	endurance, strength and muscular endurance training
1. stair climbing	1.push-up	1. leg exercise	1. dodge ball	1. jogging	1. 60-m running	1. bicycle
2. running and walking	2. bottom exercise	2. back exercise	2. tennis	2. hiking	2. swimming (short distance)	2. badminton
3. badminton	3. jumping	3. waist exercise	3. badminton	3. swimming (long distance)	3. table tennis	3. tennis
4.basketball	4. volleyball	4.bicycle	4.volleyball	4. rope skipping	4. bicycle	4. basketball
5. bicycle	5. badminton	5. badminton	5.dancing	5. badminton	5.badminton	5.volleyball
6. swimming(long distance)	6. swimming(short distance)	6. basketball		6. bicycle	6.volleyball	6.swimming
7.jogging	7. dodge ball	7.swimming		7.basketball		7.dancing
8. hiking		8.jogging				8.rope skipping
9.dancing		9.hiking				
10. rope skipping		10. rope skipping				
11. dodge ball						

(3) Including the characteristic analysis of recreational sports in knowledge base: This study included the remarks such as training purpose, operational principle, attention, age and principle of physical fitness in knowledge base so that the users could recognize the essence of the exercises that would be more interested and prevent the injury due to the unfamiliarity with the operation. Characteristic analysis of recreational sports resulted in useful knowledge and thus the participants would not be injured or have long-term injury. They would cultivate the good habit of life-long exercise and it was one of the concepts of this expert system.

3.1.2 Design of User Consultation and System Inference Decision

Simple input forms in sub-system of user consultation aimed to result in three categories of analysis and inference figures.

(1) BMI

The users were asked to input the data such as gender, age, body, weight for the calculation of weight-height index suitable for the teenagers and “standard weight ” suitable for the adults in order to classify BMI.

- (a) For the “obese” users according to the analytical result of individual BMI, this study suggested the exercises of endurance training with significant effect.
- (b) For the “thin and weak” users according to (b) the analytical result of individual BMI, this study suggested the exercises of strength training to enhance the muscle.

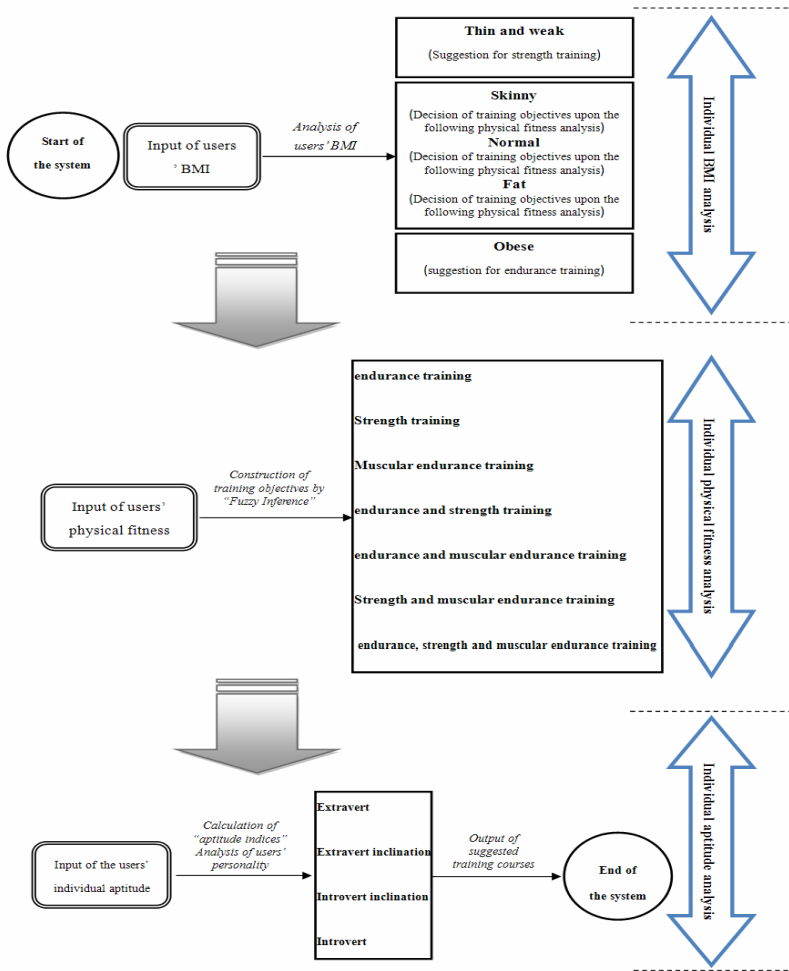


Fig. 2. Decision process of the system

(c) The users in the categories of “skinny” “normal” and “fat”, upon the following “individual physical fitness”, this study would continue the analysis and training suggestion by “fuzzy inference”.

(2) Individual physical fitness

It aimed to obtain the figures for physical fitness analysis in the system. In this study, there were four categories of age: “7~8 years old”, “9~12 years old”, “13~23 years old” and “24~65 years old”. Different measurements and figures input were required for different categories.

This critical step of sports instruction aimed to provide more precise inference by effective and flexible approach. Thus, only by “fuzzy inference”, the knowledge of Sports Training knowledge would lead to the training objectives needed by the users

the most. This study would analyze the levels of users’ physical fitness and finally save the inference results in working memory. The design of fuzzy inference will be described in section 3.1.3.

(3) Individual aptitude

Aptitude would significantly influence the selection of the types of recreational sports. If the system could analyze the users’ aptitude and recommend the exercises according to their aptitude, the participation would be more interesting and the users’ motives would be enhanced.

This study designed the checking and filling-in forms to obtain the figures: “checking for individual personality traits”, “checking for personal favorite exercises” and “filling in other favorite exercises”. Individual aptitude indices were “extravert”, “extravert inclination”, “introvert inclination” and “introvert”. According to these four categories of personality traits and fuzzy inference results in working memory, Expert System would recommend the suitable exercise and training courses for the users.

3.1.3 Design of Fuzzy Inference

This paper initially analyzed users’ physical fitness by fuzzy inference. First of all, this study constructed the rules of Fuzzy Control according to Sports Training knowledge in the books and of the experts and output the suitable exercises and training courses upon fuzzy inference engine. The steps of process were below:

(1) Definition of variables of fuzzy inference

According to the training objectives, the exercises in system knowledge base were divided into 7 categories, as shown in Table 2:

Table 2. Codes of strength training objectives

No.	Code	Items to enhance physical fitness
1	EE	endurance
2	SS	Strength
3	MM	Muscular endurance
4	ES	endurance +strength
5	EM	endurance +muscular endurance
6	SM	Strength +muscular endurance
7	ESM	endurance +strength +muscular endurance

(2) Definition of “fuzzy partition” of fuzzy inference

Three input variables (S, M, E) represented the data of users’ strength, muscular endurance and endurance. Norm of physical fitness was conversed according to the percentage. Thus, three variables were transformed in the range (-10~10).

Fuzzy partition of three variables was based on Fuzzy Sets (Low, Middle and High). Low meant the users’ physical fitness was low; Middle meant medium level and High meant high level. Fuzzy partition was shown in Figure 3.

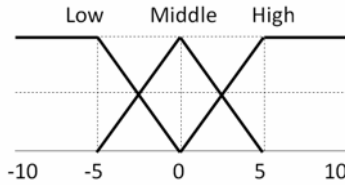


Fig. 3. Fuzzy partition

(3) Construction of “Fuzzy Rule Base” of fuzzy inference

The construction of Fuzzy Control rules was based on the literatures and experts’ knowledge and experience, as shown below:

R_i: If S is A_i¹ and M is A_i² and E is A_i³ then out is B_i

i: number of control rule

S, M, E: Input variables of thesis

A_i: Fuzzy Sets

B_i: Output variable of inference

This study designed fuzzy system with three input variables (S and M and E) and one output (EE or SS or MM or ES or EM or SM or ESM). The control rules are shown as following examples:

R1: if S is Low and M is Low and E is Low then output is ESM

R2: if S is Low and M is Low and E is Middle then output is SM

R3: if S is Low and M is Low and E is High then output is SM

R4: if S is Low and M is Middle and E is Low then output is ES

R5: if S is Low and M is Middle and E is Middle then output is SS

(4) Norm percentage conversion of physical fitness

Using a 12-year-old male student as an example, the student input the figures of “individual physical fitness”: standing long jump (S)149cm, sit-ups (M) 32 times/60 seconds, 800-m running and walking (E)228seconds. After conversing the students’ norm by the following formula, S=-3, M=-2, E=4.

$$\text{Norm percentage conversion: } \frac{Y - m}{X - (-10)} = \frac{M - Y}{10 - X} \equiv X = \frac{10(2Y - m - M)}{M - m}$$

Y: Physical fitness figures of input of “individual physical fitness analysis”.

m: Minimum figure of physical fitness norm of the age range.

M: Maximum figure of physical fitness norm of the age range.

X: Standard figure (-10~10) of physical fitness after the conversion.

(5) Calculation of the corresponding membership of triangular membership function

This study calculated corresponding membership by triangular membership function of Fuzzy. The formula was shown below:



Fig. 4. triangular membership functions

In figure 4, when e is at the right of triangular membership function, the corresponding membership is: $\frac{x_2 - e}{x_2 - x_1}$

when e is at the left of triangular membership function, the corresponding membership is: $\frac{e - x_3}{x_1 - x_3}$

Range of membership is 0~1. Larger membership indicates high level of correlation.

(a) S= -3:

Membership of “Low fuzzy partition” of Low physical fitness:
 $Low(-3)=(0-(-3))/(0-(-5))=3/5=0.6$

Membership of “Middle fuzzy partition” of Middle physical fitness:
 $Middle(-3)=(-3-(-5))/(0-(-5))=2/5=0.4$

(b) M= -2:

Membership of “Low fuzzy partition” of Low physical fitness:
 $Low(-2)=(0-(-2))/(0-(-5))=2/5=0.4$

Membership of “Middle fuzzy partition” of Middle physical fitness:
 $Middle(-2)=(-2-(-5))/(0-(-5))=3/5=0.6$

(c) E= 4:

Membership of “Middle fuzzy partition” of Middle physical fitness:
 $Middle(4)=(5-4)/(5-0)=1/5=0.2$

Membership of “High fuzzy partition” of High physical fitness:
 $High(4)=(4-0)/(5-0)=4/5=0.8$

(6) Fuzzy inference

Only the following control rules (R2, R3, R5, R6, R11, R12, R14 and R15) would influence system control. This study calculated the correlation of these control rules by “max-min operation”.

(a)Correlation with Rule 2: $W2=\min\{ Low(-3), Low(-2), Middle(4)\}=0.2$

(b) Correlation with Rule 3: $W3=\min\{ Low(-3), Low(-2), High(4)\}=0.4$

(c) Correlation with Rule 5: $W5=\min\{ Low(-3), Middle(-2), Middle(4)\}=0.2$

(d) Correlation with Rule 6: $W6=\min\{ Low(-3), Middle(-2), High(4)\}=0.6$

(e) Correlation with Rule 11: $W11=\min\{ Middle(-3), Low(-2), Middle(4)\}=0.2$

(f) Correlation with Rule 12: $W12=\min\{ Middle(-2), Low(-2), High(4)\}=0.4$

(g) Correlation with Rule 14: $W14=\min\{ Middle(-3), Middle(-2), Middle(4)\}=0.2$

(h) Correlation with Rule 15: $W15=\min\{ Middle(-3), Middle(-2), High(4)\}=0.4$

(7) “Defuzzification” resulting precise training objectives

Precise output of fuzzy inference could be obtained by “Defuzzification”. The example of the calculation was shown below:

Rule 2: $W2=0.2$, Out=SM, 0.2 SM

Rule 3: $W3=0.4$, Out=SM, 0.4 SM

Rule 5: $W5=0.2$, Out=SS, 0.2 SS

Rule 6: $W6=0.6$, Out=SS, 0.6 SS

Rule 11: $W11=0.2$, Out=MM, 0.2 MM

Rule 12: $W12=0.4$, Out=MM, 0.4 MM

Rule 14: $W14=0.2$, Out=ESM, 0.2 ESM

Rule 15: $W15=0.4$, Out=SM, 0.4 SM

“Defuzzification” result demonstrated that SM weight was 1, SS weight was 0.8, MM weight was 0.6, ESM weight was 0.2. Fuzzy inference engine treated SM with the highest output weight (“strength and muscular endurance”) as the exercise of training objectives.

(8) Analysis of users’ physical fitness:

Fuzzy inference engine defined training objectives upon SM and it must further define the users’ physical fitness level of SM in training course planning to suggest the proper training courses.

- (a) After “norm percentage conversion”, strength (S)= -3, muscular endurance (M)= -2, mean = -2.5.
- (b) Based on fuzzy partition in Figure 3, definition range of physical fitness level was in Table 3.
- (c) According to the table, physical fitness level was defined as “Middle physical fitness”.

Table 3. Definition range of physical fitness levels

Physical fitness levels	Range of mean
High physical fitness	mean >2.5
Middle physical fitness	-2.5<=mean <=2.5
Low physical fitness	mean <-2.5

After “fuzzy inference”, Expert System of recreational sports instruction saved “training objectives upon strength and muscular endurance” and “Middle physical fitness” in working memory for the following inference.

3.2 System Modification and Function Extension

(1) Design of system feedback

In order to allow this system to more precisely meet the reality and match the change of times, this study would design feedback sheet for the users and further save their satisfaction and suggestion as the criteria of inference modification, system renewal and modification of knowledge base.

(2) Design of checking function of recreational sports

The function was the practical extension of Expert System of recreational sports instruction in this study. Some users had their favorite or long-term regular exercises; however, they did not know how to plan the training courses. Thus, with this checking function, the system would plan the training courses suitable for individual physical fitness.

4 Experiment Design

This study proposed the suggestions according to the students’ individual conditions by initial model of Expert System of recreational sports instruction in order to improve the

students' physical fitness and function as the criterion for system evaluation and modification.

(1) Research targets:

Using the primary school education in Taiwan as an example, this study first selected one class for the experiment and treated the students with inferior physical fitness and needed physical fitness planning as the targets.

(2) Training courses by Expert System of recreational sports instruction:

In the constructed initial system, this study obtained the individual training course planning according to the individual students' figures input.

(3) Physical fitness training according to the courses suggested by Expert System:

The students were allowed to select the favorite exercises from several suggested courses. Thus, the study only had to arrange the training time with instruction according to the courses output from the system and further conducted the posttest after the end of the training cycle.

The approach could provide more effective consultation and assistance for strength training personnel, reduce the dependency and needs for the experts and rapidly control the direction key points of sports instruction.

5 Validation and Results

5.1 Improvement of the Physical Fitness

After the cycle of the targets' physical fitness training, this study compares the results before and after the training in posttest of physical fitness with regard to "BMI", "flexibility" and "physical fitness" and investigates the targets' satisfaction, as shown below:

(1) BMI improvement

The targets upon the losing-weight courses suggested by this system obtain "5.30% improvement" in average. Comparing with the targets without treating losing weight as the training objective (average improvement -2.87%), the former reveals the effect of improvement.

(2) Improvement of flexibility

The targets upon the flexibility training suggested by this system obtain "15.74% improvement" in average. Comparing with the targets without treating flexibility as training objective (average improvement 2.34%), the former reveals the effect of improvement..

(3) Improvement of three categories of physical fitness

The targets have training courses according to the inferred training objectives. The result of posttest demonstrates that the targets upon strength training objectives suggested by this system obtain "12.87% improvement" in average. Comparing with those without physical fitness as training objectives (5.29%improvement), the former reveals the effect of improvement.

(4) The targets' satisfaction survey

After the training cycle and posttest of the targets' physical fitness, this study conducts questionnaire survey on the targets' satisfaction. Total average satisfaction is 77% and it demonstrates that users are willing to participate in the exercises by the suggestion of the system and they are satisfied with the improved physical fitness.

6 Conclusions

In order to achieve the effect of recreational sport, people required expertise for choosing the sport and instruction. However, most of people spend time for sport without expertise. Thus, this study firstly introduced the approach of artificial intelligence and developed the expert knowledge-base that construct system framework. The study then analyzed the users' varied complicated factors by "fuzzy inference" to further suggest the plans of recreational sports instruction in order to improve physical fitness. It is not only solved the difficulty of selection of appropriate recreational sports, but also instruct the learning direction while people lack of knowledge to the particular sport. By proper course design, people could efficiently obtain the sport instruction for improving their regular exercise habit.

With regard to Sports Training and instruction, this study intends to propose a model and approaches to further facilitate the people's exercise habits. Thus, everyone can have proper recreational sports at any time and any place upon objectives and plans to further develop regular exercise habit and enhance physical fitness, prevent the diseases and ultimately enjoy healthy lives.

In the future, the study would further emphasized the consolidating expert knowledge-base for the system overall preciseness. It will also plan customized interface and function with regard to different groups and benefit more exercisers or Sports Training researchers by online system.

References

- [1] Wang, J.D., Hsiao, D.C.: Neural Network and Introduction of Fuzzy Control. Chuan Hwa Books (1994)
- [2] Kao, M.D., Huang, H.Y., Tzeng, M.S., Li, N.N., Hsieh, M.C.: Survey on the People's Nutrition in Taiwan from 1986 to 1988 –Measurement of Physical Conditions(I) Height and Weight. *Journal of Nutrition* 16, 63–84 (1991)
- [3] Chen, Y.C., Ouyang, C.R.: Decision Support and Expert System. Chuan Hwa Publishing (1991)
- [4] Chen, J.C.: Physical Fitness and Disease Prevention. Teacher Physical Fitness Instruction, Ministry of Education (2000)
- [5] Su, M.C., Chang, H.D.: Machine Learning: Neural Network, Fuzzy System and Genetic Algorithm. Chuan Hwa Books (2004)
- [6] Aerobics and Fitness Association of America, Personal Fitness Training: Theory&Practice (2007), <http://afaa.com>
- [7] Stockdale, A., Wood, M.: Building a Small Expert Systems for a Routine Task: A Case Study. *Management Decision* 16(3) (1992)

- [8] Liebowitz, J.: Knowledge-based/Expert Systems technology in life support systems Kybernetes. *The International Journal of Systems & Cybernetics* 26(5), 555–573 (1997)
- [9] Bompa, T.O.: *Periodization: Theory and Methodology of Training*, 4th edn. Human Kinetics Publishers Inc. (1999)
- [10] Jayaraman, V., Srivastava, R.: Expert Systems in production and operations management: Current applications and future prospects. *International Journal of Operations & Production Management* 16(12), 27–44 (1996)
- [11] Hoeger, W.W.K., Hoeger, S.A.: *Principles and Labs for Fitness and Wellness*, Wadsworth a division of Thomson Learning, Inc. (2007)

Using 6LowPAN UPnP and OSGi to Implement Adaptable Home Ambient Intelligence Network Platform

Zhang Hui-bing^{1,2} and Zhang Jing-wei²

¹ Embedded Software and Systems Institute
Beijing University of Technology, Beijing, China

zhanghuibing@emails.bjut.edu.cn

² Guilin University of Electronic Technology, Guangxi, China

Abstract. By means of exposing the intrinsic relationship between home Ambient Intelligence(AmI)application system and its base network infrastructure, the paper puts forward a solution of adaptive home AmI network infrastructure by using 6LowPANUPnP and OSGi technologies. The novel model supports heterogeneous devices dynamic joining, Multiple Control Center automatic switching, uniform service & controlling and has good mobility. According to the solution, the methods for implementing the architecture and its sub-systems, as well as the technology theory for it are presented. At last, a prototype system based on Jn5139 6LowPAN is designed.

1 Introduction

Today, information devices are rapid developing along the trend of miniature, wireless, embedded and mobile, and also speedup entering everywhere of the physical spaces. People will live in intelligence information spaces which are made up of heterogeneous electronic devices at anytime in anywhere. While enjoying all kinds of services, end-users, especially the home users, are facing great challenges: they have little or no special knowledge about the computers and networks, but have to choose, deploy, integrate and operate most of the products and technologies in their home[1]. This reveals the shortcoming of traditional computing model: computer-centered and people must adapt to the system, which keeps many information devices out of home.

To solve the problem, many researchers are interested in the next generation computing model: Ambient Intelligence. It is a vision on the future of consumer electronics, telecommunications and computing that was originally developed in the late 1990s for the time frame 2010-2020[2]. The most important thing in AmI systems is to “make the system adapt to people”, as opposed to “people adapting to the system” [3], so the system is human-centric. AmI is becoming a hot research field[3,4,5,6,7], but designing and realizing an adaptable home AmI network platform is of great difficulty because of the following challenges:

- 1) **Self-organization and extensibility:** In home, all kinds of devices are huge differences. Heterogeneous devices' self-organization and extensibility under "zero-configuration" are important for AmI applications.
- 2) **Depth embedded and mobility:** Telemedicine is becoming more and more popular [8]. To meet the requirement, wearable medicine devices must be implemented in the form of depth embedded, networked and mobile connected.
- 3) **Dynamic discover and obtain services:** To the layman, especially elder, it is important that the information devices can automatically discover and send services to the user by a simple and uniform mode.
- 4) **Remote management and distribute services:** If device manufacturers, telecom operators and service providers can manage and distribute services through the Internet, its corresponding costs will be huge substantially reduced and the end-users will be greatly facilitated.

To settle the above challenges and provide an advanced infrastructure for home AmI application system, the paper designs an adaptable network platform based on IPv6 [9,10], IPv6 over Low power Wireless PAN(6LowPAN) [11], Universal Plug and Play(UPnP) [12] and Open Service Gateway initiative (OSGi) [13], etc.

The remainder of this paper is organized as follows: Section 2 introduces the related technologies and works; Then we present an adaptable home network infrastructure and detail the realization methods, technology theory, application scenario in the section 3; Section 4 designs a simple prototype system based on Jn5139 6LowPAN development kits; At last, conclusions are elaborated in section 5.

2 Related Technologies and Works

2.1 UPnP and OSGi

UPnP is a protocol architecture which is for pervasive peer-to-peer network connectivity of networked appliances, audio and video equipment, sensors/actors and PCs of all shapes and sizes whether they use wire or wireless transmission [14]. In order to support "zero-configuration", invisible networking and automatic discovery for heterogeneous devices, UPnP provides a flexible, standardized and easy-to-use scheme for networked devices. This means device can dynamically join the network, automatically obtain IP address, advertise its services, discover other devices' services and leave the network "elegantly", all of which are automatic [12]. Microsoft claims that UPnP will be extended to every appliance in the home, and will become a necessary architecture, protocol and interface for PC applications and intelligent appliances' interoperability.

OSGi technology is universal middleware. It provides a service-oriented, component based environment for developers and offers standardized ways to manage the software (known as bundles) lifecycle. OSGi aims at providing a unification, open and administrable environment for services providers, developers, network managers and devices manufacturers, so make them can develop, deploy and manage services in the network in the light of universal model. OSGi-compliant devices can

download, install, update and remove the bundles without influencing the current running bundle in a dynamic and scalable fashion [13]. So it can not only provide personalization services but implement the remote management.

In addition, OSGi implements devices and services discovery methods across different technologies, so other industry-standard discovery technologies (e.g., UPnP and Jini), can be integrated into OSGi framework by using OSGi's Import/Export mechanism. The services and devices which register in the OSGi framework can be exported to a local network, so the devices and services in the local network can use them by its local discovery methods; Similarly, the devices and services which are discovered by local technology also can be imported into the OSGi framework, so they can be represented as OSGi's valid entities [15].

From the above discussions we can know that UPnP only takes charge of devices and services' declaration and makes the discovery dynamically, while the OSGi provides a platform for bundles (services) to execute by a well defined and managed environment. OSGi can use the UPnP to find the networked devices and services. So OSGi and UPnP can cooperate tightly to realize the devices' automatic discover, communicate with each other and the bundles' dynamic deploy, execute.

2.2 IPv6 and 6LowPAN

As a key for next generation Internet, IPv6 has many advantages: gigantic IP address space, better mobility (include NEMO), network self-organization, QoS and security, etc. Especially due to its hardly unlimited address space, mobility and self-organization characters can reach all the requirements of adaptable home network.

6LowPAN protocol defines how to enable IPv6 packets to be carried on IEEE 802.15.4 wireless networks. In order to achieve wireless sensor network (WSN) based on IP protocol, 6LowPAN takes up with mutual-operation between IPv6 and 802.15.4 by using IEEE802.15.4 for its MAC and PHY layer and IPv6 protocol stacks for the layers above the MAC. 6LowPAN can help to improve WSN's nodes mobility and enable the nodes can be accessed directly from the Internet and vice versa. At the same time, 802.15.4 has obvious advantages than current technologies for home control. So introducing 6LowPAN is a good solution when we design the adaptable home network platform.

2.3 Related Works

In recent years, many intelligence environment architectures have been proposed by some well-known research institutions: [8] used OSGi, UPnP and Zigbee to implement a wireless ubiquitous home healthcare environment; [16] put forward a 3-layer smart-home architecture which is constituted by gateway operator, regional management center and home gateway. In their work, OSGi, UPnP and agent were used to achieve automation of services discovery, registry and management in home gateway; [17] implemented service discovery and management, dynamically deployed software and communicated with heterogeneous networks

based on OSGi and mobile-agent; [18] adopted OSGi and OWL-S to achieve service discovery and transfer among multi-OSGi framework; [19] realized a service-oriented, peer-to-peer smart-home platform by means of OSGi and multi-agent. From the aforementioned researches we can find: (1) OSGi is the mainstream technology; (2) Most of the researches focus on the smart-home from the point of home gateway; (3) Some projects are complicated and low efficiency; (4) Intelligence nodes lack mobility; (5) Adopting the single point control model based on the home gateway will cause single point failure; (6) All of the work don't mention the uniform control and service.

“Using 6LowPAN UPnP and OSGi to Implement Adaptable Home AmI network Platform” has some similar with these aforementioned works, what the main differences are follows: (1) Implementing a Multiple Control Center (MCC) which can dynamically switch based on OSGi and UPnP; (2) Introducing the IPv6 and 6LowPAN into the home network to realize the home uniform control and mobile telemedicine; (3) providing mobility and direct access ability by using IPv6.

3 AmI-Oriented Adaptable Home Network Infrastructure

According to the requirement of home AmI application system, we put forward an adaptable home network infrastructure by using 6LowPAN, UPnP and OSGi technologies. As seen in Fig.1, it makes up of 4 subsystems: MCC, mobile wearable telemedicine & sensor, home uniform control, home office & safety. This platform can support the AmI applications successfully.

3.1 Multiple Control Center

The architecture of traditional home network is single control center by only using one home gateway, all of the operations must be transferred by the gateway

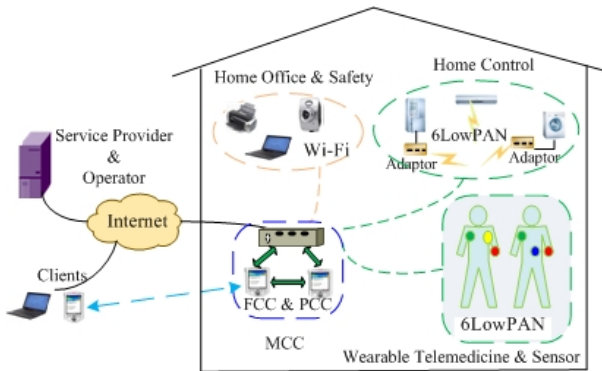


Fig. 1. AmI-oriented adaptable network platform

[8,16,20], which not only causes the single point failure [19] and communication delay, but also augments the costs and communication layers. Therefore, we propose a MCC architecture based on multi-OSGi, the MCC has two components: Fixed Control Center (FCC) and Portable Control Center (PCC), as shown in Fig. 1. FCC is similar to traditional gateway, which is responsible for communication between home network and Internet, remote management and dynamic deploy bundles. At the same time, networked devices in home can be directly accessed from Internet without parsing home gateway because they are based on IPv6 and UPnP protocols. PCC acts as both a client and a control center. On the one hand, it can access the home networked device through the FCC when it is out of home; on the other hand, it can also act as a local control center and directly access the device when it is in home. FCC and PCC have the similar 5-layer model, as described in Fig. 2. The two lower layers are responsible for network connecting and communication, while the above three layers are responsible for device and service's description, discovery and management.

Due to the OSGi framework, FCC and PCC can download, install and execute bundles automatically. So they could not only implement remote control and management, but also enable the services transfer among FCC and PCCs dynamically: when a PCC enters into the network environment, the FCC can sense it and download some services to the PCC from its framework, then the PCC can use the home network's devices and services directly.

In order to cooperate with local UPnP network's device, UPnP service bundle has been installed in the FCC and PCC's OSGi framework, as seen in Fig. 2. This bundle can map UPnP network's devices and services into the FCC or PCC's OSGi platform by using UPnP Import-Service, so they can appear as OSGi native entities; Similarly, FCC or PCC can export their UPnP-compliant entities to the UPnP network and enable them become virtual UPnP devices.

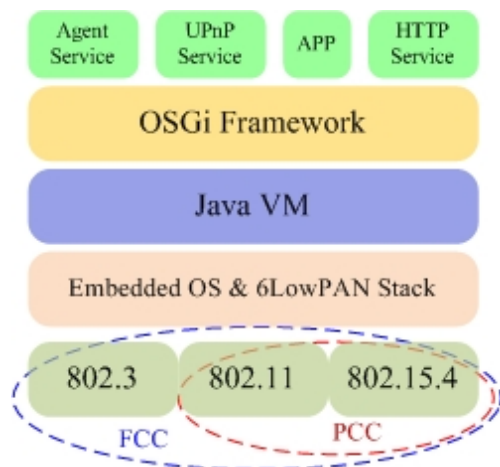


Fig. 2. FCC and PCC components model

3.2 Wearable Telemedicine and Sensors

With the growth of telemedicine technology, wearable telemedicine with network capability will be the essential ingredients of the home network. But how to automatic connect, discover these devices will be an imperative and difficult matter for researchers. [8,21,22] put forward three solutions to connect telemedicine from different points, however there are some deficiencies: complex structure, lack mobility, incapable of direct access node, etc. In order to achieve capability of IP based mobility and compact architecture for AmI-oriented electronic medical device, we design a new solution based on 6LowPAN and UPnP, which is a 7-layer architecture, as seen in Fig. 3.

In the model, MCU and RFIC layer provides the basic function for computing and wireless communication. PHY and MAC layers support signal transmission with low rate, low power and short distance based on IEEE 802.15.4 protocol. The 6LowPAN adaptor layer bridges IPv6 layer and MAC layer by its packet header compress, packet fragmentation and combination function, etc. The UDP and IPv6 layer is a tailored protocol stack which only the essential parts are reserved. The HTTPU/UPnP layer is also a tailored protocol which makes the wearable device only appear as a controlled UPnP device. The APS is a user application layer.

By using the 7-layer model, we can implement a 6LowPAN and UPnP based telemedicine device which cooperates with other devices on the UPnP network. Other sensors can also use this 7-layer model.

3.3 Home Control

To implement automatic connecting and dynamic service discovery, home electronic appliances are connected to the network directly or through adapters.

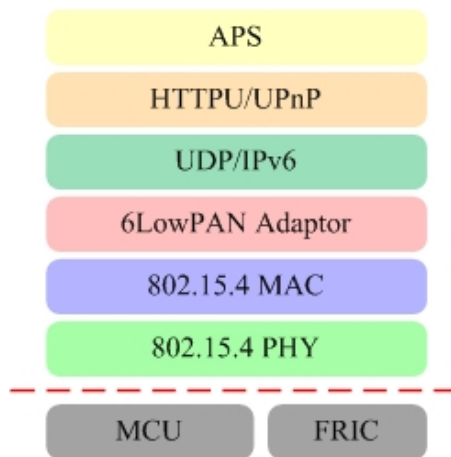


Fig. 3. Wireless telemedicine device architecture

They adopt the architecture same as the telemedicine device's, so they can be controlled in a uniform mode by using the 802.15.4 radio frequency technology. FCC or PCC can directly control them by using the global unique IPv6 address.

3.4 Home Office and Entertainment

Different from the medical device and home control, home office and entertainment have enough computing ability, memory capacity and high communication rate, etc. IEEE 802.11, IPv6 and UPnP are often the default configuration. So they can be used in the adaptable network without additional technology.

3.5 Application Scenario

Fig. 4 details a typical application scenario, where a remote terminal accesses the UPnP service through OSGi platform. It bands together related subsystems seamless.

When the wearable medical device (blood pressure meter) with UPnP function appears in UPnP home network, it first obtains an IPv6 address by address automatic configure protocol, then advertises its service(Blood-Pressure Service, BPS) over the network, the UPnP control point (CP) can deal with the BPS. In this scenario, the FCC acts as the UPnP CP. OSGi based FCC can detect, analyze and register the BPS because the UPnP base driver bundle has been installed in OSGi framework. After that, the BPS is turned into the OSGi native

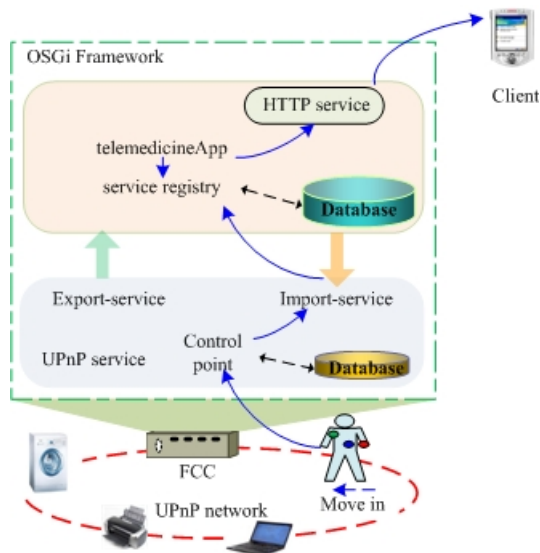


Fig. 4. Intelligence terminal accesses UPnP device

service by using the UPnP driver's Import-service; next, the OSGi's service registry can register it in the service database. Now, the UPnP BPS appears as a OSGi native service and the other OSGi services, for example telemedicineApp, can discover, invoke it by usual means. When the remote intelligence terminal wants to inspect the blood pressure, it sends a request to the HTTP Service, then the HTTP Service can respond to the request through sending current blood pressure value to intelligence terminal by cooperating with related service.

4 A Simple Prototype System

According to the proposed solution for the adaptable network architecture, we have designed a prototype system by using the Jn5139 6LowPAN development kits [23][24], as shown in Fig. 5. The wearable blood pressure meter is made up of wireless node and blood pressure meter. The measure values of blood pressure are sent to wireless node through serial port; FCC is made up of wireless access point (WAP) and PC, the WAP and PC are connected by Ethernet network. In addition, the OSGi framework is deployed on the PC and the tailored UPnP/HTTPU protocols is deployed on wireless node (WN). The remote client can access FCC through Internet. Between the WN and WAP is 6LowPAN network.

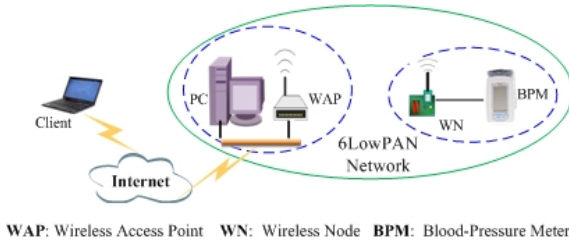


Fig. 5. Jn5139 6LowPAN based prototype system

5 Conclusion

In this paper, we have analyzed some disadvantages of the traditional smart-home network infrastructure and then have proposed an AmI-oriented adaptable network platform by using the 6LowPAN, UPnP and multi-OSGi. This model not only supports the common functions but also provides novel AmI applications: MCC's dynamic switching, unify controlling by 802.15.4 technology and mobility access of wireless wearable nodes, etc.

But this adaptable infrastructure is the first step, in the future researches, we will focus on the protocol's tailor and its optimization, for example the IPv6 protocol; we will also design and implement some appropriate and integrated devices, such as 6LowPAN based wearable blood pressure meter, OSGi and 6LowPAN based FCC.

Acknowledgement

The authors would like to thank the National Nature Science Foundation of China (No.90407017, 60773186) and Nature Science Foundation of Beijing (No.4072008) for their support to the research.

References

1. Reynolds, F.: The Ubiquitous Web, UPnP and Smart Homes (March 2007), <http://www.w3.org/2006/02/slides/reynolds.pdf>
2. <http://www.ami-lab.org>
3. Rui, C., Hou, Y.-b., Huang, Z.-Q., Jian., H.: Modeling the Ambient Intelligence Application System: Concept, Software, Data, and Network. *IEEE Transactions on Systems, Man, and Cybernetics, Part C. Applications and Reviews* 39(3), 299–314 (2009)
4. Bravo, J., Alaman, X., Riesgo, T.: Ubiquitous computing and ambient intelligence: New challenges for computing. *Journal Of Universal Computer Science* 12(3), 233–235 (2006)
5. ISTAG. Scenarios for Ambient Intelligence in 2010, <http://www.cordis.lu/istag.htm>
6. Weber, W., et al.: Ambient intelligence-key technologies in the information age. In: *IEEE International conference on Electron Devices, Hilton Washington and Towers, Washington DC, USA, December 8-10*, pp. 111–118 (2003)
7. Riva, G., Vatalaro, F., Davide, F., Alcaniz, M.: *Ambient Intelligence: The Evolution Of Technology*. In: *Communication And Cognition Towards The Future Of Human-Computer Interaction*. OCSL Press (2005)
8. Lin, W.-W., Sheng, Y.-H.: Using OSGi UPnP and Zigbee to provide a wireless ubiquitous home healthcare environment. In: *Mobile Ubiquitous Computing, Systems, Services and Technologies, UBICOMM 2008, September 29-October 4*, pp. 268–273 (2008)
9. Johnson, D., Perkins, C., Arkko, J.: *Mobility Support in IPv6*. IETF RFC 3775 (June 2004)
10. Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: *Network Mobility (NEMO) Basic Support Protocol*. IETF RFC 3963 (January 2005)
11. 6LowPAN, <http://www.ietf.org/html.charters/6lowpan-charter.html>
12. UPnP, <http://www.upnp.org>
13. OSGi, <http://www.osgi.org>
14. von Pattay, W.P., Heusinger, S.: Peer-to-peer connectivity made easy (April 2009), http://www.iso.org/iso/p._8_main_focus.pdf
15. Dobrev, P., Famolari, D., et al.: Device and Service Discovery in Home Networks with OSGi. *IEEE Communications Magazine*, 86–92 (August 2002)
16. Zhang, H., Wang, F.-Y., Ai, Y.: An OSGi and agent based control system architecture for smart home. In: *2005 IEEE Proceedings of the Networking, Sensing and Control, Tucson, Arizona, USA, March 19-22*, pp. 13–18. PARCS Research Center, University of Arizona (2005)
17. Ancuti, C., Fumerala, M., Luyten, K., Coninx, K.: General Adaptable Services Manager for Pervasive Computing Environments. In: *Proceedings of the ITI 2007, 29th Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, June 25-28*, pp. 209–214 (2007)

18. Lee, S., Lee, J.: Design of service management system in oSGi based ubiquitous computing environment. In: Etzion, O., Kuflik, T., Motro, A. (eds.) NGITS 2006. LNCS, vol. 4032, pp. 321–332. Springer, Heidelberg (2006)
19. Wu, C.-L., Liao, C.-F., Fu, L.-C.: Service-Oriented Smart-Home Architecture Based on OSGi and Mobile-Agent Technology. *Systems, Man, and Cybernetics, Part C. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 37(2), 193–205 (2007)
20. Lin, R.-T., Hsu, C.-S., Chun, T.Y., Cheng, S.-T.: OSGi-Based Smart Home Architecture for heterogeneous network. In: 3rd International Conference on Sensing Technology, ICST 2008, November 30-December 3, pp. 527–532 (2008)
21. Paksuniemi, M., Sorvoja, H., Alasaarela, E., Myllylä, R.: Wireless sensor and data transmission needs and technologies for patient monitoring in the operating room and intensive care unit. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, September 1-4, pp. 5182–5185 (2005)
22. Frehill, P., Chambers, D., Rotariu, C.: Using Zigbee to Integrate Medical Devices. In: Proceeding 29th Annual International Conference of the IEEE EMBS, August 23-26, pp. 6717–6720 (2007)
23. Jennic. JN5139-EK036, 6LoWPAN Evaluation Kit, V1.0 (2009)
24. Jennic. 6LoWPAN APIs User Guide, Revision 1.0 26 (February 2009)

Low Frequency Domain Aided Texture Synthesis for Intra Prediction

Xiaowei Sun, Baocai Yin, and Yunhui Shi

Multimedia and Intelligent Software Technology Beijing Municipal Key Lab,
College of Computer Science and Technology,
Beijing University of Technology, 100124, Beijing, China
sunxw@emails.bjut.edu.cn, {ybc, syhzm}@bjut.edu.cn

Abstract. To save bit-rate, texture synthesis can be employed in video coding to generate similar texture instead of encoding the texture region. Without mean squared error (MSE) as the distortion criterion, high quality but different texture can be synthesized with few bits. One key problem that embarrasses the application of texture synthesis in video coding is the annoying artifacts. In this paper, we describe a new synthetic region selection strategy which can simultaneously save bit-rate and avoid the annoying artifacts. Our method is based on low frequency consistent. That is, a low resolution version of video sequence is coded and sent to the decoder side, and the texture synthesizer should strictly accord with the decoded low resolution picture. In the high resolution layer, structure and sampler blocks are identified and encoded using MSE as the distortion criterion. We implement our scheme in JSVM and the bit-rate saving account to 18% compared with SVC with the same subjective quality.

Keywords: Intra prediction, texture synthesis, low frequency domain consistent.

1 Introduction

Video compression is one of the hottest research areas that have received considerable attention over the last few decades. Various techniques, such as motion compensation, intra prediction and transform are employed in the state-of-the-art video coding standards, which regard the compression as a signal processing task, to exploit statistical redundancy based on the MSE criterion. Focusing on pixel-wise redundancy rather than perceptual redundancy, the types of redundancies exploited by these video coding schemes are rather limited. The second generation image and video coding attempt to identify features within the image and use the features to achieve compression. All second-generation techniques incorporate properties of the human visual system (HVS) into the coding strategy in order to achieve high compression ratios while still maintaining acceptable image quality [5][6].

In recent years, advancements in texture synthesis are leading to promising efforts to exploit visual redundancy. In video scene, large amounts of texture regions can be regarded as subjectively irrelevant. That is, whatever the original texture is, if the similar texture is used to replace the original ones, they can convey the same information as the real ones do. In most cases, those regions contain a high amount of visible

details but carry little information relative to the structure regions. On the other hand, when using MSE as the distortion criterion, it usually requires high bit-rates to encode those regions. To simultaneously save the bits used in these regions and preserve high quality of the decoded sequence, texture synthesis methodology is employed to replace these regions with high quality synthetic texture. It has been reported that improvement is achieved by employing texture synthesis in compression even though in a straight-forward fashion.

In a series publication [2][3], Liu etc. propose to utilize the edge-based inpainting as an intra prediction method in the context of block-based compression. Specifically, the current block to be predicted is divided by edges into some unknown regions, each of which is independently generated by the Laplace equation. Two differences can be found Compared with H.264. Firstly, edges are used in this method to replace the predefined directional modes. It is generally believed that edges represent salient image singularities, as well as kinds of directional information in pixel values. Moreover, edges are extracted according to local image features, instead of predefined, which makes the prediction adaptive to different image structures. Secondly, the prediction is generated by the Laplace equation, which is inspired by the PDE-based inpainting techniques and is believed to improve the accuracy of prediction. In order to drive the edge-based inpainting towards better rate-distortion performance, an edge estimation method is also presented to revise the traditionally extracted edges. In general, the intra prediction is uniformly defined and it is adaptive to local image features.

In [4], a structure-aware inpainting (SAI) method is developed to restore the skipped structural regions by taking advantage of the available portion of the decoded image. A binary structure map is extracted and compressed into the generated bit-stream to indicate the skipped regions with salient structures. By making use of the decoded texture information together with the structure map, the SAI method can recover the skipped structural regions as well as the non-structural ones effectively at the decoder. Given the binary structure info, the structure-aware inpainting method is proposed that will propagate the structure along the received structure info and then synthesize texture for the rest part at decoder. Patch-based inpainting, which consider the average pixel values between synthetic and non-synthetic patches, is used in the proposed scheme to recover the texture information lost in the unknown region.

Without MSE as its distortion criterion, high quality texture, which is different but similar to the original texture, can be generated under the instruction of few assistant bits. However, from the published experimental results, it's hard to say method via texture synthesis or inpainting can have the advantage of the current block based intra-prediction when using MSE as the distortion criterion. To further improve the coding efficiency, researchers are trying to break through the restraint of MSE.

It's common sense that we can acquire almost the same information from the low resolution video sequence as the high resolution one. In other words, the low frequency domain contains the most important information for the understanding of moving picture. Based on this assumption, we present a low frequency domain consistent based texture synthesis method for intra prediction and image compression. A low resolution version of video sequence is firstly coded and sent to the decoder side. All blocks in high resolution layer are classified into texture and structure ones. Texture blocks are synthesized according to the low frequency information to save bit-rate. The structure blocks which contain more information than the texture blocks and texture sampler are encoded using MSE as its distortion criterion.

On the other hand, the resolution diversity of current display devices also motivates the need for spatial scalability. With the expectation that future applications will support a diverse range of display resolutions, the Joint Video Team (JVT) has developed a scalable extension to the state-of-the-art H.264/AVC video coding standard [1]. This extension is commonly known as Scalable Video Coding (SVC) and it provides support for multiple display resolutions within a single compressed bit stream. Due to the similar purpose and architecture, we implement our proposed method on the platform of SVC.

The remainder of this paper is organized as follows. An in-depth description of our framework is given in Section 2. We discuss the encoder and decoder design in Section 3 and 4. Finally experimental results and discussion of our experiments are demonstrated in Section 5.

2 Framework of Our Proposal

In this section, we present the framework of our low frequency domain consistent based intra prediction scheme. Following the definition in SVC, We refer the lowest resolution video data in the spatially scalable system as the base layer, and the higher resolution video data as the enhancement layer. There is no special technique, like texture synthesis and inpainting, employed in the base layer, and it is encoded by an ordinary single-layer encoder. The decoded lower resolution version of an image is served as assistant information for the texture synthesis of a corresponding higher resolution sequence.

As same as the other texture-synthesis based video coding scheme, video sequence is analyzed in encoder to identify the sampler and synthetic region for the enhance layer. Non-homogeneity texture and texture sampler region is encoded with normal encoder. Side information like sampler and synthetic region mask is also encoded and sent to the decoder. Sampler and synthetic region mask define the position of the selected sampler and the dropped texture region. The difference between our proposal and SCV is that no residual information is encoded for selected synthetic region of enhancement layer. Instead, texture is synthesized in the decoder under the instruction of side information. Fig. 1 shows the framework of our scheme.

As we can see from Fig. 1(a), we assume the input picture is always the high resolution sequence, and it is firstly down sampled through the tool DownConvertStatic which is provided by SVC. The low resolution sequence is directly sent to the SVC

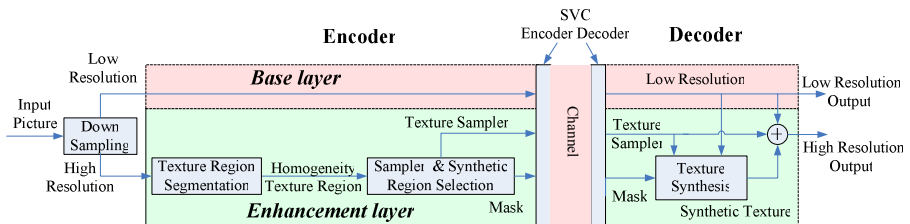


Fig. 1. Framework of our scheme

encoder. Texture analysis is performed for the enhancement layer. The high resolution picture is segmented into several homogeneity texture regions. For each homogeneity texture region, sampler blocks are identified and the rest is regarded as synthetic region. Texture sampler and its position information is encoded and passed to the decoder. Under the instruction of the decoded base layer, texture sampler and mask information, the decoder can generate high quality synthetic texture for the enhancement layer.

3 Encoder Design

The encoder generates all the assistant information for decoder to direct the synthesis process. Homogeneity clustering and sampler selection compose the main part of the encoder.

3.1 Homogeneity Clustering

Luma, chroma and edge information is considered jointly to estimate whether the blocks have homogeneous content. A multi-resolution quad-tree is built to speed up this process. When homogeneity texture has different color, e.g. grass with flowers on it, best possible segmentation results can be achieved on the top of the quad-tree, since in the top of the quad-tree we can filter the high frequency component of texture that may disturb the process of segmentation.

Two texture regions are considered to be similar if the distance between the corresponding feature vectors lies below a given threshold. For each block, the feature vector includes the average of luma, average of Cb, average of Cr, and variation of luma. The similarity threshold is optimized manually for some key frames of a given sequence. The optimal threshold is then used for all frames of the video. Color histogram can help an automatic segmentation algorithm for some sequence.

Chroma component with variation info is firstly used to classify the texture region at the top of the quad-tree, which gives a general distribution of homogeneous texture. The identified texture region is shown in Fig. 3 (c).

After that, canny operator is used to detect the edge info for every frame. To preserve the edge info which contain more information than the pure texture region, all the blocks containing edge are excluded from the removable texture region. Fig. 3 (d) is the edge info detected by canny operator, and Fig. 3 (e) is the texture region excluding the edge blocks.

3.2 Sampler Selection

In the synthesis-based compression schemes, how to select removable regions and useful samplers is significant to successful synthesis. In the following of this section, we will explain how to select exemplar regions and synthetic regions based on block categorization.

To get a proper sampler collection, inspired by inverse texture synthesis technique in [8], we cast our texture sampler selection process as an optimization problem. Specially, given an original texture X , our goal is to calculate a small texture collection Z with user-specified input parameter, minimizing

$$\Phi(x; z) = \frac{1}{|X|} \sum_{p \in X} |x_p - z|^2 + \alpha |Z|. \quad (2)$$

where z/x represents the sample values for Z/X , p is a texture index which runs through of X , x_p indicates the selection unit indexed by p , z is a subset of Z having the same shape with respect to x_p , and α is a user tunable weighting which is used to control the area of synthetic region.

The first term has the same function as described in [8]. It measures the local similarity for a set of samples in X with respect to Z . By calculating a Z that minimizes this energy term, we attempt to ensure that for every homogeneous texture region x_p , we could find a corresponding set z_p that is similar to x_p . The first term is called the inverse term due to its inverse synthesis nature. With this inverse term, the texture collection can greatly preserve the most representative features of the original texture region. Because we assume that the area of sampler region is less than the area of synthetic region, we can always divide x into several x_p which have the same or similar shape of the current sampler z . In this case, the comparison is performed between part of X and the whole of Z , which have the same result as the forward synthesis do. The second term is added into the energy function to provide a user interface to control the synthetic area.

Sampler selection in our proposal is achieved by solving Equation 1. Our basic solver is inspired by texture optimization [9], but since our energy function contains inverse and sampler cost terms, we have to provide a different solver. For clarity, the core part of our solver still follow the usage of E/M-steps as described in [9]. At E-steps, we solve for z to minimize the energy function. At M-steps, we calculate the difference between Z and all subset x of X which has the same shape with Z . If the current x is smaller than Z in area, the difference is derived from x and a subset of Z which has the same shape and the most similar pixel values with x . The output of one step feeds as input to the other, and we iterate this process several times until a pre-determined number of iterations is reached.

4 Decoder Design

In the decoder, base layer, sampler of the enhancement layer and the mask is decoded for the texture synthesis step. Forward texture synthesis is performed between the synthetic and sampler region of base layer to find a texture index for each synthetic unit in the enhancement layer, which is shown in Fig. 2. To avoid the annoying artifacts on the boundary of the current synthetic block, the synthetic unit in our proposal always has an irregular shape. To be specific, each block of base layer is divided into several sub-regions according to their similarity of pixel value, and each sub-region performs synthesis independently.

The sampler texture with minimum diversity in corresponding base layer with the current synthetic region is regarded as current sub-region's synthetic texture. The diversity is measured by D defined as follows:

$$D = \sum_{x_p \in X} SSD(x_p, z_{pq}) / |z_{pq}| \quad \text{with} \quad (3)$$

$$q = \arg \min_q (SSD(x_p, z_{pq} + Ax_p - Az_p))$$

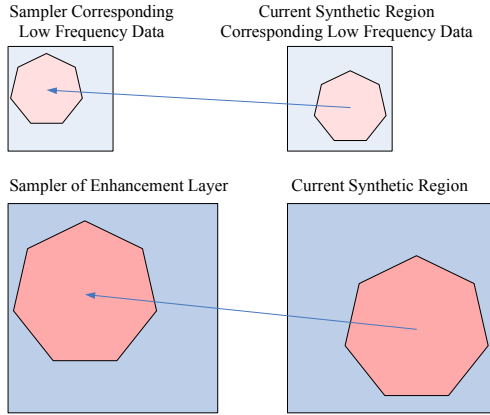


Fig. 2. Low frequency Domain Aided forward synthesis on Sub-region

Where X is the identified texture region of the input frame, SSD denotes the sum of squared difference of the input pixel, x_p and z_p represent the pixel values of the synthetic and sampler region, Ax_p and Az_p is the DC value of the current synthetic and sampler unit. To reduce the computational complexity, we use the upsampling data as its synthetic texture for blocks with small variation.

After the texture synthesis, low frequency compensation is performed to keep the low frequency character of synthetic texture identical to that in the base layer. The reason for this additional low frequency compensation is that no matter how close is the synthetic texture to the base layer texture, there is always difference between them. To eliminate this difference and preserve the reality of synthetic texture, all the synthetic texture in the enhancement layer should be revised according to the base layer pixel value.

5 Experimental Results

We have integrated the proposed coding scheme into the SVC reference software and have compressed the frames of video sequences “Container”, “foreman”, “Coast-guard” and several standard still test pictures. In our experiments, we regard QCIF as base layer and CIF as the enhancement layer. Besides, rate distortion optimization (RDO) and entropy coding is turned on.

Fig. 3 (f) shows the reconstructed picture of “cameraman”. Although texture region is identified strictly according to the luma, chroma and edge information, some structure blocks will still be classified into texture region because their statistical properties fall in the range of the texture region. But this won’t cause negative impact on the visual quality of synthetic texture due to low frequency compensation. And the low frequency aided forward synthesis algorithm may also help to choose a similar patch for the current sub-region.

Our approach has a similar visual quality level compared with the reconstruction of SVC. Average bit-rate saving of our approach for all testing sequences account to 18% under different quantization parameters. Due to large amounts of bits saved in the texture region, the structure region which is more sensitive to HVS can be allocated more bits, and the quality of encoded picture can be ensured.

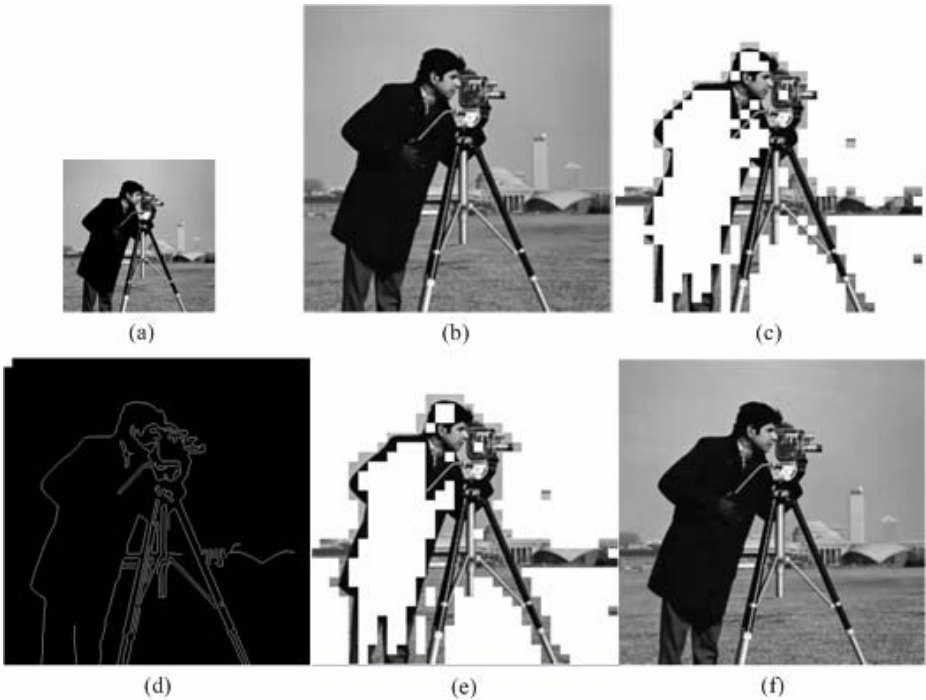


Fig. 3. Output of our experiments (a) base layer; (b) upsampled by SVC; (c) identified texture region represented by the white blocks; (d) edges extracted by canny operator; (e) revised homogeneity texture by canny operator; (f) output of texture synthesis

Acknowledgements

This paper is supported by National Hi-Technology Research and Development Program (863) of China (2006AA01Z317), National Natural Science Foundation of China (60825203, 60533030) and Scientific Research Common Program of Beijing Municipal Commission of Education (KM200710005017).

References

1. Segall, C.A., Sullivan, G.J.: Spatial Scalability within the H.264/AVC Scalable Video Coding Extension. *IEEE Transactions on Circuits and Systems for Video Technology* 17(9), 1121–1135 (2007)
2. Liu, D., Sun, X., Wu, F., Li, S., Zhang, Y.Q.: Image Compression With Edge-Based Inpainting. *IEEE Transactions on Circuits and Systems for Video Technology* 17(10), 1273–1287 (2007)
3. Liu, D., Sun, X., Wu, F.: Intra Prediction via Edge-Based Inpainting. In: *Data Compression Conference*, pp. 282–291 (2008)
4. Wang, C., Sun, X., Wu, F., Xiong, H.: Image compression with structure-aware inpainting. In: *Proc. IEEE International Symposium on Circuits and Systems*, pp. 1816–1819 (2006)

5. Torres, L., Kunt, M., Pereira, F.: Second Generation Video Coding Schemes And Their Role In Mpeg-4. In: European Conference on Multimedia Applications, Services and Techniques, pp. 799–824 (1996)
6. Reid, M.M., Millar, R.J., Black, D.N.: Second-generation image coding: an overview. *ACM Computing Surveys (CSUR)* 29(1), 3–29 (1997)
7. Dumitras, A., Haskell, B.G.: An encoder-decoder texture replacement method with application to content-based movie coding. *IEEE Transactions on Circuits and Systems for Video Technology* 14, 825–840 (2004)
8. Wei, L.Y., Han, J., Zhou, K., Bao, H., Guo, B., Shum, H.Y.: Inverse texture synthesis. In: *Proc. ACM SIGGRAPH* (2008)
9. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture Optimization for Example-based Synthesis. In: *Proc. ACM SIGGRAPH*, pp. 795–802 (2005)
10. Wang, H., Wexler, Y., Ofek, E., Hoppe, H.: Factoring Repeated Content Within and Among Images. In: *Proc. ACM SIGGRAPH* (2008)

A Novel Geometry Image Coding

Yunhui Shi, Wen Wen, Baocai Yin, and Jijun Shi

Beijing Key Laboratory of Multimedia and Intelligent Software Technology,
College of Computer Science and Technology,
Beijing University of Technology, Beijing 100124, China
syhzm@bjut.edu.cn, wenwen0305@emails.bjut.edu.cn,
ybc@bjut.edu.cn, jijun.shi@gmail.com

Abstract. Nowadays computer graphics over the Internet has attracted a lot of attentions; however transmitting 3D model takes a long time. Geometry image is a completely regular structure to remesh an arbitrary surface. It captures the geometry information of the mesh as an array of $[x, y, z]$ values. In this paper, a novel framework of geometry image coding is proposed which utilized the image compression standard JPEG2000 to represent 3D model. To improve the quality of encoding the geometry image, a directional lifting wavelet transform-based image compression algorithm is utilized. Besides, we use an iterative algorithm in the parameterization phase to improve the quality of geometry image. According to the experiments, both objective and subjective performances of 3D model are enhanced, the photorealistic rendering is improved and the time of transmission is reduced.

1 Introduction

With the development of computer graphics, highly detailed 3D models with billions of surfaces are presented and computer graphic is widely applied on the Internet. However, due to the limitations of storage, transmission or display, the compression, especially progressive compression of 3D model, is becoming an urgent research topic. With progressive compression user can download and render 3D model at real-time, upon that, a lot of works have been done both in literature and in industry.

Geometry image [1] is one of those techniques which use a completely regular structure to remesh an arbitrary surface. Geometry image can capture the geometry information as a 2D array. Other surface signals like normals and colors can also be stored in similar 2D arrays which share the same surface parameterization. The benefit of geometry image is that it can be encoded using traditional image compress algorithms such as wavelet coders. So, it is very suited for hardware rendering and used in many applications including compression, level-of-detail, and remeshing.

Normal mapping is a prevalent rendering technique to reveal the details of complex geometry object, and improves the visual quality. For the normal map can be represented as true-color image, some compression techniques have been developed. In [2], a VQ-based compression algorithm is proposed by taking the advantage of the limited spatial distribution of the normal vectors. However, it is unsuitable for normal map image to compress associating with geometry image.

In this paper, for the purpose of a better reconstruction quality of 3D model and making this method more adapt to industrialize, we propose a scheme to encode the geometry image using improved JPEG2000. We use an iterative algorithm [3] in the parameterization phase to improve the quality of geometry image. Furthermore, an efficient directional lifting wavelet transform based image coding is utilized for increasing the reconstruction quality of geometry image and normal map image, consequently improving the photorealistic rendering.

2 Proposed Framework of Geometry Image Coding

2.1 Introduction to the Proposed Framework

The proposed framework to compress geometry image using JPEG2000 is shown in Fig 1, and the major steps of the proposed scheme are as follows:

- 1) Obtain the optimal initial cut path of the original mesh. In order to parameterize the original mesh onto a 2D square, an optimal cut path must be found to ensure the quality of the parameterization.
- 2) Parameterize the mesh onto a 2D square with the cut path obtained in step 1.
- 3) Resample the surface at grid points in the square domain, and calculate the vertex coordinates $[x, y, z]$ of the sampling points, which can be seen as the $[r, g, b]$ values of an image, then geometry image was created. Other 2D arrays of geometric attributes such as normal maps and texture coordinates can be created also.
- 4) Encode and parallel transmit geometry image and other 2D image of geometric attribute using JPEG2000 which based on directional lifting wavelet transform.
- 5) The geometry image can be reconstructed by the decoder and the coordinate of every sampling point can be reconstructed.

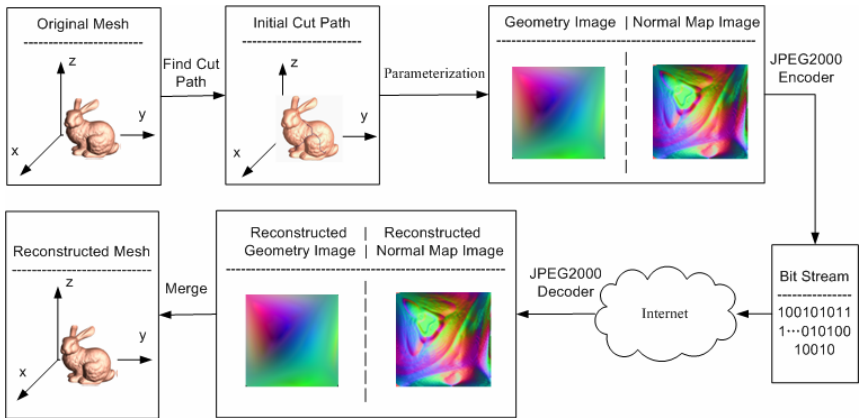


Fig. 1. Framework of the proposed method

The rest of this section is organized as follows. Section 2.2 describes how to find the optimal initial cut path. Section 2.3 gives a description of the parameterization algorithm. Section 2.4 details the directional lifting wavelet transform algorithm. The advantages of using the proposed framework are discussed in Section 2.5.

2.2 The Algorithm to Find the Optimal Initial Cut Path of the Mesh

The parameterization of a whole mesh onto 2D square will generate great distortions. So, how to find an optimal cut path is very important for the parameterization. In this paper, we use the algorithm proposed in [1] to find the optimal cut path.

The procedure of finding the optimal cut path is as follows. First, we use the algorithm mentioned in [1] to find the initial path. Then, an iterated cut path augmentation step is performed to find the optimal cut path. Fig 2 shows the iterations of finding the optimal cut path. It is important to use Floater parameterization [4] in this algorithm because the Floater parameterization can evenly distribute the stretch of the parameterization.

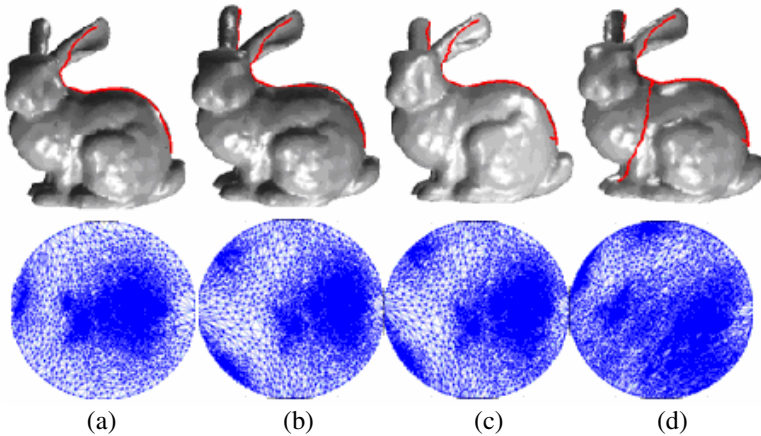


Fig. 2. Columns (a-d) show iterations of the iterated cut path improvement algorithm. Upper images show the model Bunny with the current cut path (red). Bottom images show the Floater parameterization (over circle) under the current cut path.

2.3 Iterative Parameterization Algorithm

The parameterization algorithm used in [1] can generate high quality mesh parameterizations, but the mesh parameterization often generates regions of high anisotropic stretch, consisting of slim triangles. So we use a fast and simple stretch-minimizing mesh parameterization algorithm proposed in [3]. The benefit of this algorithm is that it does not generate slim triangles in the parameterization region which cause crack in the reconstructed mesh.

The parameterization algorithm proposed in [3] is heuristic. Fig 3 shows the iterations of parameterization algorithm. As can be seen from Fig 3, the iterations improve the quality of the parameterization.

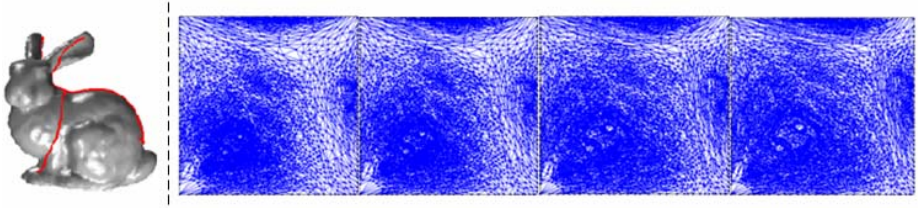


Fig. 3. Illustration of the iterative parameterization algorithm. Left image is the model with optimal cut path and right images show the iterations of obtaining the best parameterization.

2.4 Directional Lifting Wavelet Transform

In order to obtain a better reconstruction quality of 3D model, a compression technique based on directional lifting wavelet transform [5] is utilized for the geometry and normal map image coding. As for geometry images, detailed information is important which have an effect on the reconstructed 3D model. Moreover, normal map image contains a mass of high-frequency details, and encoding these details is difficult. Despite all that, directional lifting wavelet transform is adapted to enhancing the quality of reconstructed images, and then the reconstruction quality of 3D model is increasing also.

The traditional lifting technique is an efficient implementation of wavelet transform. However, the prevailing 2D lifting wavelet transform always predicts and updates in vertical or horizontal directions. Thus, it is unsuited to the image with arbitrary orientation, or even result in large-magnitude high-frequency coefficients.

The directional lifting technique aims at exploiting the spatial correlation among neighboring pixels. First is to analyze the spatial correlation in all directions for a pixel, and then selects the direction with the minimum prediction errors instead of always predicting along vertical or horizontal direction. Like classical lifting, it is consisting of three steps: split, predict and update. Let $x(m,n)_{m,n \in \mathbb{Z}}$ be a 2D image, and then split all samples into two sub-sets: the even sample set and the odd sample set.

$$\begin{cases} x_e(m,n) = x(m,2n) \\ x_o(m,n) = x(m,2n+1) \end{cases} \quad (1)$$

In the predict step, pixels in the even set are used to predict the pixels in the odd set, and the direction with minimum prediction residual is selected, so the spatial prediction can be more accurately. In [5], quarter pixels are used to achieve high resolution and they can be calculated by any existing interpolation method. However, different fractional pixel precision is suitable for different image. For the geometry image, quarter pixels are needless by reason that it is relatively smooth. Upon that, half pixel in the even set is enough to predict pixel in odd set. However, quarter pixels are needed for coding normal map image, since it is relative less and contains a mass of details. The directions for geometry image and normal map image are shown respectively in Fig 4. Consequently, the odd samples are predicted as follows.

$$P_e(m,n) = \sum_i p_i x_e(m + \text{sign}(i-1) \frac{\text{dir} - d}{d}, n + i) \quad (2)$$

Where $sign(x)$ is 1 for $x \geq 0$ and -1 otherwise, for geometry image, $d=2$, $dir=0, \dots, 4$, and for the normal map image, $d=4$, $dir=0, \dots, 8$. Then, the high-frequency coefficient $h(m,n)$ is calculated with

$$h(m,n) = x_o(m,n) - P_x(m,n) \tag{3}$$

In update step, for economize the bits to code the side information of direction, the update step of directional lifting is performed along the same direction as that in the predicting step. Consequently, the updating value $UB_{hB}(m,n)$ for even samples is as follows.

$$U_h(m,n) = \sum_j u_j h(m + sign(j) \frac{dir-d}{d}, n+j) \tag{4}$$

The parameters of the formula contain the same meaning like the predict step. After that, the sample at the even set is updated to produce low-frequency coefficient $l(m,n)$, which is calculated with

$$l(m,n) = x_e(m,n) + U_h(m,n) \tag{5}$$

Similar to the compression technique in [5], rate distortion optimization and direction prediction are utilized to reducing the bits of coding direction also. Moreover, we try to find an approximate offset δ to estimate the R-D slope, in order to make it suitable for geometry image and normal map image coding.

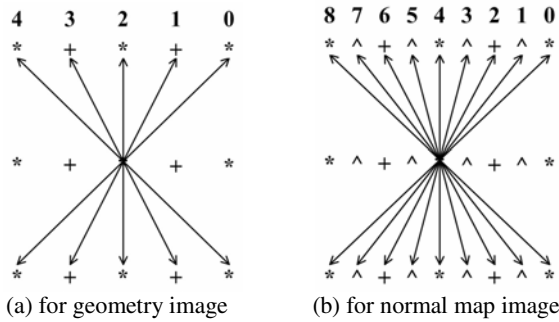


Fig. 4. Directions in predict and update step where integer pixels are marked by '*', half pixels by '+', and quarter pixels by '^'

2.5 Advantages in Applying the Proposed Framework

The main advantage offered by JPEG2000 is the flexibility of the code stream. The code stream can be decoded by truncating the code stream at any point. As this characteristic of progressive compression and decompression JPEG2000 supports, it allows us to achieve 3D model progressing compression. When an image at lower resolution obtained, the representation of 3D model can be render firstly. Then, the more code stream received, the higher 3D model resolution gained. Upon that, 3D model can be transmitted over the Internet progressively to reduce the waiting time.

Similar to the compression technique for geometry image, normal map image can be compressed and parallel transmitted with geometry image progressively. Therefore, the 3D model is reconstructed by geometry image and normal map image which are transmitted synchronously, and the photorealistic effect is improved progressively.

3 Experimental Results

To verify the effectiveness of the proposed method, this proposed scheme has been implemented into JPEG2000 reference software vm9.0 for geometry and normal map images coding and regenerate the 3D model from those reconstructed images.

The first experiment evaluates the ability of directional lifting wavelet transform for geometry and normal map images coding compared with the original JPEG2000. We simply replace the conventional lifting wavelet transform module with the directional lifting wavelet transform and use others technique as same as JPEG2000, so as to evaluate the proposed performance objectively. Fig 5 shows the result of encoding geometry image and normal map image. According to experiment, the coding gain of proposed normal map image coding can be up to 1.0 dB.

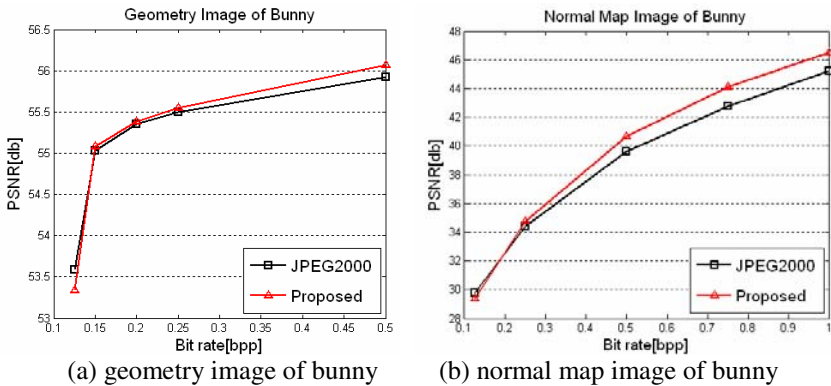
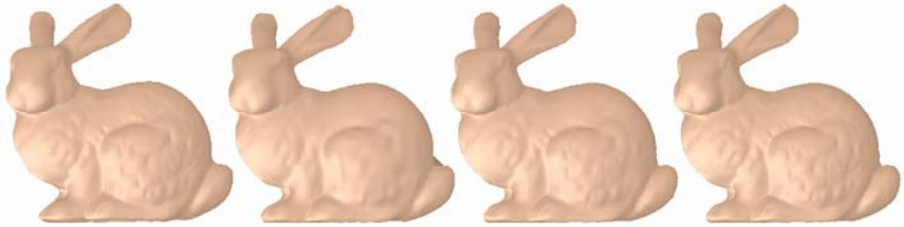


Fig. 5. Rate distortion for geometry and normal map image

In the second experiment, we regenerate the 3D model from reconstructed geometry image, and calculate the Hausdorff distance [6] between the original and reconstructed model to quantify the visual differences. The Hausdorff error is calculated with the function as follows.

$$error_{Hausdorff} = \frac{distance_{Hausdorff}}{bounding\ box\ diagonal} \tag{6}$$

With transmitting progressively for bunny model, in Fig 6, (b), (c), (d) are the progressive rebuilt model from coarse to fine, (d) is the final reconstructed model and their Hausdorff error are 0.32198%, 0.30798%, and 0.29452%. As can be seen, the visual differences are getting smaller and smaller, and the final reconstructed one has very good quality compared with the original model. Besides, by applied better reconstructed normal map, the photorealistic rendering of 3D model is improved.



(a) original model (b) (c) progressive reconstructed model (d) final reconstructed model

Fig. 6. 3D model render with progressive transmission

4 Conclusions

In this paper, we propose a novel framework to compress geometry image using JPEG2000. We first find the optimal cut path of the mesh and then perform parameterization. Then we compress the geometry image and normal map image with directional lifting wavelet transform. Since it is implemented in JPEG2000, we can utilize matured international standard of image compression to deal with progressively transmit and render 3D model. Most importantly, a better reconstruction quality of 3D model is achieved, and the photorealistic effect is improved.

Acknowledgements. This work was supported by: National Natural Science Foundation of China (60825203, 60533030); National High Technology Research and Development Program of China (2006AA01Z317); Scientific Research Common Program of Beijing Municipal Commission of Education (KM200710005017).

References

1. Gu, X., Gortler, S.J., Hoppe, H.: Geometry images. In: ACM SIGGRAPH 2002, pp. 355–361 (2002)
2. Yamasaki, T., Aizawa, K.: Fast and Efficient Normal Map Compression. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 9–12 (2006)
3. Shin, Y., Alexander, B., Hans-Peter, S.: A fast and simple stretch-minimizing mesh parameterization. In: Proceeding Shape Modeling International 2004, pp. 200–208 (2004)
4. Floater, M.: Parameterization and smooth approximation of surface triangulations. In: CAGD 1997, pp. 231–250 (1997)
5. Ding, W., Wu, F.: Adaptive Directional Lifting-Based Wavelet Transform for Image Coding. *IEEE Transactions on Image Processing* 16(2), 16–427 (2007)
6. Aspert, N., Santa-Cruz, D., Ebrahimi, T.: Mesh: Measuring Errors between Surfaces using the Hausdorff distance. In: IEEE International Conference in Multimedia and Expo, vol. 1, pp. 705–708 (2002)

Musical Style Classification Using Low-Level Features

Armando Buzzanca, Giovanna Castellano, and Anna Maria Fanelli

Department of Computer Science, University of Bari, Italy

Abstract. In this paper we address the problem of musical style classification. This problem has several applications like indexing in musical databases or development of automatic composition systems. Starting from MIDI files of real-world improvisations, we extract the melody track and cut it into overlapping segments of equal length. From these fragments, numerical features are extracted as descriptors of style samples. Then a cascade correlation neural network is adopted to build an effective musical style classifier. Preliminary experimental results show the effectiveness of the developed classifier that represents the first component of a musical audio retrieval system.

1 Introduction

Musical style as well as the mechanisms underlying style classification are relatively ill-defined [1]. Several definitions of musical style have been formulated so far. In [3] musical style is defined as “the identifiable characteristics of a composer’s music which are recognizably similar from one work to another”. Another definition of musical style is given in [5]: “Style is a replication or patterning, either in human behavior or in the artifacts produced by human behavior, that results from a series of choices made within one set of constraints”. In [4] theoretical comprehensive guidelines for style analysis are provided by dissecting musical style into three dimensions: large (groups of works, work, movement), middle (part, section, paragraph, sentence) and small (motive, subphrase, phrase).

Whatever the definition, musical style (and its classification) is something related to human nature: any layperson can recognize the difference among simple stylistic features. Conversely, automatic recognition of musical style is not an easy task.

Despite its intrinsic difficulty, automatic classification of musical styles is gaining more and more importance since it may serve as a way to structure and organize the increasingly large number of music files available on the Web. Actually, styles and genres, typically created manually by human experts, are used to organize music content on the Web. Automatic musical style classification can potentially automate this process and provide an important component for a complete music information retrieval system. Therefore, building a classifier that can recognize different musical styles is of primary interest.

In this paper, we address the automatic classification of musical styles by means of a feature-driven approach. More specifically, a feature set for representing a musical monophonic excerpt is considered. The significance of the proposed features is investigated by training a neural network with a particular architecture, using real-world

data collected from actual performances of a musician playing the piano. Using the proposed feature set, good classification for all the considered musical styles is achieved. These results are comparable to those reported from human musical style classification.

The paper is structured as follows. Section 2 describes the process to extract features useful to describe the content of a musical excerpt played according to a specific style. Section 3 deals with the automatic classification of styles using a Cascade Correlation neural network trained on the extracted features. In Section 4 preliminary results are given. Finally, in Section 5 some conclusions and future directions are drawn.

2 Extraction of Low-Level Features

The first step for automatic musical style classification is feature extraction, that is the process of computing a compact numerical representation to be used for characterizing a musical excerpt. To this aim, we consider a number of styles that a performer could improvise playing a musical instrument like the piano and using only one melodic line. We assume standard MIDI files as the source of monophonic melodies.

In order to extract a number of samples for each style, and a number of features for describing each sample, we first transform the MIDI file in text format using the Midi2txt tool [7] that allows to extract information such as midi key number, duration (in ms) and volume. Then, we apply a parser that analyzes the text file and extracts useful information from the file. The parser was developed in Borland Delphi 7.0 and presents a graphical interface (see fig. 1) that allows the user to visualize useful information extracted from the MIDI file. Specifically, the parser extracts information contained in the MIDI file header:

- Midi File Type (in our case always 0 because we assume only one melodic line)
- Number of tracks;
- Number of ticks;
- Tempo (here we deal only with melodies written in 4/4);
- Beats.

as well as information concerning the MIDI tracks:

- Key Number: it is the note that was pressed;
- Duration: the distance in pulses from the event that onsets the sound of a note to the finishing event;
- Duty Factor: the ratio of the time between midi note_on and midi note_off;
- Pitch: the value each note can take and ranges from 0 to 127;
- Volume;
- Counts of notes.

Summarizing, the parser provides two text files, a file named “general_info” containing general information extracted from the midi file (note on, note off, volume, duration in ms) and a file named “track_info” containing information extracted from

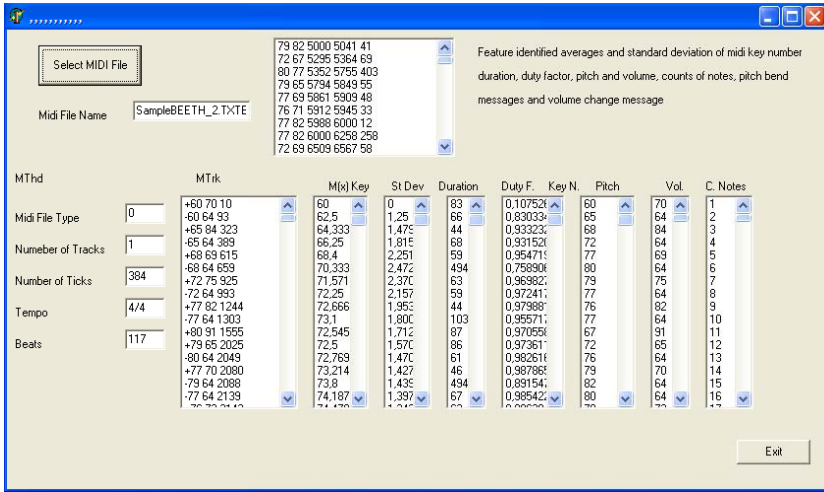


Fig. 1. Graphical interface of the parser developed to extract information from MIDI files

the midi tracks. In order to derive a number of samples for each style, data contained in the “track_info” file of each musical improvisation are processed according to the following scheme.

Each minute of performance is divided into six blocks of 10 seconds, as proposed in [11]. Each 10-sec segment is then divided into six overlapping intervals, as depicted in fig. 2, with duration of 5 seconds. On the overall, from each minute of performance, we extract 36 samples. Then, for each 5-sec interval the following statistics are computed:

- average and standard deviation of the Key Number
- average and standard deviation of the Duration
- average and standard deviation of Duty Factor
- average and standard deviation of Pitch
- average and standard deviation of Volume
- count of notes

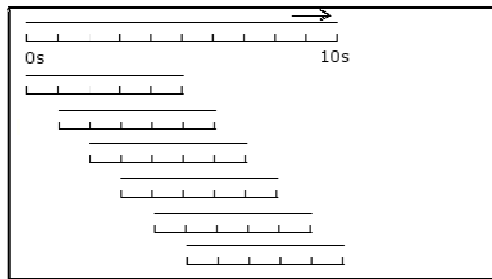


Fig. 2. Overlapping intervals extracted from the 10-sec segment

To summarize, each sample is a 11-dimensional vector describing a 5-sec fragment of a played excerpt. Since the problem of musical style classification is mainly a supervised one, each sample is labeled with a number indicating the class (musical style) it belongs to.

3 Classification of Musical Style

Once features are extracted, a pattern recognition technique area can be used to classify musical styles. In this work, we employ the Cascade-Correlation neural networks [12]. This kind of network is based on a cascade architecture, in which hidden units are added to the network one at a time and do not change after they have been added. During learning, new hidden units are created. For each new hidden unit, the magnitude of the correlation between the new unit's output and the residual error signal is maximized. The cascade architecture is illustrated in fig. 3.

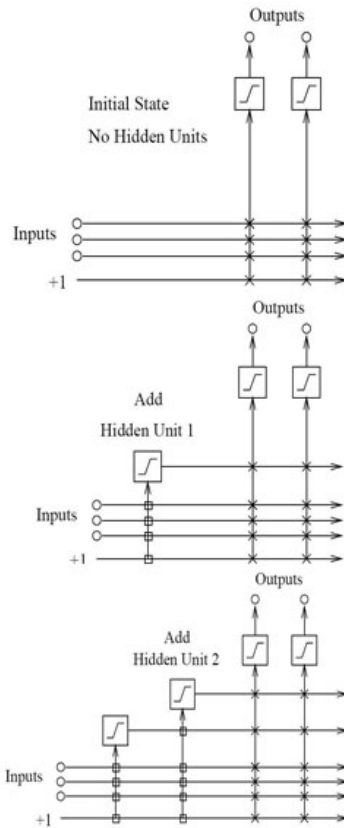


Fig. 3. The Cascade Correlation Architecture

The network structure is initialized with some inputs and one or more output units, with no hidden units. The number of inputs and outputs is dictated by the problem and by the I/O representation. Every input is connected to every output unit by a connection with an adjustable weight. There is also a bias input, permanently set to +1. The output units may just produce a linear sum of their weighted inputs, or they may employ some non-linear activation function. In our work, we use a symmetric sigmoidal activation function (hyperbolic tangent) whose output range is -1.0 to +1.0.

During the learning phase, hidden units are added to the network one by one. Each new hidden unit receives a connection from each of the network's original inputs and also from every pre-existing hidden unit. The hidden unit's input weights are frozen at the time the unit is added to the net; only the output connections are trained repeatedly. Each new unit therefore adds a new one-unit "layer" to the network, unless some of its incoming weights happen to be zero. This leads to the creation of very powerful high-order feature detectors, but it also may lead to very deep networks and high fan-in to the hidden units. There are a number of possible strategies for minimizing the network depth and fan-in as new units are added [12]. We are currently exploring some of these strategies.

The learning algorithm begins with no hidden units. The direct input-output connections are trained as well as possible over the entire training set. To train the output weights, we use the quickprop algorithm [13]. With no hidden units, quickprop acts essentially like the Delta rule, but it converges much faster. At some point, this training will approach an asymptote. If no significant error reduction occurs after a maximum number of training cycles (controlled by a "patience" parameter), we run the network learning one last time over the entire training set to measure the error. If the network performance is satisfying, the process stops; otherwise, some residual error has to be further reduced. We attempt to achieve this by adding a new hidden unit to the network, using the unit-creation algorithm described below. The new unit is added to the net, its input weights are frozen, and all the output weights are once again trained using quickprop. This cycle repeats until the error is acceptably small (or until we give up). To create a new hidden unit, we begin with a candidate unit that receives trainable input connections from all of the network's external inputs and from all pre-existing hidden units. The output of this candidate unit is not yet connected to the active network. We run a number of passes over the examples of the training set, adjusting the candidate unit's input weights after each pass. The goal of this adjustment is to maximize S , the sum over all outputs units of the magnitude of the correlation between V , the candidate unit's value, and E_o the residual output error observed at unit o . We define S as

$$S = \sum_o \left| \sum_p (V_p - \bar{V})(E_{p,o} - \bar{E}_o) \right| \quad (1)$$

where o is the network output at which the error is measured and p denotes the training sample. The quantities \bar{V} and \bar{E}_o are averaged over all samples. In order to maximize S , we must compute

$$\frac{\partial S}{\partial w_i} = E_{p,o} \sum_{p,o} \sigma_o(-\overline{E_o}) f'_p \quad (2)$$

where σ_o is the sign of the correlation between the candidate's value and output o , f'_p is the derivative for sample p of the candidate unit's activation function with respect to the sum of its inputs.

4 Experimental Results

In order to test our approach, we have chosen seven classical musical styles obtained by playing the piano. Data have been collected at the 'N. Piccinni' State Conservatory of Music in Bari, Italy, where a musician performed for us, playing the piano using only one melodic line, a total amount of about 35 minutes of improvisation into the seven selected styles. For each style, the musician played an excerpt corresponding to a part of an opera written by a not coeval composer according to that style. Table 1 reports details about the excerpts played. Melodies have been sequenced in real time using the piano Yamaha 48" Mark III Series Upright Disklavier that allows the recording of the played excerpt directly in MIDI format.

Table 1. Details of the excerpt played by a musician according to different musical styles

Author	Opera	Historical period	Duration (in ms)	No.of samples	Style
Czerny	Etude op.740 nr.3	1791–1857	183669	107	1
Chopin	Studio Opera 10 nr.1	1810-1849	195000	113	2
Bach	English Suite BWv 808: Prelude	1685–1750	235708	138	3
Beethoven	Piano Sonata op. 27 nr.2 (1st movement)	1770-1827	274111	161	4
Bach	English Suite BWv 807: Prelude	1685–1750	361242	215	5
Schubert	Impromptu op. 90 nr. 3	1797–1828	386961	228	6
Beethoven	Sonata op.57 "Appassionata" 3rd movement	1770-1827	430480	251	7

For each of the seven musical styles, a number of representative samples were derived, by processing the corresponding excerpt according to the feature extraction scheme described in Section 2. Due to different duration of played excerpt, a different number of samples was derived for each style (see Table 1). In order to have a uniform distribution of samples for each style, we decided to extract exactly 100 samples for each style.

Once the feature extraction process was completed, we built different datasets, in order to evaluate the classification accuracy with different number of musical styles. Specifically, we constructed the following datasets:

- 21 two-class datasets, including all combinations of two styles among seven.
- 35 three-class datasets, including all combinations of three styles among seven.
- 35 four-class datasets, including all combinations of four styles among seven.
- 20 five-class datasets, including all combinations of five styles among seven.
- 7 six-class datasets, including all combinations of six styles among seven.
- 1 seven-class dataset, including samples of all styles.

To evaluate the classification performance, a scheme based on leave-k-out was carried out. In our case $k=20\%$ of the size of the dataset. In other words, each dataset was randomly partitioned so that $4/5$ of data were used as training set and the remaining $1/5$ as testing set. The Cascade correlation neural network, implemented in Matlab 2008a, was applied to five different random partitions and the results were averaged. This ensures the calculated accuracy to be not biased because of a particular partitioning of training and testing. Indeed, if the datasets are representative of the corresponding musical styles then these results are also indicative of the classification performance with real-world unknown performances.

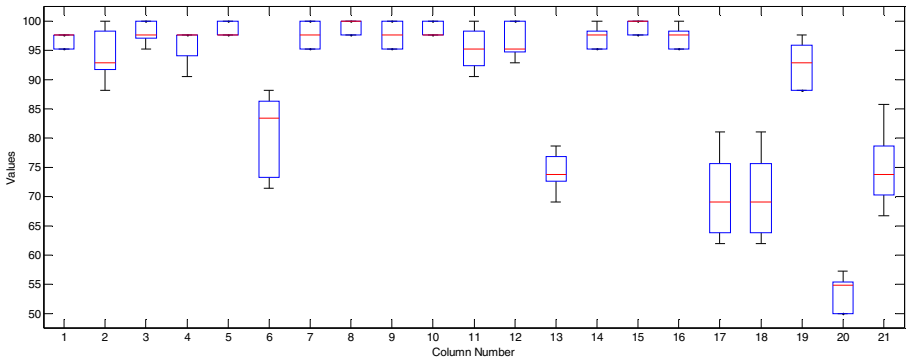
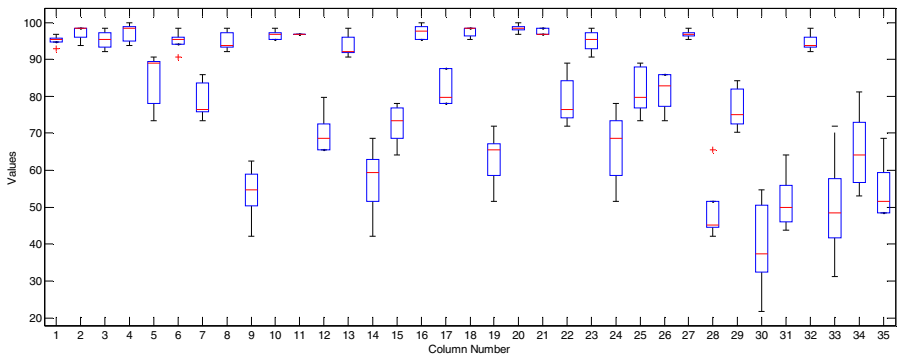
Classification results for each classifier are plotted in figures 4-8. As concerns the 2-style classifier, the best average performance was obtained with styles 1 and 2, with a classification rate of 99,57%. For 3-style classification, the best result is achieved with styles 1,2 and 4, with a classification rate of 98,89%. The best result for 4-style classifier was obtained with styles 1,2,4 and 5, with a classification rate of 96,07 %. Classifiers for 5 and 6 styles perform quite worst. Specifically, for 5 styles the best result was obtained with style 1,2,3,4 and 5, with a classification rate of 94,43%. As concerns the 6-style classifier, the best average performance was obtained with style 1,2,3,4,5 and 6, with a classification rate of 80,25%.

As concerns classification of all the seven musical styles, figure 9 shows more detailed information about the classification accuracy in the form of a confusion matrix. In a confusion matrix, the columns correspond to the actual style and the rows to the classified style. For example, the cell of row 7, column 3 with value 19 means that 100% of the style 3 (column 3) was wrongly classified as style 7 (row 7). The percentages of correct classification lie in the diagonal of the confusion matrix. The confusion matrix shows that the misclassifications of the system are similar to what a human would do. For example, samples of style 3 (Bach) are misclassified as style 7 (Beethoven). By analyzing the confusion matrix, it can be seen that Style 3 has the worst classification accuracy since it is easily confused with other styles. This result was somehow expected because of the broad nature of Style 3, which includes very general rules of composition.

Summarizing, the classifier developed on the basis of the extracted features has produced very interesting results for all the styles recognized, as depicted in table 2.

Table 2. Average classification accuracy for each classifier

Classifier	Cascade Correlation. Accuracy
2-style classifier	92,38
3-style classifier	81,46
4-style classifier	74,35
5-style classifier	66,73
6-style classifier	56,84
7-style classifier	50,21

**Fig. 4.** The box whisker chart with 2 styles**Fig. 5.** The box whisker chart with 3 styles

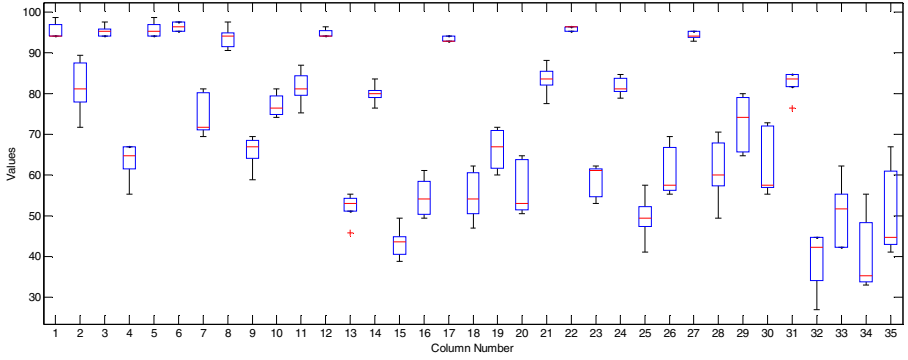


Fig. 6. The box whisker chart with 4 styles

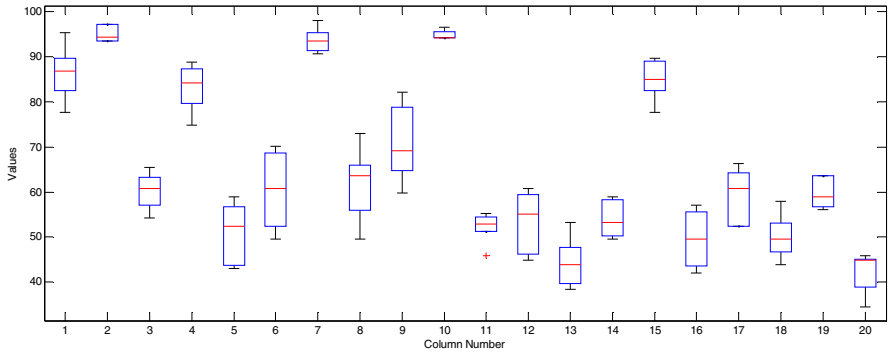


Fig. 7. The box whisker chart with 5 styles

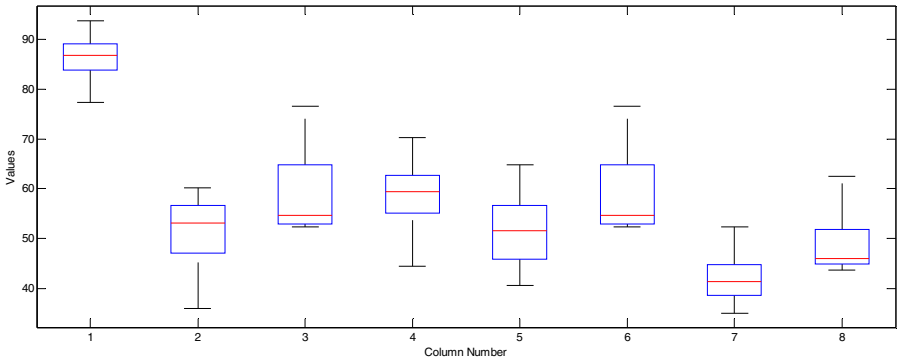


Fig. 8. The box whisker chart with 6 styles

	Style 1	Style 2	Style 3	Style 4	Style 5	Style 6	Style 7
Style 1	15	0	0	0	0	0	0
Style 2	0	17	0	0	0	0	0
Style 3	0	0	0	0	0	0	0
Style 4	0	0	0	15	0	0	0
Style 5	0	0	0	0	2	0	2
Style 6	0	0	0	0	0	11	0
Style 7	10	3	19	11	14	7	23

Fig. 9. Confusion Matrix concerning the 7-style classifier

5 Conclusions

This work has shown that a neural network classifier can be used to classify the musical style of an author once significant features are properly extracted. In our case the approach used has been tested with a free interpretation (improvisational style) with one melodic line.

Some of the misclassification can be caused by the lack of a smart method for melody segmentation. The music samples have been arbitrarily restricted to a duration of 5 sec, getting fragments not necessarily related to musical motives. The main goal of this work was to test the feasibility of the feature extraction approach, and an average recognition of 92,38% with the cascade correlation neural network is a very encouraging result keeping in mind these limitations. This is just a preliminary work, and a number of possibilities are still to be explored, such as the study of other features as significant descriptors. Moreover, in order to draw significant conclusions about the validity of the proposed approach, a large database of musical excerpts representing different styles has to be created and tested using our approach.

References

1. Crump, M.: A principal components approach to the perception of musical style. Paper at the Banff Annual Seminar in Cognitive Science (BASICS), Banff, Alberta (2002)
2. Brook, B.S.: Style and Content Analysis in Music. In: Gerbner, G., et al. (eds.) *The simplified Plaine and Easie Code in The Analysis of Communication Content*. Wiley, Chichester (1969)
3. Cope, D.: *Computers and Musical style*. A-R Editions, Inc., Madison (1991)
4. La Rue, J.: *Guidelines for style analysis*. W. W. Norton & Company, New York (1970)
5. Meyer, L.: *Style and Music*. University of Pennsylvania Press, Philadelphia (1989)
6. Buzzanca, G.: A supervised learning approach to musical style recognition. In: *Additional proceedings of the Second International Conference ICMAl*, Edinburgh, Scotland (2002)
7. Nagler, G.: MIDI2TXT v1.14 midi binaries to text mnemonic (1995), <http://www.gnmidi.com>
8. MacKay, D.J.C.: Information-based objective functions for active data selection. *Neural Computation* 4(4), 590–604 (1992)
9. MacKay, D.J.C.: Bayesian methods for backpropagation networks. In: Domany, E., van Hemmen, J.L., Schulten, K. (eds.) *Models of Neural Networks III*, ch. 6. Springer, New York (1994)

10. Bishop, C.M.: Neural Networks for pattern Recognition. Clarendon Press, Oxford (1995)
11. Dannenberg, R.B., Watson, T.D.: A Machine Learning Approach to Musical Style Recognition. School of Computer Science, Carnegie Mellon (1997)
12. Fahlman, S.E., Lebiere, C.: The cascade correlation learning architecture. Tech. Rep. CMU-CS-90-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1990)
13. Fahlman, S.E.: An Empirical Study of Learning Speed in Back-Propagation Networks Tech. Rep. CMU-CS-88-162, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1988)

Enterprise Cluster Dynamics and Innovation Diffusion: A New Scientific Approach

Marco Remondino, Marco Pironti, and Paola Pisano

University of Turin, e-Business L@B
Cso Svizzera 185, Turin, Italy
{remond,pironti,pisano}@di.unito.it

Abstract. A model in the field of enterprise management is described in this work. Its main goal is to represent and analyze the dynamics and interrelations among innovation diffusion and enterprise clusters formation and modifications. A formal description of the model is given, along that of its main parameters. Qualitative results are described. Clustering is definable as the tendency of vertically and/or horizontally integrated firms in related lines of business to concentrate geographically, or, to a more general extent, virtually. Innovation is a critical factor for the competitiveness of a National System, especially when the economy of the latter has come to maturity. However, the diffusion of innovations among its potential adopters is a complex phenomenon.

Keywords: Innovation diffusion, enterprise network, simulation, model.

1 Introduction

The studies about innovation prove that, beside the creation of innovations, it is also crucial to study their diffusion in the system in which the firms work and cooperate, i.e.: the network.

At that level, it is important to clarify what an enterprise network is and why the firms start to cooperate inside the network for diffusing an innovation.

A collaborative network is a whole of nodes and ties with a different configuration based on of what it has to achieve. These concepts are often displayed in a social network diagram, where nodes are the points and ties are the lines. The idea of drawing a picture (called a “sociogram”) of who is connected to whom for a specific set of people is credited to [5], an early social psychologist who envisioned mapping the entire population of New York City. Cultural anthropologists independently invented the notion of social networks to provide a new way to think about social structure and the concepts of role and position [7], [8], [23], an approach that culminated in rigorous algebraic treatments of kinship systems [29]. At the same time, in mathematics, the nascent field of graph theory began to grow rapidly, providing the underpinnings for the analytical techniques of modern social network analysis.

The nodes represent the different organizations that interact inside the network and the links represent the type of collaboration between different organizations.

The organizations could be Suppliers, Distributors, Competitors, Customers, Consultants, Professional Associations, Science Partners, Incubators, University, and so

on. The kind of partner firms linked over a network looks to be related to the type of innovation occurring: for example incremental innovators rely more frequently on their customers as innovation partners, whereas firms that have totally new products for a given market are more likely to collaborate with suppliers and consultants. Advanced innovators and the development of radical innovations tend to require a tighter interaction with universities. This point is supported by [15] in a survey of 4.564 firms in the Lake Constance region (on the border between Austria, Germany and Switzerland). By examining the interactions among firms, customers, suppliers and universities it emerges that firms that do not integrate their internal resources and competences with complementary external resources and knowledge show a lower capability of releasing innovations [14].

Philippen and Riccaboni [26], in their work on “radical innovation and network evolution” focus on the importance of local link formation and the process of distant link formation. Regarding the formation of new linkages Gulati [20] finds that this phenomenon is heavily embedded in an actor’s existing network. This means that new ties are often formed with prior partners or with partners of prior partners, indicating network growth to be a local process. Particularly when considering inter-firm alliances, new link formation is considered “risky business” and actors prefer alliances that are embedded in a dense clique were norms are more likely to be enforceable and opportunistic behavior to be punished [18], [21], [28], [2]. Distant link formation implies that new linkages are created with partners whom are not known to the existing partners of an actor. At the enterprise level, [6] shows that distant linkage that serve as bridge between dense local clique of enterprises, can provide access to new source of information and favorable strategic negotiation position, which improves the firms’ position in the network and industry.

In order to analyze the complex dynamics behind link formation and innovation diffusion, as long as their relationships, an agent based model is introduced in this work, and is formally analyzed.

2 Network Shape, Collaboration and Innovation Diffusion

The ties representing collaborations among firms can be different in structure, type and number.

- type of ties: strong or weak (depending on the type of collaboration: contracted development, licensing, research partnerships, joint venture, acquisition of an owner of a given technology);
- structure of ties: long or short (for example industrial districts in which firms are geographic clusters or virtual clusters); reciprocal or not (firms that exchange competences each other or simply give/take);
- number of ties: dense or not (depending on the number of links among the firms).

The type and the number of ties affect the network efficiency: for example, a network composed of relationships with partners comprising few ties among them would enable control for the principle partner. A network of many non-overlapping ties would provide information benefits: in [30] the authors suggest that the number of collaborative

relationships a firm is involved in, is positively related to innovation output while, conversely, closed networks have been found to foster innovation more than open ones [9]. A network composed of partners with many interlocking and redundant ties would facilitate the development of trust and cooperation.

The firm's position inside the network is as important as the number and type of ties. In [6] the authors find that rather than maximizing the number of ties, firms should strive to position themselves strategically in gaps between different nodes, so to become intermediaries. Contrary to this perspective, [3] propose that the best position is one where all the firms are tied only to the focal actor. On the other side, [4] suggests that the benefits of increasing trust, developing and improving collaboration and reducing opportunism shapes network structures creating cohesive interconnected partnerships. These consequent studies highlight that there is no consensus about which the optimal networking configuration should be. The configuration depends on the actions that the structure seeks to facilitate.

The firms start to collaborate inside a network for different reasons:

- risk sharing [16]
- obtaining access to new markets and technologies [17];
- speeding products to market [1];
- pooling complementary skills [12];
- safeguarding property rights when complete or contingent contracts are not possible [22];
- acting as a key vehicle for obtaining access to external knowledge [28], [10].

The literature on network formation and networking activity therefore clearly demonstrates that whilst firms collaborate in networks for many different reasons the most common reason to do so is to gain access to new or complementary competencies and technologies. Those firms which do not cooperate and which do not formally or informally exchange knowledge and capabilities limit their knowledge base on a long-term and ultimately reduce their ability to access exchange relationships.

When the innovation start to circulate, it can affect the network collaboration efficiency: firms can decide to cooperate inside the network by developing an external exploration behavior, meaning that a firm decides to be related to other organizations in order to exchange competences and innovations. Otherwise if the firm considers its internal capability to create innovation as a point of strength, or if the cost of external exploration is perceived as higher than that of internal research, then it could prefer to assume an internally explorative behavior in which it tries to create new competences (and possibly innovations) inside the organization itself.

During the process of innovation diffusion the network can change in the number of actors (exit and entry), and in numbers and patterns of link information [2]. The network can expand, churn, strengthen or shrink. Each network change is brought about by specific combination of changes in tie creation, tie deletion, and by changes in an actor's portfolio size (number of link) and portfolio range (numbers of partners) [2]. It's normal that the modification depends on the original structure of the network.

Also the propensity to collaborate inside a network affects innovation diffusion. When a network is a highly collaborative one, the innovation tends to diffuse more quickly, if the ties are dense, non redundant, strong and reciprocal. If the network is a

collaborative one, but the ties are weak or unidirectional, the innovation spreads slowly and could not reach all the nodes in the network.

To explore and analyze these complex social dynamics, an agent based model is described in the following paragraphs, that keeps into account most network and enterprise variables.

3 Agent Based Simulation

Why do enterprises team up? There can be many reasons for this strategy, leading, in its widest extent, to the creation of joint-ventures, i.e.: a new economical subject formed by two or more enterprises with the goal of new projects, or of clusters and networks of enterprises. The leading cause for these phenomena is the optimization of the production, by resources and competences sharing. Agent based simulation is an effective paradigm for studying complex systems. It allows the creation of virtual societies, in which each agent can interact with others basing on certain rules. The agents are basic entities, endowed with the capacity of performing certain actions, and with certain variables defining their state. In the model presented here, the agents are reactive, meaning that they simply react to the stimuli coming from the environment and from other agents, without elaborating their own strategies. When the model is formally built and implemented, it can be run by changing a parameter at a time, and emergence of a complex behavior occurs.

Agent based Modeling is thus one of most interesting and advanced approaches for simulating a complex system: in a social context, the single parts and the whole are often very hard to describe in detail. Besides, there are agent-based formalisms which allow studying the emergence of social behavior through the creation and study of models, known as artificial societies. Thanks to the ever increasing computational power, it has been possible to use such models to create software, based on intelligent agents, whose aggregate behavior is complex and difficult to predict, and which can be used in open and distributed systems.

In [11] we read that: “An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future”.

Another very general, yet comprehensive definition is provided by [24]: “...the term [agent] is usually applied to describe self-contained programs which can control their own actions based on their perceptions of their operating environment”.

Agents have traditionally been categorized as one of the following types [16]: Reactive; Cognitive/Deliberative; Hybrid.

When designing any agent-based system, it is important to determine how sophisticated the agents' reasoning will be. Reactive agents simply retrieve pre-set behaviors similar to reflexes, without maintaining any internal state. On the other hand, deliberative agents behave more like they are thinking, by searching through a space of behaviors, maintaining internal state, and predicting the effect of actions. Although the line between reactive and deliberative agents can be somewhat blurry, an agent with no internal state is certainly reactive, and one that bases its actions on the predicted actions of other agents is deliberative.

The agents used in this paper are reactive, but organized in the form of a MAS (Multi Agent System), which can be thought of as a group of interacting agents working together or communicating among each other. To maximize the efficiency of the system, each agent must be able to reason about other agents' actions in addition to its own. A dynamic and unpredictable environment creates a need for an agent to employ flexible strategies. The more flexible the strategies however, the more difficult it becomes to predict what the other agents are going to do. For this reason, coordination mechanisms have been developed to help the agents interact when performing complex actions requiring teamwork. These mechanisms must ensure that the plans of individual agents do not conflict, while guiding the agents in pursuit of the system goals. Many simulation paradigms exist; agent-based simulation is probably the one that best captures the human factor behind decisions. This is because the model is not organized with explicit equations, but is made up of many different entities with their own behavior. The macro results emerge naturally through the interaction of these micro behaviors and are often more than the algebraic sum of them. This is why this paradigm is optimal for the purposes of modeling complex systems and of capturing the human factor. The model presented in this paper strictly follows the agent based paradigm and employs reactive agents, as detailed in the following paragraph.

4 The Model

The model is built in Java, thus following the Object Oriented philosophy and has been engineered and built at the e-business L@B, University of Turin. This is suitable for agent based modeling, since the individual agents can be seen as objects coming from a prototypal class, interacting among them basing on the internal rules (methods). While the reactive nature of the agents may seem a limitation, it's indeed a way to keep track of the aggregate behavior of a large number of entities acting in the same system at the same time. All the numerical parameters can be decided at the beginning of each simulation (e.g.: number of enterprises, and so on). Everything in the model is seen as an agent; thus we have three kinds of agents: Environment, Enterprises and Emissaries (E^3). This is done since each of them, even the environment, is endowed with some actions to perform.

4.1 Heat Metaphor and the Agents

In order to represent the advantage of an enterprise in owning different competences, the "heat" metaphor is introduced. In agent based models for Economics, the metaphor based approach [19] is an established way of representing real phenomena through computational and physical metaphors. In this case, a quantum of heat is assigned for each competence at each simulation turn. If the competence is internal (i.e.: developed by the enterprise) this value is higher. If the competence is external (i.e.: borrowed from another enterprise) this value is lower. This is realistic, since in the model we don't have any form of variable cost for competencies, and thus an internal competence is rewarded more. Heat is thus a metaphor not only for the profit that an enterprise can derive from owning many competences, but also for the managing and synergic part (e.g.: economy of scale).

Heat is also expendable in the process of creating new internal competences (exploitation) and of looking for partner with whom to share them in exchange of external competences (exploration). At each time-step, a part of the heat is scattered (this can be regarded as a set of costs for the enterprise). If the individual heat gets under a threshold, the enterprise ceases its activity and disappears from the environment. At an aggregate level, average environmental heat is a good and synthetic measure to monitor the state of the system.

The *Environment* is a meta-agent, representing the environment in which the proper agents act. It's considered an agent itself, since it can perform some actions on the others and on the heat. It features the following properties: a grid (X,Y), i.e.: a lattice in the form of a matrix, containing cells; a dispersion value, i.e.: a real number used to calculate the dissipated heat at each step; the heat threshold under which an enterprise ceases; a value defining the infrastructure level and quality; a threshold over which new enterprises are introduced; a function polling the average heat (of the whole grid). The environment affects the heat dispersion over the grid and, based on the parameter described above, allows new enterprises to join the world.

The *Enterprise* is the most important and central type of agent in the model. Its behavior is based on the reactive paradigm, i.e.: stimulus-reaction. The goal for these agents is that of surviving in the environment (i.e.: never go under the minimum allowed heat threshold). They are endowed with a heat level (energy) that will be consumed when performing actions. They feature a unique ID, a coordinate system (to track their position on the lattice), and a real number identifying the heat they own. The most important feature of the enterprise agent is a matrix identifying which competences (processes) it can dispose of. In the first row, each position of the vector identifies a specific competence, and is equal to 1, if disposed of, or to 0 if lacking. A second row is used to identify internal competences or outsourced ones (in that case, the ID of the lender is memorized). A third row is used to store a value to identify the owned competences developed after a phase of internal exploration, to distinguish them from those possessed from the beginning. Besides, an enterprise can be "settled", or "not settled", meaning that it joined the world, but is still looking for the best position on the territory through its emissary. The enterprise features a wired original behavior: internally or externally explorative. This is the default behavior, the one with which an enterprise is born, but it can be changed under certain circumstances. This means that an enterprise can be naturally oriented to internal explorative strategy (preferring to develop new processes internally), but can act the opposite way, if it considers it can be more convenient. Of course, the externally explorative enterprises have a different bias from internally explorative ones, when deciding what strategy to actually take.

Finally, the enterprise keeps track of its collaborators (i.e.: the list of enterprise with whom it is exchanging competencies and making synergies) and has a parameters defining the minimum number of competencies it expects to find, in order to form a joint. The main goal for each enterprise is that of acquiring competences, both through internal (e.g.: research and development) and external exploration (e.g.: forming new links with other enterprises). The enterprises are rewarded with heat based on the number of competences they possess (different, parameterized weights for internal or external ones), that is spread in the surrounding territory, thus slowly evaporating, and is used for internal and external exploration tasks.

The *Emissaries* are agents that strictly belong to the enterprises, and are to be seen as probes able to move on the territory and detect information about it. They are used in two different situation: 1) if the enterprise is not settled yet (just appeared on the territory) it's sent out to find the best place where to settle. 2) if the enterprise is settled and chooses to explore externally, an emissary is sent out to find the best possible partners. In both cases, the emissary, that has a field of vision limited to the surrounding 8 cells, probes the territory for heat and moves following the hottest cells. When it finds an enterprise in a cell, it probes its competencies and compares them to those possessed by its chief enterprise verifying if these are a good complement (according to the parameter described in the previous section). In the first case, the enterprise is settled in a cell which is near the best enterprise found during the movement. In the second case, the enterprise asks the best found for collaboration).

While moving, the emissary consumes a quantum of heat, that is directly dependant on the quality of infrastructures of the environment.

The movement of the emissaries is based on reactive rules; it follows the hotter cells it meets on its path and, if an enterprise is found, it checks for the complementary competences, in order to propose a link with the parent enterprise.

In the following paragraph a formal insight of the model is given through a set of defining equations, for the agents and the general rules.

5 Underlying Formal Equations

In order to formally describe the model, a set of equations is described in the following.

The multi agent system at time T is defined as:

$$MAS_T = \langle \bar{E}, \bar{e}, \bar{\varepsilon}, \overline{\text{link}} \rangle . \quad (1)$$

Where \bar{E} represents the environment and is formed by a grid $n * m$, and a set \bar{k} :

$$\left\{ \begin{array}{l} \bar{E} = \langle n * m, \bar{k} \rangle \\ n, m > 0 \end{array} \right. . \quad (2)$$

Where the set \bar{k} defines the heat for each cell, \bar{e} is the set of enterprises with coordinates on the grid, and $\bar{\varepsilon}$ is the set of the emissaries, also scattered on the grid:

$$\left\{ \begin{array}{l} \bar{k} = \langle k_{i,j} \rangle \\ \bar{e} = \langle e_{i',j'} \rangle \\ \bar{\varepsilon} = \langle \varepsilon_{i'',j''} \rangle \\ 0 < i, i', i'' \leq n \\ 0 < j, j', j'' \leq m \end{array} \right. . \quad (3)$$

Each enterprise is composed by a vector \vec{c} , and an emissary (ε_e). The vector \vec{c} defines the owned competences, with a length L and competences C_l represented by a boolean variable (where 1 means that the l^{th} competence is owned, while 0 means that it's lacking):

$$\begin{cases} e_{i,j} \ni \vec{c}, \varepsilon_e \\ \vec{c} = (L, C_l) \\ 0 \leq l \leq L \\ C_l = \text{Boolean} \end{cases} \quad (4)$$

In $T = t > 0$, $k_{i,j}$ that's the heat of each cell on the grid, depends on the heat produced by the enterprises (K_e) and the dispersion effect (d). The heat of each enterprise is function of the competences it possesses and of the behavior it carried on in the last turns (b_e).

$$\begin{cases} k_{i,j} = f(K_e, d) \\ K_e = f(\vec{c}_e, b_e) \\ b \in \bar{b} \\ \bar{b} = \langle \text{set of behaviors} \rangle \end{cases} \quad (5)$$

In particular, a certain behavior can be successful, meaning that at the end of a phase of internal or external exploration, a new competence (internal or outsourced, respectively) will be possessed. Otherwise, a it's unsuccessful when, after some steps of research and development (internal exploration) or external market research to find a partner, nothing new is found, and thus the l^{th} competence remains zero.

$$\begin{cases} \text{if } (b = \text{success}) \text{ then } C_l = 1 \\ \text{else } C_l = 0 \\ b \in \bar{b} \end{cases} \quad (6)$$

At each time-step the set of links (connecting two enterprises together) is updated basing on the competences of the enterprises.

$$\begin{cases} \overline{\text{link}} = \langle \text{link}(e_{i,j}, e_{i',j'}) \rangle \\ \text{link}(e_{i,j}, e_{i',j'}) = f(\overrightarrow{c_{e_{i,j}}}, \overrightarrow{c_{e_{i',j'}}}) \end{cases} \quad (7)$$

Specifically, when an enterprise does external exploration, it looks for a good partner, i.e.: an enterprise with a number of competences to share. So, if an enterprise with a vector like $\boxed{1\ 0\ 0\ 0\ 1}$ meets one with a vector like $\boxed{0\ 1\ 1\ 1\ 0}$ then there is a perfect match and the two enterprises will create a link among them, to share the reciprocally missing competences. This is the perfect situation, but not the only one in which two enterprise can create a link; in fact, it's enough that there is at least one competence to reciprocally share. The strength of the link is directly proportional to the exchanged competences. This set of equations and rules is enough to explore the effects on the network of the behaviors of the enterprises, namely the way in which the firms are managed (externally or internally focused). Though the model allows also to explore the effects on innovation (i.e.: a competence that's possessed only by one enterprise).

In $T = t' > t$ a radical innovation can be metaphorically introduced in the system (this is called "shock mode", since this is decided by the user, at an arbitrary step) by means of increasing the length of the vector of competences of a specific enterprise:

$$\left\{ \begin{array}{l} L \leftarrow L + 1 \\ C_{i+1}(\bar{e}) = 1 \\ C_{i+1}(\bar{e} - \underline{e}) = 0 \end{array} \right. \quad (8)$$

Meaning that the competence C_{i+1} will be possessed by only one enterprise, at that time, while the same competence will be lacking to all the others; though, all the enterprises' vectors will increase in length, meaning that potentially all of them will be able to internally develop that new competence through R&D, from then on.

The vector length metaphorically represents the complexity of the sector (industry) in which the enterprises operate; an highly technological sector has many more potential competences than a non-technological one. So, another kind of "shock effect" to the system is that of increasing the length of the vector by more than one component, and by leaving all the new components to zero for all the enterprises. In this way, they'll have to develop themselves the new competences by means of internal exploration. The analysis phase is carried on after several steps after t' , in order to see how the introduction of the innovation impacted the network and the enterprise in which the innovation was first introduced. So we have an analysis phase in $T = t'' > t'$ defined as:

$$\left\{ \begin{array}{l} \text{MAS}_{t'} \text{ vs } \text{MAS}_{t''} \\ l \rightarrow d\theta \text{ link}; d\theta e; d\theta k \end{array} \right. \quad (9)$$

Namely, the comparison among the system at time t' and the same system at time t'' , since the innovation has differential effects on the number (and nature) of the links, on the number of enterprises and the heat of the cells composing the environment, always depending on the managerial behavior of the involved enterprises. At the beginning of a simulation, the user can change the core parameters, in order to create a particular scenario to study and analyze.

6 Conclusion and Qualitative Results

The impact of innovation diffusion on the network depends on the collaboration degree of the system. If the network is collaborative the diffusion of innovation strengthens the ties and increases the number of the links among organizations. The firms are more inclined to exchange competences than to create them inside the organization: they favor an externally explorative behavior that obviously strengthens the network. In order to study the complex social dynamics and interrelations among innovation diffusion and collaborative/non-collaborative networks, an agent based model is introduced in this work and described in details. Even if beyond the purpose of the present work, some qualitative results coming from the simulator are given here, in order to show that this model can be effectively used as a tool for studying the dynamics of different base scenarios. As shown in figure 1 and figure 2, where some output graphs obtained from the E^3 simulation model are depicted, a collaborative network (A1) is defined by the existence of a large number of strong ties (compared to the number of enterprises). In our example, there are 10 strong ties among the enterprises. In a network structured in this way, the introduction and consequent diffusion of an innovation strengthens the collaborations through:

- An higher number of ties
- Ties that get even stronger (A2). In particular, the existing links get stronger and new ties are created ex novo.

In this case, the “shock effect” described in the previous paragraph introduces effects in the networks that affect the decree of collaboration of the network itself. The introduction of an innovation in the network strengthens the links among the enterprises and the collaboration efficiency increases.

On the other side, in the case of a network with low propensity to collaboration the strong links do not exist or are a few when compared to the number of enterprises. The introduction of innovation in a network structured in this way can affect the degree of collaboration of the enterprises, according to industry complexity. In this situation (B1), it’s possible to notice two different scenarios. If industry complexity is not too high (e.g.: the textile industry), as represented in B2, the number of ties is low and the firms prefer to create innovation inside the organization than receiving it from other organizations: in this case the firms favor internal exploration. So, when the complexity is low, the propensity to collaboration does not change and the enterprises are still loosely connected. The number of links could even increase, but much more slowly compared to the case of a collaborative network (B2 vs A2).

If industry complexity is high (B3), the diffusion of innovation increases the number of ties (but less than in a collaborative network) but the structure of ties is weak: in this case, again, the firms prefer an externally explorative behavior. So, in this case, the propensity to collaborate gets higher than before after the introduction of an innovation, but the links are always weaker when compared to the case of a collaborative network (B3 vs A2).

The analysis carried on through an agent based model allow to study “in the lab” a social system, like an enterprise network, and to study the effects of an innovation on collaborative and non-collaborative networks. While the purpose of this work is the description of the model itself, the qualitative results show that the innovation diffusion in a network can create new ties among the enterprises (can thus be regarded as a driver for ties creation in a network). Though, only in a collaborative network, or in a non-collaborative network acting in a complex industry, the number of the links increases significantly, while in non-collaborative networks acting in an industry which is not too complex, the number of links among the enterprises stays more or less the same, even after the introduction of the innovation (the enterprises being more focused on internal explorative behavior).

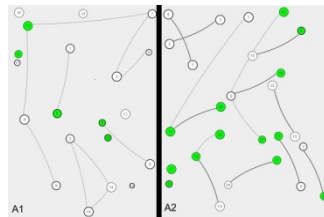


Fig. 1. Collaborative network before (A1) and after (A2) the introduction of an innovation

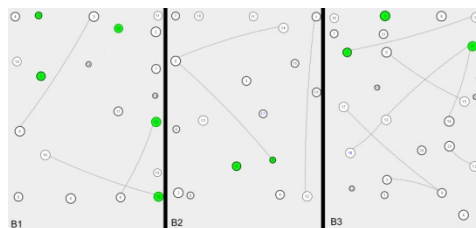


Fig. 2. Non-collaborative network before (B1) the introduction of an innovation. After (B2) in case of non complex industry, and after (B3) in case of complex industry.

The presented model is comprehensive and its scope is wide; it could be used to study the behavior of enterprises clusters and networks in many different scenarios and situations. In future works quantitative results will be given, and different situations will be analyzed.

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147, 195–197 (1981)
2. May, P., Ehrlich, H.-C., Steinke, T.: ZIB structure prediction pipeline: Composing a complex biological workflow through web services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) *Euro-Par 2006*. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: *The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration*. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov>
7. Almeida, P., Kogut, B.: Localization and knowledge and the mobility of engineers in regional networks. *Management Science* 45(7), 905 (1999)
8. Koka, R.B., Madhavan, R., Prescott, J.E.: The evolution of Inter-firm Network: environmental effect on patterns of network change. *Academy of Management Review* 31(3), 721–737 (2006)
9. Brass, D., Burkhardt, M.: Centrality and power in organizations. In: Nohria, N., Eccles, R. (eds.) *Networks and Organizations*, p. 191. Harvard University Press, Boston (1992)
10. Ahuja, G.: The duality of collaboration: Inducements and opportunities in the formation of interfirm linkages. *Strategic Management Journal* 21(3), 317 (2000)
11. Moreno, J.L.: *Who Shall Survive? Nervous and Mental Disease Publishing Company*, Washington D.C (1934)
12. Burt, R.: *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge (1992)
13. Nadel, S.F.: *The Theory of Social Structure*. Free Press, New York (1957)

14. Mitchell, J.C.: The Concept and Use of Social Networks. In: Clyde Mitchell, J. (ed.) *Social Networks in Urban Situations*. Manchester University Press, Manchester (1969)
15. Coleman, J.: Social capital in the creation of human capital. *American Journal of Sociology*, 94, 95 (1988)
16. Cooke, P.: The new wave of regional innovation networks: Analysis, characteristics and strategy. *Small Business Economics* 8(2), 159 (1996)
17. Franklin, S., Graesser, A.: Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In: *Proceedings of the Agent Theories, Architectures, and Languages Workshop*, pp. 193–206. Springer, Heidelberg (1997)
18. Eisenhardt, K., Schoonhoven, C.: Resource-based view of strategic alliance formation: strategic and social effects in entrepreneurial firms. *Organization Science* 7, 136 (1996)
19. Jennings, N.R.: Software agents. *IEEE Review*, 17–20 (1996)
20. Gemünden, H., Heydebreck, P., Herden, R.: Technological interweavement: A means of achieving innovation success. *R & D Management* 22(4), 359 (1992)
21. Gemünden, H.G., Ritter, T., Heydebreck, P.: Network configuration and innovation success: An empirical analysis in German high-tech industries. *International Journal of Research in Marketing* 13(5), 449 (1996)
22. Grandori, A.: An organizational assessment of interfirm coordination modes. *Organization Studies* 18(6), 897 (1997)
23. Grandori, A., Soda, G.: Inter-firm networks: Antecedents, mechanisms and forms. *Organization Studies* 16(2), 183 (1995)
24. Granovetter, M.: Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91, 481 (1985)
25. Remondino, M.: Agent Based Process Simulation and Metaphors Based Approach for Enterprise and Social Modeling. In: *ABS 4 Proceedings*. SCS Europ. Publish. House (2003)
26. Gulati, R.: Alliances and networks. *Strategic Management Journal*, Special Issue 19(4), 293–317 (1998)
27. Gulati, R.: Network location and learning: the influence of network resources and firm capabilities on alliance formation. *Strategic Management Journal* 20(5), 397–420 (1999)
28. Liebeskind, J., Porter, O., Zucker, L., Brewer, M.: Social networks learning and flexibility: Sourcing scientific knowledge in new biotechnology firms. *Organization Science* 7(4), 428 (1996)
29. Boyd, J.P.: The Algebra of Group Kinship. *Journal of Mathematical Psychology* 6, 139–167 (1969)
30. Woolridge, M., Jennings, N.: Intelligent agents: Theory and practice. *Knowledge Engineering Review* 10(2), 115–152 (1995)
31. Harary, F.: *Graph Theory*. Addison-Wesley, Reading (1969)
32. Phlippen, S., Riccaboni, M.: *Radical Innovation and Network Evolution* (2007)
33. Boyd, J.P.: The Algebra of Group Kinship. *Journal of Mathematical Psychology* 6, 139–167 (1969)
34. Powell, W.W., Koput, K.W., Smith-Doerr, L.: Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly* 41(1), 116 (1996)
35. White, H.C.: *An Anatomy of Kinship*. Prentice-Hall, Englewood Cliffs (1963)
36. Shan, W., Walker, G., Kogut, B.: Interfirm cooperation and startup innovation in the biotechnology industry. *Strategic Management Journal* 15, 387 (1994)

A Novel Application of Organic Plant Farming Analysis System – Using Game Theory and Multi-Agent Technique

Chih-Yao Lo^{1,2} and Yu-Teng Chang^{1,2}

¹ School of management, Huazhong University of Science & Technology

² Department of Information Management, Yu-Da University,

Miaoli County, Taiwan 36143, R.O.C.

{jacklo, cht}@ydu.edu.tw

Abstract. In the recent decade, people have attached increasingly greater importance to healthy diet demands; facilitating rapid growth of organic agriculture. However, methods for organic cultivation have great differences from traditional methods. How to effectively choose appropriate organic crops, for season changes and different farming conditions, is a difficult issue for the majority of planters. In addition, achieving the goal of maximum economic profits for the year must target all crops, and should be evaluated as the basis of priority selection. Taking organic vegetable farming as an example, this research uses knowledge and rule-based methods, while applying a game and multi-agent theory, and develops a set of graphic intellectual decision evaluation mechanisms with ASP.NET and MS-SQL. In the system, a merger of game theory and multi-agent system was tested and verified to provide suggestions that are 84.25% effective, as compared to the suggestions provided by experts. Future studies will apply the theory on other crop planting issues.

Keywords: Multi-Agent, Game Theory, Decision Evaluation System, Organic Vegetables.

1 Introduction

Along with the promotion of national incomes, people's level of living has been gradually enhanced, in addition, environmental protection and consciousness of environments is gaining ground, causing agricultural production methods to disappear. Organic agriculture is in accordance with this kind of market direction, and develops new agricultural methods, and thus, expresses hidden opportunities for organic agriculture of the future [1].

At present, decision evaluation for system applications in agricultural are related to scholarly research and are aimed at plant or vegetables that require the application of fertilizers for plant disease prevention, often the user must have agricultural aspects that are suitable to specialized knowledge, and often result in diagnosis irregularities [2]. Developed science and technology information, in use for transportation in the organic agricultural industry, is fastidious about present knowledge, should be engaged in for

organic agricultural computerization of systems, as the farmer is the quite important for policy-making and the promotion of efficiency as well as various aspects of production of competitive powers [3].

This research aimed to integrate an information and technology decision evaluation system and multi-agent mechanism to construct references for use in organic agriculture, and develop a decision evaluation system for organic vegetable farmers to deal with these issues.

2 Theoretical Background

2.1 Organic Agriculture

“Organic agriculture” is a method of agricultural cultivation that uses minimal or no chemical fertilizers. This is coupled with a crop rotation system that plants green fertilizer crops, while using natural wastes, both in and out of the farm, as well as rock beddings that are rich in plant nutrients, in order to maintain a balance and stable style of growing crops. Although current information regarding organic agriculture is insufficient, farmers who have begun organic cultivation are actively researching related production technology market information updates. However, there are still many problems with circulating information, such as accuracy, practicability, specialties, and integrality.

2.2 Decision Evaluation System

The Decision Evaluation System is a system that uses statistic and knowledge to integrate a useful behavioral model of its users [4]. The purpose is to establish a system, which imitates human behaviors, to provide proposals or information to users in a certain domain. The proposals may include merchandise proposals, personalized merchandise information, and comments on specific groups. For companies, the output of the decision evaluation system may act as a reference. However, to make the application of a decision evaluation system more valuable and practical, the support of the Internet is indispensable. The characteristics of the Internet, such as its ability to provide quick dispersal of information, low cost, universality, diversity, and rich contents of its data, has helped the Internet to become the best channel for spreading agricultural information.

2.3 Bargain Game Theory

The greatest difference between game theory and general policy-decision theory is that, game theory explores the problems faced by a group of policy-decision makers in a specific situation. It has helped economists, politicians, and financial experts to solve many problems as it has the exactitude of a mathematical model and simplified complex interactions in the real world by providing research decision-policy makers with methods of strategic behavior analysis [5-6].

2.4 Multi-Agent System

The so-called Multi-Agent System (MAS) was derived from Distributed Artificial Intelligence (DAI) years ago. The term simply means “a distributed environment

formed by multi-agents.” The primary concept of a multi-agent system is to take advantage of characteristics of cooperation in the division of labor, individual techniques, knowledge, and plans from various types of agents in a multi-agent environment, in which each agent does its part to solve distributed problems and achieve the goal of an integral multi-agent [7-8].

3 Methodology and Steps

As shown in Fig. 1, this study is divided into two parts: suitability decision evaluation, and maximum economical profit decision evaluation. The following will separately explain the methods used in each part.

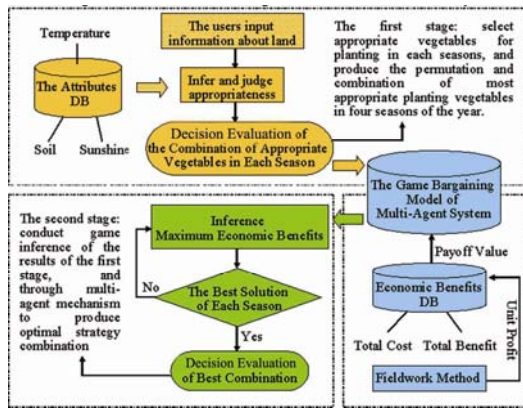


Fig. 1. The Structure of Integral Operation of the System

3.1 Decision Evaluation for Suitability

In this stage, the major task is to decide on suitable crops for planting, and then arrange the order of the crops by suitability. Each step is explained as followed:

A. The Policy-Decision Factors

This study applied the “retro link” method, which establishes a relationship that links backwards from the result. According to the terms for each crop to grow, and to create defining variables close to a professionals’ criteria for crop selection, we change the term of “land suitability” to “crop selection criteria”, as shown in Fig. 2.

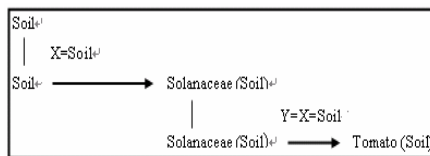


Fig. 2. The Example of Relations of Inference Chain

Due to the complexity of factors that can affect the growth of vegetables, after reviewing past studies [9-11], and with help from the Agricultural Research and Extension Station in Miaoli, Taiwan, and interviews with organic farming experts through field study, this research categorized the environmental factors that may determine the best environment for vegetable growth, and listed the four major factors by their degrees of importance, as shown in Fig. 3.

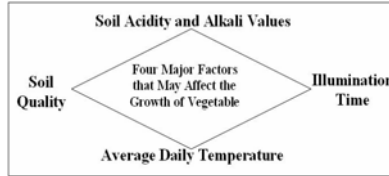


Fig. 3. Four Major Factors that Affect the Growth of Vegetable

B. The Installment of a Knowledge Base

In the development of our knowledge base for the decision evaluation system, this study interviewed 11 organic farmers, who were recommended by Mr. Tsai, Cheng-Hsiang, the section chief of the Agricultural Research and Extension Station, as our domain experts for extracting knowledge. In order to simplify our prototypical study of organic vegetables, present in markets in Taiwan, we selected a fixed list of 20 varieties, and then divided them into four groups representing the four seasons, each with five varieties of vegetables, and no vegetable used in more than two groups, with each of the seasons of an equal three-month duration. The known attributes were discussed to establish repertory grids [12], as shown in Tables 1 and 2.

Table 1. List of Attributes for Judgment of Plant ability

Name of Attribute	Data Type	Description
Soil Acidity and Alkali Values	Integer	Unit: pH (power of Hydrogen ions)
Soil Quality	Integer	1. Loam 2. Clay Loam 3. Sandy Loam 4. Various Soil
Average Temperature	Integer	Unit: Centigrade
Illumination Time	Integer	1. Slight Illumination 2. Semi-Illumination 3. Full Illumination

Table 2. The Grid of Knowledge Extraction about Gherkin

Results	Range Fit for Planting
Changes	
pH Value of Soil	pH5.5~7.2
Quality of Soil	Loam
Average Temperature	20~25
Illumination Time	Slight Illumination

C. Establishing Calculation Formulas for Rule Judgments

According to Rule (1), compare and judge the factors input by the users and each variety of vegetables in knowledge base. All factors of each vegetable must meet the requirements of Rule (1) to enter into the second step. Rule (1) is as follows:

$$\text{Min}(V_{ij}) \leq \text{Input}(j) \leq \text{Max}(V_{ij}) \quad (1)$$

V_{ij} : V is the variable of a specific vegetable, i is the order of the vegetable, while j is the ordered factors.

$\text{Input}(j)$: Input is the variable input by users; j is the factors in order.

The total objective functions for each variety of vegetable that meet the requirements of Rule (1) are obtained. The score of each factor is its sub-objective. The range of each factor is divided into 10 equal parts and compare with input factor values. Scores are given according to the distance to average value (middle value). The farthest one is 5, whereas the nearest is 0. The sub-objective function is regarded as Rule (2). As seen, the purpose of the total objective function is to seek the minimum value of the addition of all sub-objective functions. Rule (2) is shown as follows:

$$\begin{aligned} \text{Total objective function: } S_i &= \text{Min} \sum S_{ij} \\ \text{Sub-objective function: } S_{ij} &= \left[\frac{\left| \text{Input}(j) - \left(\frac{\text{Max}(V_{ij}) + \text{Min}(V_{ij})}{2} \right) \right|}{\left| \frac{\text{Max}(V_{ij}) - \text{Min}(V_{ij})}{10} \right|} \right] \end{aligned} \quad (2)$$

S_i : Total scores of the ith vegetable

S_{ij} : The scores of the jth factor of the ith vegetable

Arrange the order of the total scores of vegetables that are appropriate for each season, from smallest to the greatest. The smaller the scores are, the more appropriate the vegetable will be for the user's farmland, based on the user's own input. Finally, a proposal was reached for a set of permutation and combination of the vegetables of the most appropriate cultivation. The output of this process is compiled into a list of the ideal vegetables to be planted on the user's farmland for each season. According to the knowledge grids and rule calculation formula, we can obtain all 20 rules of the system.

3.2 Decision Evaluation for Maximum Profit

This section discusses the results of part one as the input data for the second part of the study of the bargain game theory. Then, MAS is applied to construct the MAS-game mechanism. The second part of the decision evaluation is determined from analysis of the mechanism. The construction steps are explained, as follows:

A. Formulate the Analytic Structure of Game Agreement

Based on the results of the first stage, we determined the vegetables that are most appropriate for each season, each of which has its own profit, sale status, crop switching

costs, and conflicting conditions. In order to realistically maximizing the profit, the first task in each season is to consider the most suitable crop. Therefore, through game theory, the prioritized crops for maximum profit were identified. The following are briefs on the game participants and the strategies to be applied in their bargaining processes.

Game participants (seasons):

1. Spring; 2. Summer; 3. Autumn; 4. Winter

Strategies: The appropriate vegetables for each season, based on the relevant norms of organic farming, the necessities and costs involved in field conversions between seasons, and the unequal varieties of vegetable cultivation.

1. Spring (spring onion, snap bean, watermelon, balsam pear, and spinach)
2. Summer (gherkin, sweet potato, pumpkin, eggplant, and sponge cucumbers)
3. Autumn (lappa, sweet corn, Chinese cabbage, broccoli, and onions)
4. Winter (rape, celery, potato, turnip, and radishes)

Limitations: (conditions)

Under the regulations of organic farming, one crop should not be planted in two consecutive seasons.

The goal of game bargaining:

In organic farming's rules, it is specified that the same crop is never planted in the same field two seasons consecutively. It is noted that the conversion of the field from one crop to another would inevitably incur "transition costs" that must be accounted for when figuring net profit.

The steps of game operation are as follows:

Step 1) Select two objectives (A and B represent two consecutive seasons out of a total of four seasons), and list the appropriate vegetables for planting in both seasons. At this point, the payoff matrix does not include any values.

Step 2) Determine the sales weight values, using the Miaoli region as an example, it is worth noting the difficulties of distributing the organic vegetables and the timeliness of delivery costs. Organic farmers in Taiwan typically send at least a portion of their organic vegetables to their local organic vegetable distribution center; the amounts sent are based on the figures determined by the farmers' union for that area of Taiwan and the sales channels there.

Step 3) In the first phase, along with the results of field surveys, each of the season's appropriate vegetables should have relatively flat profits, then the level of sales value of the weights and the units suitable for growing vegetables for profit, which correspond to sales of the vegetable weights. The pros and cons of the influencing amendments are determined, and the cultivation and conversion costs are deducted to achieve the real value of the profits according to the following computation formula:

Original Profit Value \times Weighted Number of Sales Value $-$ conversion cost = Profit Value after Modification

Step 4) The value modified in Step 3 is considered the payoff value used in the game payoff matrix; the next step is to input the payoff values of all strategy combinations, according to the corresponding columns of the matrix, to generate a payoff matrix of a 2-player multi-strategy, with payoff values (see Table 3). Based on the results, this study determined the dominant strategy of the matrix. If a dominant strategy exists, the matrix is simplified first before proceeding further.

Table 3. Payoff Matrix

		B				
		B-1	B-2	B-3	B-4	B-5
A	A-1	(Pa1,Pb1)	(Pa2,Pb2)	(Pa3,Pb3)	(Pa4,Pb4)	(Pa5,Pb5)
	A-2	(Pa6,Pb6)	(Pa7,Pb7)	(Pa8,Pb8)	(Pa9,Pb9)	(Pa10,Pb10)
	A-3	(Pa11,Pb11)	(Pa12,Pb12)	(Pa13,Pb13)	(Pa14,Pb14)	(Pa15,Pb15)
	A-4	(Pa16,Pb16)	(Pa17,Pb17)	(Pa18,Pb18)	(Pa19,Pb19)	(Pa20,Pb20)
	A-5	(Pa21,Pb21)	(Pa22,Pb22)	(Pa23,Pb23)	(Pa24,Pb24)	(Pa25,Pb25)

Step 5) Using the maximum payoff value of Objective A as a starting point, confirm whether Objective B may receive the maximum profit if adopting the responding strategy, as Objective A chooses a certain strategy of the maximum payoff value. If so, the strategy combination is an equilibrium solution; if not, then Objective B responds with the other strategy that could obtain a better payoff; then Objective A adopts the same action; thus the actions continue until an equilibrium solution is discovered.

Step 6) The convergence situations of the search are as follows:

A single equilibrium solution (converge to an intersection) is a 1×1 matrix; in this situation, the optimal equilibrium solution is the output of game bargains.

Approximate equilibrium solution (a situation in which one party’s strategy option converges to a single strategy): This is a 1×N or N×1 matrix; in this situation, we only need to consider the responding strategy option of a higher payoff value, which an undecided player may select. For a player who has already decided, adopting the strategy will definitely secure a higher payoff.

Double equilibrium solutions (evaluate solution by mixed strategies): This is a 2×2 matrix; this situation is a symmetrical equilibrium that produces a feasible strategic combination by employing a mixture of strategies.

No equilibrium solution (seek for solution by mixed strategy): It is a 2×2 matrix; in this situation, in order to form a cycle, an equilibrium strategy combination that secures the maximum expected payoff may be obtained by employing a mixture of strategies.

Step 7) All objectives conduct mutual game actions and integrate the results. Only four games are held in this research, which are: spring, summer, autumn, and winter. Thus, in this manner, we solve the problems of having no straight planting and meet the requirements of the considering maximum profit. First, we have to judge the situation of

our final decision. The optimal situation involves reviewing the statistics for all objectives, when only one strategic option remains, the one with the highest probability of satisfying all parties is selected. At this point, take the strategy that has the highest probability value as the final equilibrium solution. If one or more roles in the game have several strategy options, which all have the highest probability values, reserve all strategy options for later consideration; then select the strategy that has the least effect on all players (in which all parties receive minimal negative effects, but the highest expected profit) to be the final strategy.

B. The concept of Multi-Agent Implementation for Conflict Resolutions

This research adopts the concept of a multi-agent to conduct our game, in which the characteristics of a multi-agent help players to cooperate and complete their tasks and achieve their final objective, which enhances the implementation efficiency of coordination for the entire game. The processes of operation for game coordination are shown in Fig. 4. When the bargains fail, both parties should consult again and secure an internal agreement before proceeding to the next negotiation.

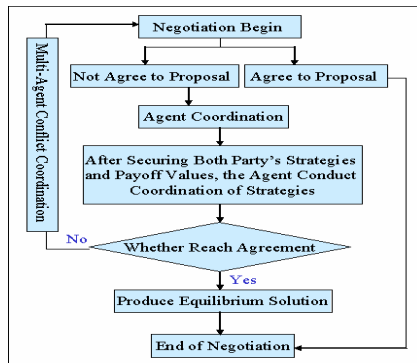


Fig. 4. The Coordination and Operation

In the research structure, for the convenience of analyses, we assume Seasons A and B are the two players in a 2-player multi-strategy game. Consider Season A as Agent A, which has five strategies. In the same way, we could consider Season B as Agent B, with its own five strategies. Both positions of negotiation will be affected by profit factors. Because both parties have high expectations for a profitable outcome to the negotiations, each party may present conditions that are far from what the other party expects. Nevertheless, the premise is that they are willing to negotiate for the best game equilibrium achieved so far. Therefore, we will need another agent to act as coordinator (the third party), who is called the conflict resolution agent, designated in this negotiation example as Agent C. Agents A and B will share a common database, but each has its own private knowledge base, which the other agent is not privy to.

The structure of the multi-agent is shown in Fig. 5.

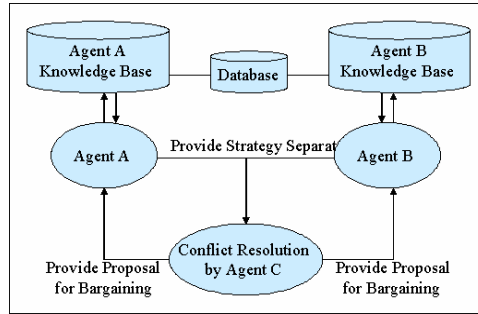


Fig. 5. The Structure of Multi-Agent System

C. The Game Bargaining Model of a Multi-Agent System

Based on the above-constructed game of analytical structure and multi-agent conflict coordination, we can build a multi-agent game bargain model for research. The entire structure of the model is shown in Fig. 6.

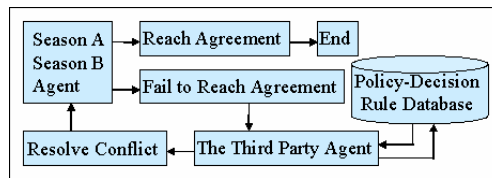


Fig. 6. The Game Bargaining Model of Multi-Agent System

The steps of the game bargaining model of a multi-agent system are as follows:

Step 1) Design three agent mechanisms – a mechanism of mediation with the season A agent, the season B agent, and a third party agent representing, respectively, the first two parties and the third party.

Step 2) Discover the strategy reward values under the limits of individual demands from the two parties.

Step 3) The season A agent and the season B agent simultaneously convey their strategy reward values to the third party agent.

Step 4) The third party agent judges whether the strategy reward values received are conflicting.

Step 5) If they are conflicting, the third party selects a resolution rule from the policy-decision rule database to implement conflict resolution and returns them to the two-party agent mechanisms – the season A agent and the season B agent – for judgment to determine whether they satisfy their individual demands.

Step 6) If the results from conflict resolution implemented by the third party agent cannot satisfy the individual demands of any agent, then another resolution rule from

the policy-decision rule database must be selected to implement conflict resolution and produce a new result.

The process is repeated until a result satisfies the demands of both agents.

5. Constructing an online decision evaluation system

This system provides decision evaluation in two steps, one toward suitable organic crops, and a second toward the maximum economic profit. Through the convenience of the internet, users are provided an instant decision evaluation service, as shown in Fig. 7.

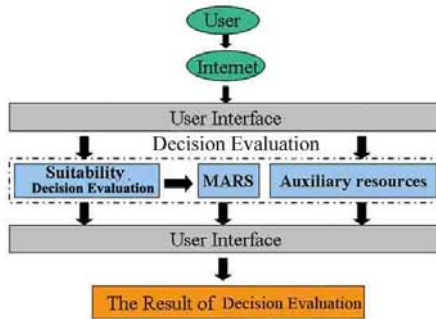


Fig. 7. The Structure of the Intelligent Decision Evaluation System

4 Test and Verification

4.1 Systematic Test

The systematic test subjects are provided by the Section Chief of Miaoli Agricultural Research and Extension Station, Mr. Cheng-Hsiang Tsai. The subjects include eleven organic crop farmers, and the conditions of the lands are recorded as data for the research. The test results are shown as Fig. 8.



Fig. 8. Operating System Decision Evaluation Screen

4.2 Systematic Verification

To verify the value and accuracy, this study compared the systematic output with professionals in the Taiwan Outstanding Agriculturists Association. The data from the

eleven farmers' interviews were quantified and organized into questionnaires for obtaining suggestions from one hundred professionals', based on suitability and maximum profit. After organizing and analyzing the data, and comparing the results with the systematic output, it was concluded that the output agrees with the greater portion of the professionals' suggestions. The results have an 84.25% level of agreement. Therefore, it was confirmed that the decision evaluation system has the accuracy of a professional level.

5 Conclusions

The contributions and results of this research are as follows:

1. Combining the decision evaluation system with a multi-agent mechanism: This study integrated the two techniques and applied them to the installment of decision evaluation tools for choosing organic vegetables. In the first stage, we adopted a knowledge base and rules in the decision evaluation system for judgment, which was followed by analyses reached through the game bargaining model of a multi-agent system in the second stage. The primary purpose was to break away from the now-popular decision evaluation system of simple evaluation and policy decision-making. In addition to the knowledge base of domain knowledge, the new system also includes a new model that can provide users, faced with several possible choices, more objective and effective decision evaluation.
2. Combining the model of the game theory: We transformed the game theory of economics into a model and applied it into the decision evaluation system, and then, further applied it to organic planting fields. That is, the system used different viewpoints to analyze and judge models, which considered a combination of theory and practice, and represents an innovative way of thinking about related academic researches.
3. Establishing a knowledge base for making judgments on organic planting: In mass researches of related books and interviews with domain experts, we retrieved knowledge, based on a knowledge table, to construct a knowledge base of choices and judgments regarding organic planting. The knowledge base can also provide people who are responsible for planning and designing the knowledge base with reference that will help them to make later policy decisions. The system has shared resources through webpages, and the installment of the website also expands the sources of knowledge retrieval. Meanwhile, due to the understanding that the information stored in any knowledge base may require occasional updates, the system has considered the question of periodic renewal, including ways to expand or rule out some of the rules in order to retain and maintain the applicability and convenience of the system.

References

1. Hammer, G.L., Hansenb, J.W., Phillipsb, J.G., Mjeldec, J.W., Hillc, H., Lovec, A., Potgieter, A.: Advances in application of climate prediction in agriculture. *Agricultural Systems* 70(2-3), 515–553 (2001)

2. Omamo, S.W., Lynam, J.K.: Agricultural science and technology policy in Africa. *Research Policy* 32(9), 1681–1694 (2003)
3. Leftwich, G.M.: Science and the humanities: the case for state humanities councils. *Technology in Society* 24(4), 523–530 (2002)
4. Resnick, P., Varian, H.R.: Recommender systems. *Communication of ACM* 40(3), 56–58 (1997)
5. Cheng, C.H., Soo, V.W.: Multi-Agent Recommendation Systems Based on Group Preferences. In: *Proc. of PRIMA* (2003)
6. Rasmuse, E., Young, J.Y., Chang, J.Y., Wu, L.H.: *Game Theory and Information Economics*. Wu Nan Books, Reading, R.O.C. (2003)
7. Bradshaw, J.M.: An Introduction to Software Agents. In: *Software Agents*, Menlo Park, pp. 3–46 (1997)
8. Jennings, N.R., Sycara, K., Wooldridge, M.: A Roadmap of Agent Research and Development. *Autonomous Agents and Multi-Agent Systems Journal* 1(1), 7–38 (1998)
9. Yu, C.Q., Lee, M.G.: *Organic Vegetables DIY*. Gardening World Publishing, Reading, R.O.C. (2001)
10. Araki, M., translated by Yeh, S.W.: *Graphic Vegetable Cultivation Skills*. Wen Guo Books, Reading, R.O.C. (2003)
11. Itagi, T., translated by Chou, L.L.: *Encyclopedia for Family Vegetable Garden*. Morning Star Publishing, Reading, R.O.C. (2003)
12. Tseng, H.H., Huang, G.C.: *Artificial Intelligence and Expert System*. Chi-Biao Publishing, Reading, R.O.C. (2005)

A Dynamic Trust Network for Autonomy-Oriented Partner Finding

Hongjun Qiu¹, Jiming Liu^{1,2}, and Ning Zhong^{1,3}

¹ International WIC Institute, Beijing University of Technology, Beijing, 100124
jiming@comp.hkbu.edu.hk

² Department of Computer Science, Hong Kong Baptist University,
Kowloon Tong, HK

³ Department of Life Science and Informatics, Maebashi Institute of Technology,
Maebashi, Japan

Abstract. The problem of finding partners is to identify which entities (agents) can provide requested services from a group of entities. It can be found in open and distributed environments for such tasks as file sharing and resource allocation. Previous studies have shown that entities can refine and determine partners through measuring trust relationships, i.e., the beliefs of entities that others will accomplish a request for assigned services at hand. Entities dynamically change their beliefs through recalling their past experiences in order to quickly identify partners for new requests. This paper aims to observe whether those changes can enable entities efficiently find partners and hence provide services. To this end, we propose a dynamic network of trust-based entities. Then, we investigate the dynamics of its structure and efficiency in the above-mentioned aspects. Autonomy-Oriented Computing (AOC) is applied to observe how the dynamics emerge from local behaviors. A notion of autonomy is embodied in defining how entities activate their partner finding behaviors, whereas self-organization is realized to update the strength of trust relationships. Experimental results explicitly display a dynamic process of this network, changing from containing no link (trust relationship) to having some stable links. Specially, in this process, the efficiency gradually gets enhanced.

1 Introduction

In a multi-entity network, entities (e.g., companies) usually look for some others to provide services beyond their own limited abilities, i.e., finding partners. The partners determine whether and how entities can successfully provide services. For example, a future online-library network will be an open, distributed environment. People can visit any library in this network to request digital copies of papers or other digital resources in real time. If a digital library only subscribes post-1980 papers and is requested to provide a paper published in 1965, it can ask another library to immediately send the requested paper.

Currently, researchers have introduced a notion of trust relationship as a measurement for identifying partners, to share files or solve problems in open, distributed and dynamic environments [1]. Trust relationships refer to entities' beliefs

that others will provide services as needed [2]. In most cases, entities are inclined to choose partners from the ones in which their beliefs are relatively strong. Entities weaken their beliefs if the found partners cannot provide services. Otherwise, they strengthen their beliefs. They dynamically change the strength values of their beliefs with more and more such experiences. Thus, they can quickly identify partners at any time.

Many studies have been done in applying trust relationships to find partners. Golbeck [1] argues that trust relationships can be used to identify who can produce trustworthy online information. Sen et al. [3] compare three schemes of computing trust relationships in reducing the cost of finding partners from a small group of entities. Specially, they introduce a nearest neighbor algorithm as a scheme through recalling a fixed number of latest experiences of finding partners. To get rid of malicious entities from the group of potential partners, Kamvar et al. [4] build a matrix of weighted trust relationships through aggregating and normalizing based on a history of uploading files. They compute the left principal eigenvector of this matrix as their metric. More related work can be found in [5,6,7]. In such work, entities either measure the probabilities that others have successfully provided services before, or ask past partners for their beliefs about third-party ones.

However, it remains to clarify why and how trust relationships can make entities efficiently find partners in dynamic, distributed, real-world networks. This paper aims to solve this problem through proposing a dynamic trust network based on the bottom-up methodology of Autonomy-Oriented Computing (AOC). AOC can help us characterize how network dynamics emerge from local interactions through utilizing the ideas of autonomy and self-organization [8,9,10]. Network dynamics include the dynamics of network structure, network performances in providing services and so forth. Specially, Zhang et al. [11] explore such dynamics of a dynamic, distributed network for providing services. Their work shows the feasibility and rationality for the work in this paper.

In this study, the autonomy is embodied for entities (nodes) to find partners while a process of self-organization is realized to update trust relationships (links). Entities decide to activate which behaviors for partner search or selection through measuring their abilities and beliefs. They also immediately update their beliefs, once they receive the feedback from their latest partners. Specially, a positive feedback mechanism is emphasized, i.e., the strong beliefs become stronger while the weak become weaker. Experimental results show that our network quickly converges to be scale-free. With such a structure, entities quickly find partners for any request. These results explain why dynamic trust relationships are helpful for entities to find partners.

The reminder of this paper is organized as follows: Section 2 gives a detailed problem statement. Section 3 formulates a network on the basis of AOC. The network contains autonomous entities as nodes and self-organized trust relationships as links. In Section 4, we observe the structural emergence of this network and the dynamics of trust relationships. The efficiency in finding partners is also measured. Finally, the paper is concluded in Section 5.

2 Problem Statements

As mentioned above, the goal of this study is to examine the effects of dynamic trust relationships on finding partners in open, dynamic, and distributed networks, such as the Internet. It requires us to answer the following questions:

- How do entities update their trust relationships? What will they do if they never interact with their newly-found partners before? Will entities memorize the information about all of their past partners?
- What are the mechanisms for entities search and select partners? How do entities refine partners based on updated beliefs? How do entities identify their partners from their refinement results?
- Which parameters can we introduce to characterize the change of trust relationships, besides the variation of their strength? Is it enough to count the number of entities' trustees, associated by their beliefs?
- What is the efficiency in finding partners based on dynamically-changing trust relationships?

Previous studies have confirmed two preliminary phenomena. One is that entities prefer to interact with their trustees rather than strangers. The other is that entities strengthen their beliefs about another entity once it accomplishes a service request and vice versa. Based on the two phenomena, we attempt to answer the above questions in the following scenario.

Here, we assume that there exists a virtual network, which contains trust relationships as links and entities as nodes. In this network, entities are assigned with fixed abilities of providing certain services. Each of them can find a partner to satisfy a whole service request, which cannot be finished by itself. The request has a cost limit, i.e., the maximum number of times for finding partners. The partner will honestly inform whether it can finish this request. Then the finder can decide to strength or weaken its belief about this partner. If this partner cannot finish, it will transfer the request to the third entity, which is discovered as its partner. In other words, this request propagates among entities until it is accomplished or the number of finding times reaches the given cost limit.

Accordingly, the above questions can be translated into the following tasks:

1. Modeling a virtual network, in which distributed entities can find partners and update their beliefs.
2. Measuring structural characteristics at different moments to display the dynamics of the links among entities. The characteristics include the clustering coefficient and the harmonic mean of average path length (APL^{-1}).
3. Examining the efficiency in finding partners with two parameters, i.e.,
 - Accomplishment ratio: the ratio of the number of accomplished requests to the total number of requests;
 - Average propagation step: the average times that entities find partners for finishing a request.

3 Modeling a Dynamic Trust Network

3.1 Basic Ideas

In this section, we will model a virtual network based on the methodology of AOC, i.e., a dynamic trust network. In this network, trust relationships are self-organized as entities continually find partners in an autonomous manner.

For finding partners, entities activate one of well-defined search or selection behaviors in a probabilistic manner. The probability of activating each behavior is adaptive to new requests and entities' states (e.g., their beliefs).

A process of self-organization is realized to update trust relationships. At first, there is no trust relationship in the network. After a period of continually finding partners, entities maintain several trust relationships. The realization of this process lies in two aspects. On one side, entities' behaviors are defined to be exploratory or even stochastic. They may find strangers as their partners and create their beliefs about the newly-found partners. However, with positive feedback mechanisms, the newly-created beliefs may be weakened or even eliminated later. Therefore, only a part of generated relationships can remain. The remaining relationships can help entities effectively find partners to finish new requests. So, any request can be accomplished with a higher probability.

The network, entities and service requests are defined as follows:

Definition 1. Let $G'(t) = \langle E, L'(t) \rangle$ denote a dynamic network of trust-based entities on the basis of AOC. For brevity, we call it a dynamic trust network. $E = \{e_1, \dots, e_{N_e}\}$ denotes the set of entities and N_e is the number of entities in the network. $L'(t) = \{l'_{ij}(t) | e_i, e_j \in E\}$ is the set of trust relationships at time t .

Definition 2. Let $e_i = \langle ID, ability, rules \rangle$ denote an entity where ID is its identifier, i.e., $e_i.ID = i, i \in [1, N_e]$. It can provide some services with different performances, i.e., $e_i.ability = \langle cw_1, \dots, cw_{N_{service}} \rangle$, where cw_j denotes the performance on the j th type of service t_j . $N_{service}$ is the number of service types. The rules define when and how entities activate their behaviors.

Definition 3. A request can be formulated as $r_m = \langle rw_1, \dots, rw_{N_{service}}, T_{max} \rangle$ where rw_j denotes the need of the type of service t_j . T_{max} is the maximum number of times that entities are allowed to find partners.

The two ideas from AOC are illustrated through describing the process of finding partners for finishing a request, which is specified in Section 3.5. Once a new request r_m is submitted to an entity e_i , this entity will perform the following:

1. **Evaluating.** Firstly, the entity will determine whether it can solely accomplish the new request by means of matching its ability with this request using a cosine-based similarity function $e_i.simRE(r_m)$. The request can be considered as accomplished once the value of $e_i.simRE(r_m)$ is larger than a threshold. The evaluation functions are specified in Section 3.4.
2. **Partner Search and Selection.** If the request cannot be accomplished and its T_{max} is not reached, the entity e_i will start finding partners. Firstly, it will

search some candidates by means of activating one of search behaviors in a probabilistic manner. Then, it will select a candidate as its partner through activating one of selection behaviors also probabilistically. The probabilities are adaptive to the request and the states of the entity, e.g., its trust relationships. So, trust relationships will explicitly affect entities' partner finding. The detailed behaviors are given in Section 3.2.

3. **Updating.** The entity e_i will change its states once a partner is found. Firstly, it will deliver the whole request to its partner. Then, the partner e_j will honestly feedback its evaluating results $e_j.simRE(r_m)$. Finally, this entity will generate the trust relationship l'_{ij} , if $l'_{ij}(t)$ is not in the network. Otherwise, it will strengthen the relationship if the ability of this partner is relatively closer to this request, or weaken the trust relationship if not. The mechanisms of trust relationships are specified in Section 3.3.

3.2 The Local Autonomy of Entities

Three search behaviors and two selection behaviors are defined in this section. They will be probabilistically activated to realize the autonomy of entities. The probability of activating each behavior is determined with the degree of similarity between requests and entities' abilities, i.e., $e_i.simRE(r_m)$. When the degree is high, entities' neighbors will be discovered as partners with a high probability. Stochastic behaviors are given for entities to avoid being trapped in local-optima and for newcomers to join this network.

- **Neighbor-based search.** An entity e_i will find some neighbors, which are not involved in the current request r_m , with the probability of $e_i.simRE(r_m)$. The probability will be adaptive to requests accordingly. Neighbors refer to its trustees. i.e., there are trust relationships from this entity to them.
- **Recommendation-based search.** An entity e_i will discover candidates within a distance $maxD$ around itself with a probability of $1 - e_i.simRE(r_m)$ when the receiving request is relatively near the entity's ability. That is, the entity will enlarge its search area from its neighbors to the area within a given distance. This behavior is inspired by two observations: 1) besides direct experiences, indirect experiences are also important for entities to make decisions; 2) communities have been discovered in many real-world networks, i.e., agents with similar interests cluster together. Here, the request far beyond the ability of an entity is likely out of its neighbors' abilities.
- **Random search.** An entity e_i will search in the whole network when it is a newcomer or its neighbors are all involved in the current request.
- **Trust-based select.** The entity e_i will select an entity e_j with the maximal degree of trust, $e_i.trust(e_j, t)$, from its candidates as its partner with a probability of $e_i.simRE(r_m)$. Entities are supposed to be more confident in selecting partners for requests which are relatively closer to its abilities.
- **Random select.** The entity e_i will stochastically select an entity from the candidate set $e_i.cand(r_m)$ as its partner with a probability of $1 - e_i.simRE(r_m)$.

3.3 The Mechanisms of Trust Relationships

We have introduced how trust relationships are applied in activating entities' behaviors for finding partners in the previous section. Now, we will introduce how they are changed based on the feedback from partners. Trust relationships are defined as follows:

Definition 4. A link $l'_{ij} = \{ \langle e_i, e_j, e_i.succ(e_j, t), e_i.fail(e_j, t), e_i.latestTime(e_j), e_i.trust(e_j, t) \rangle \mid 1 \leq i, j \leq N_e \}$ reflects a trust relationship from e_i to e_j where

- $e_i.succ(e_j, t)$ and $e_i.fail(e_j, t)$ denote the numbers of successful and failed interactions, respectively;
- $e_i.latestTime(e_j)$ is the time of their latest interaction;
- $e_i.trust(e_j, t)$ is the degree of trust at time t , quantifying the belief that e_j will accomplish a new request from e_i . It is a numeric value in $[\varepsilon, 1.0]$, where the threshold ε will be described below.

Once an entity e_i receives the feedback of its partner $e_j.simRE(r_m)$, a new relationship $l_{ij}(t)$ will be generated if it is not in the network. Otherwise, the existing relationship will be updated. The parameters of this relationship will be assigned as introduced below.

Firstly, the time of their latest interaction is set as the current time. Then, e_i will evaluate whether the partner is more suitable for the current request with Eq. [1](#) where λ is a threshold, $\lambda \in (0, 1)$.

$$e_i.QoI(e_j, r_m) = \begin{cases} true & \frac{e_j.simRE(r_m)}{e_i.simRE(r_m)} > (1 + \lambda); \\ false & otherwise \end{cases} \quad (1)$$

Other parameters will be assigned according to the result of $e_i.QoI(e_j)$. When the partner is more suitable for this request, i.e., $e_i.QoI(e_j, r_m) = true$, this interaction is regarded as successful and the number of their successful interactions increases. Otherwise, the number of failed interactions increases. Then, the degree of trust $e_i.trust(e_j, t)$ can be set with updated $e_i.succ(e_j, t)$ and $e_i.fail(e_j, t)$ as follows:

$$e_i.trust(e_j, t) = \frac{e_i.succ(e_j, t)}{e_i.succ(e_j, t) + e_i.fail(e_j, t)} \quad (2)$$

Trust relationships may be removed since the degree of trust is supposed to decay over time as follows:

$$e_i.trust(e_j, t) = e_i.trust(e_j, e_i.latestTime(e_j)) - \eta * (t - e_i.latestTime(e_j)) \quad (3)$$

where η is the decay factor, $\eta \in (0.0, 1.0)$. Once the degree of trust is less than a small negative numeric value ε , i.e., $e_i.trust(e_j, t) < \varepsilon$, the neighbor is interpreted as being no longer able to accomplish new requests from this entity, and the corresponding relationship will be eliminated. The threshold ε is empirically set, $\varepsilon \in (-0.05, 0)$.

In addition, entities will derive new trust relationships from the relationships already in the network. If the shortest distance from entity e_i to entity e_j is

larger than a given threshold $maxD$, i.e., $e_i.shortestDis(e_j, t) > maxD$, e_i has no idea about the ability of e_j and the degree of trust $e_i.trust(e_j, t)$ is set as 0. Otherwise, the following functions can be used to compute the degree of trust for the derived relationships:

$$e_i.trust(e_j, t) = \frac{\sum_{k=1}^{n_{ij}(t)} e_i^k.trustPath(e_j, t)}{n_{ij}(t)}$$

$$e_i^k.trustPath(e_j, t) = \sum_{l'_m j(t) \in L'(t)} e_i^k.trustPath(e_m, t) * e_m.trust(e_j, t) \quad (4)$$

where $n_{ij}(t)$ is the number of the shortest paths from e_i to e_j and the function $e_i^k.trustPath(e_j, t)$ denotes the degree of trust derived from the k th shortest path. If e_j is a neighbor of e_i , $e_i^k.trustPath(e_j, t)$ will be computed with Eq. 3.

3.4 Evaluation Functions

We define two functions, $e_i.simRE(r_m)$ and $e_i.match(r_m)$, for entities to evaluate whether they can accomplish received requests in this section. A cosine-based function is given to compute the degree of similarity between request r_m and the ability of entity e_i , i.e.,

$$e_i.simRE(r_m) = \frac{\sum_{k=1}^{N_{service}} (r_m.rw_k * e_i.cw_k)}{\sqrt{\sum_{k=1}^{N_{service}} r_m.rw_k^2 * \sum_{k=1}^{N_{service}} e_i.cw_k^2}} \quad (5)$$

When the value of $e_i.simRE(r_m)$ is larger than a threshold δ , $\delta \in (0, 1)$, the request r_m is assumed to be accomplished by the entity e_i , i.e.,

$$e_i.match(r_m) = \begin{cases} true & simRE(r_m) > \delta; \\ false & otherwise \end{cases} \quad (6)$$

3.5 The Algorithm

The trust network will evolve through continually finding partners based on the above-defined behaviors and trust relationships. Algorithm 1 describes how entities find partners once a new request is stochastically submitted. $search()$ and $select()$ represent entities activate a behavior of search and selection, respectively. $updateState()$ denotes that entities update their trust relationships. As defined before, $G'(t)$ represents the trust network at time t , i.e., t requests have been submitted to the network. If no request has ever been submitted, the network $G'(t)$ only contains N_e independent entities, i.e., $G'(0) = \langle \{e_1, \dots, e_{N_e}\}, \emptyset \rangle$.

Algorithm 1. The Autonomy-Oriented Partner Finding

```

Input: A new request  $r_m$ , the dynamic trust network  $G'(t)$ 
Output: The evolved network  $G'(t+1)$ 
begin
  // initialization phase
  stochastically select an entity  $e_i$  as the receiver of new request  $r_m$ ;
   $current\_entity \leftarrow e_i$ ;    $next\_entity \leftarrow null$ ;    $flag \leftarrow true$ ;
  // self-organized computing phase
  while  $r_m.T_{max} > 0$  do
     $matchResult \leftarrow current\_entity.match(r_m)$  based on Eq. 6;
    if  $matchResult \neq true$  then
       $current\_entity.search()$ ;
       $next\_entity \leftarrow current\_entity.select()$ ;
      // update based on feedback
       $flag \leftarrow current\_entity.QoI(next\_entity, r_m)$  based on Eq. 11;
       $current\_entity.updateState(flag)$  based on Eq. 2;
       $current\_entity \leftarrow next\_entity$ ;    $r_m.T_{max} = r_m.T_{max} - 1$ ;
    else
       $r_m.T_{max} = 0$ ;
    end
  end
end

```

4 Experiments

In the above, we have presented a dynamic trust network, which contains autonomous entities as nodes and self-organized trust relationships as links. This section gives some experiments with the following objectives:

1. to observe the structural emergence of the dynamic trust network;
2. to illustrate the dynamics of trust relationships;
3. to measure the efficiency of the trust network in finding partners.

4.1 The Structural Emergence of the Trust Network

Here, we explicitly display the structure of a dynamic trust network after finishing several requests. We also construct a scale-free benchmark network based on the abilities of the same entities. Experimental parameters are listed in Table 11. Entities' performances on providing each type of service are supposed to follow a power-law distribution and the power falls in the range of [1, 2]. Given that one cycle denotes the time spent on finishing a request, more one trust relationship are changed in each cycle.

The phenomenon of scale-free has been testified in many real-world networks as a consequence of preferential attachment, i.e., entities with a higher indegree (outdegree) will be connected with a higher probability. Extended from the GLP generator [12], the scale-free network is built based on the following principles:

Table 1. The experimental parameters

	parameters
	$N_{service} = 5$
scale-free network	$N_e = 1000, N_{edge} = 4200, m = 1.13,$ $\alpha = -1.7, minSimD = 0.3$
trust network	$N_{request} = 10000, \lambda = 0.2, maxD = 4$ $\delta = 0.95, \eta = 0.0003, \varepsilon = -0.02$

Table 2. The structural characteristics of two networks

structural characteristics	scale-free network	trust network
network indegree/outdegree	4.373	3.800
clustering coefficient	0.172	0.012
APL^{-1}	0.387	0.221

- A link will be created unless the degree of cosine-based similarity between the abilities of its two end nodes is not less than a threshold $minSimD$. Entities are assumed to cluster with similar interests.
- The preferred probability of an entity as an end node of a new directed link is computed with Eq. 7, in which β is an adjusting factor in $(0, 1)$.

$$\begin{aligned}
 e_i.sourcePreferProb &= \frac{e_i.outdegree - \beta}{\sum_{e \in E} (e.outdegree - \beta)} \\
 e_i.targetPreferProb &= \frac{e_i.indegree - \beta}{\sum_{e \in E} (e.indegree - \beta)}
 \end{aligned} \tag{7}$$

The structural characteristics of two networks are illustrated in Table 2 and Fig. 1. The results show that these two networks are quite similar. It can be concluded that the dynamic trust network converges to be scale-free. Moreover, this convergence is not irrelevant with the distribution of entities' abilities, which is briefly mentioned due to space limitation. The trust network contains 3800 trust relationships. Its average path length is about 4.5. In this network, more than 300 entities have only one trustee while a few entities trust more than 30. Specially, the two distributions largely overlap in Fig. 1b. That is, the two distributions approximate to each other.

4.2 The Dynamics of Trust Relationships

We have observed the emergence of our trust network in the previous section. Now, we examine the dynamics of trust relationships by measuring two structure characteristics: the clustering coefficient and the harmonic mean of average path length APL^{-1} . The results are shown in Fig. 2. The horizontal axes denote the

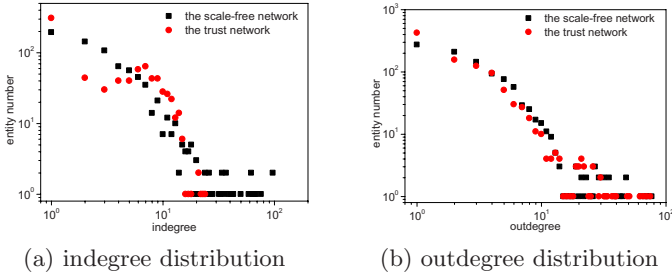


Fig. 1. The distributions of entities’ indegree (a) and outdegree (b) in the two networks. One is the dynamic trust network after finishing 10000 requests. The other is a standard scale-free network.

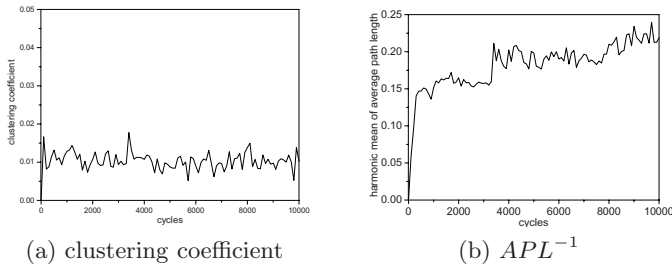


Fig. 2. The dynamics of trust relationships with the measurement of structural characteristics. (a) clustering coefficient, and (b) the harmonic mean of average path length.

cycle number (request number) while the vertical axes represent the clustering coefficient and APL^{-1} , respectively. We can find that their values fluctuate slightly after changing rapidly in the first 200 cycles. It means that entities quickly find their relatively stable partners for various requests.

We can observe that the clustering coefficient increases rapidly at first and then rises slowly, even keeps steady at some time in Fig. 2a. Likewise, the APL^{-1} increases rapidly at first and fluctuates relatively little, i.e., the average path length decreases rapidly and then fluctuates relatively less. At the beginning, many relationships are generated since entities mostly activate stochastic behaviors. The behaviors are exploratory in nature, which results in that many generated relationships are useless and eliminated later. Only a few trust relationships remain, i.e., entities refine some entities as potential partners. Therefore, the probability of adding new relationships is quite low and the network structure shows a slight change.

4.3 The Efficiency of the Trust Network in Finding Partners

In this section, we will examine the efficiency of our network in finding partners with two measurements: the accomplishment ratio and the average propagation

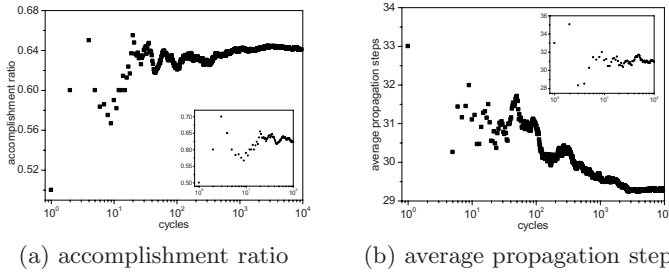


Fig. 3. The dynamics of the efficiency in finding partners for finishing 10000 requests (cycles). The sub-graphs denote the efficiency in the first 100 cycles. (a) the accomplishment ratio of requests, and (b) the average times that entities find partners for a request.

step. The values of the measurements are averaged over 10 iterations. In each iteration, 10000 requests are stochastically submitted to our network. Besides the parameters in Table I, the T_{max} of each request is fixed as 50. As to be described later, experimental results show that entities can accomplish requests with a higher probability in fewer steps, along with the dynamic trust relationships.

The dynamics of the efficiency are presented in Fig. 3. The accomplishment ratio converges to 0.641 while the average propagation step converges to 29.266. Fig. 3a shows that the ratio quickly increases to its convergence level. Fig. 3b displays that the average step decreases rapidly at first and later fluctuates slightly. This dynamics can be explained based on trust relationships. At first, there is few relationships, wherefore entities activate stochastic behaviors with a higher probability. The accomplishment ratio is low and the average propagation step is accordingly large. After a certain period of time, entities find some trustworthy ones and maintain corresponding trust relationships. Hence, entities can quickly find partners to finish new requests based on existing relationships. Therefore, the ratio keeps relatively higher and the average step is relatively lower.

5 Conclusions

In this study, we are interested in characterizing dynamically-changing trust relationships in real-world applications. We attempt to understand why trust relationships can enable distributed entities quickly find partners to provide their requested services.

Specifically, we have proposed a dynamic trust network by means of utilizing the ideas of autonomy and self-organization from AOC. These ideas can help us understand how network-level phenomena emerge from entities' local behaviors. In this study, the notion of autonomy means that distributed entities (nodes) activate different behaviors based on their abilities and trust relationships. Self-organization refers to the process that entities change their relationships (links) with positive feedback mechanisms, according to the feedback from their partners.

Experimental results have shown that the network quickly converges to be scale-free. In other words, parts of dynamically-generated trust relationships remain relatively stable while others are eliminated. Our results have also shown the accomplishment ratio quickly arises to a level while the average propagation step gradually decreases. Hence, it is reasonable to say that the process of self-organization makes trust relationships emerge quickly. Moreover, this process leads to the high efficiency of entities in successfully finding partners.

Acknowledgment

This work was supported by the National Natural Science Foundation of China Grant 60673015, the Hong Kong Research Grants Council grant (210508/32-08-105), and the Major State Basic Research Development Program of China (973 Program)(2003CB317001).

References

1. Golbeck, J.: Weaving a web of trust. *Science* 321(5896), 1640–1641 (2008)
2. Marsh, S.P.: Formalising Trust as a Computational Concept. PhD thesis, University of Stirling (1994)
3. Erete, L., Ferguson, E., Sen, S.: Learning task-specific trust decisions. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multi-Agent Systems, pp. 1477–1480 (2008)
4. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The eigentrust algorithm for reputation management in P2P networks. In: Proceedings of the 12th International World Wide Web Conference, pp. 640–651 (2003)
5. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. *Decision Support Systems* 43, 618–644 (2007)
6. Ramchurn, S.D., Huynh, T.D., Jennings, N.R.: Trust in multi-agent systems. *The Knowledge Engineering Review* 19(1), 1–25 (2004)
7. Sabater, J., Sierra, C.: Review on computational trust and reputation models. *Artificial Intelligence Review* 24, 33–60 (2005)
8. Liu, J.: Autonomy-oriented computing (AOC): The nature and implications of a paradigm for self-organized computing (keynote talk). In: Proceedings of the 4th International Conference on Natural Computation and the 5th International Conference on Fuzzy Systems and Knowledge Discovery, pp. 3–11 (2008)
9. Liu, J., Jin, X., Tsui, K.C.: *Autonomy Oriented Computing: From Problem Solving to Complex Systems Modeling*. Springer, Heidelberg (2005)
10. Liu, J., Jin, X., Tsui, K.C.: Formulating computational systems with autonomous components. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35(6), 879–902 (2005)
11. Zhang, S., Liu, J.: Autonomy-oriented social networks modeling: Discovering the dynamics of emergent structure and performance. *International Journal of Pattern Recognition and Artificial Intelligence* 21(4), 611–638 (2007)
12. Bu, T., Towsley, D.: On distinguishing between internet power law topology generators. In: Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communication Societies, vol. 2, pp. 638–647 (2002)

Modeling an Educational Multi-Agent System in MaSE

Izabela Salotti Braga Gago, Vera M.B. Werneck, and Rosa M. Costa

Universidade do Estado do Rio de Janeiro – UERJ-IME
Rua São Francisco Xavier 524, 6o Andar Bloco B – Rio de Janeiro – RJ, Brazil
{vera,rcosta}@ime.uerj.br

Abstract. The growing complexity of modern systems is demanding more and more time from software designers and developers. The challenge of constructing complex systems is complicated by the geographic distribution of the systems. An agent oriented paradigm reinforces the software flexibility and the agent's social possibilities, taking space as a solution in software engineering. The idea of independent and autonomous entities capable of relating in the search of the system goals originated the multi-agent systems, where at least two agents are capable of interaction. Such systems possess great applicability in different knowledge areas, as for example, in Education. This work defines an Education Multi-Agent System, which is a learning education environment with multi-agents. The system aims at helping the teaching process on a specific topic. The system agents were modeled using the MaSE (Multi-Agent System Engineering) method.

1 Introduction

The educational software has been developed since the mid 60s to use Artificial Intelligence (AI) techniques. The software intends to create more flexible programs able to support individualized instructions. The Intelligent Tutoring Systems (ITS) is one of the programs that stand out because it permits the particularized teaching of specific domains by diagnosing the students' structures and knowledge levels and by respecting their learning style and rhythm. In addition, it promotes the shaping of well-structured educational interactions, which are adapted to different kinds of users [3].

The increasing use of multi-agent systems brings challenges that have not been studied yet, such as: how should we adapt elicitation of requirements to cope with agent properties like autonomy, sociability and proactiveness. The agent-oriented modeling is proposed as a suitable software engineering approach for complex organizational application domains that deal with the need for new applications. These requirements are not broadly considered by current paradigms. Autonomy and sociability aspects such as the dependency of an agent on another, and how critical this condition should be, have to be analyzed from the early stages of the software development process [13].

In this work the agent orientation paradigm is adopted in the modeling of an Intelligent Educational Environment, using the MaSE methodology for defining and designing the multi-agent system. We have chosen MaSE because this methodology uses some UML concepts and diagrams that make the process easier to learn and the

support tool free for use. It also has a good performance, leading to a high quality modeling.

This work is organized into 5 sections. Section 2 gives an overview of the MaSE methodology and section 3 defines the purpose of Intelligent Educational Multi-Agent Systems, discussing the system models and agents. Section 4 presents the model requirements and the project design of the system using MaSE. Finally section 5 concludes the work and presents future works.

2 MaSE Methodology

The main purpose of the Multi-agent System Engineering (MaSE) methodology is to support the designer in catching a set of initial requirements and analyzing, drawing, and implementing a multi-agent system (MAS). This methodology is independent of any agent's architecture, programming language, or communication framework. The MaSE treats the agents as a deeper object orientation paradigm, where the agents are object specializations. Instead of simple objects, with methods that can be invoked by other objects, the agents talk among themselves and proactively act in order to reach goals [5], [9].

MaSE is a specialization of traditional software engineering methodologies with two phases (Analysis and Design) and several activities [5]. The MaSE Analysis phase has three steps: Capturing Goals, Applying *Use Cases*, and Refining Roles. The Design phase has four activities: Creating Agent Classes, Constructing Conversations, Assembling Agent Classes, and System Design.

The Analysis phase aims at defining a set of roles that can be used to achieve the goals of the system level. These roles are detailed by a series of tasks, which are described by finite-state models. This process is performed in three steps represented by a Goal Hierarchy: *Use Cases*, *Sequence Diagrams*, *Concurrency Tasks* and *Role Model*.

Goals are what the system is trying to achieve and they generally remain constant throughout the rest of the Analysis and Design phases. Capturing Goals can be divided into two parts: identifying and structuring goals. The Goal representation is a form of hierarchy where the goals are decomposed.

Use Cases describe the behavior of agents for each situation in MAS. In the step *Applying Use Cases*, situations of the initial requirements are acquired and represented into *Use Cases* Diagrams and Descriptions, and UML *Sequence Diagrams*. These representations help the designer show the desired system behavior and its sequences of events.

The third step of modeling is the transformation of MaSE goals into roles. A role is an expected abstract description behavior of each agent that aids in reaching the system goals. In general, the transformation of goals into roles is done in a proportion of one to one: each goal is mapped to a role. However, in some situations it is useful to combine goals into a single role.

The Role Diagram has a number of considerations regarding the representation role and communication between roles. First, each role associated with the same goals is listed below the role name. Often, these goals are represented by numbers used in the Goal Diagram. A set of tasks is associated with each role, representing the expected role behavior.

Once the roles are defined, you must describe the details of each task in the Role Diagram. The definitions of the tasks are shown in *Concurrent Tasks* Diagrams based on finite automata states. By definition, each task must be executed concurrently, while communicating with other internal or external tasks. A *concurrent task* is a set of states and transitions. The states represent the internal agent mechanism, while the transitions define tasks communications. Every transition has an origin and a destination state, a trigger, a guard condition and a transmission [8].

In general, the events that are sent as broadcasts or triggers are associated with events sent to work in the same role instance, requiring an internal coordination of each task. The representation of messages sent between agents uses two special events: send and receive. The send event represents the message sent to another agent and is denoted by send (message, agent), while the receive event is used to define the message received from another agent denoted by receive (message, agent).

The Design phase includes the following diagrams: Agent Classes, Conversations, Agent Architecture and Deployment Diagram. The first step in the design process involves the definition of each agent class, which is documented in an Agent Class Diagram. The system designer maps each role defined in the Role Diagram in at least one agent class. Since the roles are derived from the system goals and are responsible for meeting them, ensuring that each role is mapped in at least one agent class helps guarantee the implementation of the goals in the system. In general the agent classes can be thought of as templates defined in terms of roles they play and of the protocols they use to coordinate with other agents [5], [8].

An agent class has two components: one or more roles and their conversations with other agents. Conversations in an agent class are made between two players, the conversation initiator and the recipient. The conversation detail is expressed through the communication protocols based on *concurrent tasks* that were identified during the analysis in the Role Diagram.

Each task defined in the *Concurrent Task* Diagram describes a component in the agent class. The *concurrent tasks* can be translated into multiple conversations in the Conversation Diagram.

The definition of the agent architecture is performed in two steps: (i) definition of the agent architecture and (ii) components. The designer can choose the agent architecture, such as Belief-Desire-Intention (BDI), Reactive or Knowledge Base. For example, a reactive architecture includes components such as an interface to other agent classes, an internal controller class, a rule verifier, and sensors [4].

The final step in the MaSE design phase uses the Deployment Diagram to show the number, types and location of agents in the system. The diagram describes a system based on agent classes, and is very similar to the UML Deployment Diagram.

The system designer can use the Deployment Diagram to define different configurations of agents and platforms to maximize the processing power and bandwidth of a network.

3 Educational Multi-Agent System Overview

The Educational Multi-Agent System proposed in this work relies on the ITS' classical architecture, considering four models: Pedagogic, Expert, Student and Interface. Each model reflects the ability and the characteristics of the Educational System [12], [13].

The student knowledge should be measured through a questionnaire, and through concerns expressed by the student during the learning process. The student model can follow certain principles such as Differential Model, Overlay Model, Model of Disturbance, Simulation Model, Stereotype Model or Belief Model.

The pedagogical model contains teaching strategies adopted based on the input submitted by the student. Such strategies structure the way in which knowledge is represented and taught to the student, ensuring the success of the methods used by some teaching strategies such as Socrates, Coaching, Hypertext and Others. The tutor can use one or a combination of behaviors, and goals that may vary depending on the information received from the student (actions) and the goal's purpose (plan) [2].

The pedagogical agents may act as virtual tutors, virtual students or colleagues who assist in the virtual learning process. These action forms result on group formations that can distribute tasks between themselves, from greater (student modeling or selection of the strategy and tactics) or lesser (strategy responsibility by each agent) extension.

Dividing the system into smaller tasks reduces the complexity and the monitoring task executed at a higher abstraction level [10]. The learning process characterization developed in some multi-agents systems can be classified in three abstraction levels: (i) replication learning, (ii) learning by tautology or (iii) learning by interaction and shared dynamic [2].

The system is modeled using the paradigm of oriented agents, and each model's basic architecture is represented by an agent. We included other agents in the system to support the key actors. This model uses the definition of learning by replication, where each agent has a specific task, and one of them, the tutor task.

The next subsection describes the agent system models and their characteristics in the system. The whole system uses Artificial Intelligence (AI), Psychology and Pedagogy solutions, involving efficient ways to perform the most varied tasks.

3.1 Student Model

At first, the student model is not represented by an agent in the system architecture. While in the process of modeling it can be represented as an agent, in the proposed student model it is treated as an object that communicates with other elements of the system through messages because we consider this model the representation of the human user in the system. Thus, at certain times of the tutor learning process the student can behave in a reactive, as well as, an active form, indicating concern about a particular point of the subject [4], [12].

The user of the system is represented by the same type of student that looks more like an object, or may have more than one instance and communicates with the other elements of the system through messages.

The stereotype model is used in the modeling of the student, considering the student's initial knowledge in the following classification: beginner, intermediate, advanced and expert. This model type is a simplified model. The strategies of the teaching system are developed and modified according to the level of the student. At the end of each topic, an open question session is opened and a test is applied. After that the tutor starts the process of classifying the user in one of the levels.

3.2 Educational Model

The pedagogical model adopted in this system aims at guiding the student. The tutor sets the teaching strategies, knowledge representation and correction of the tests. At the end it shall be responsible for grading the student to a new level and selecting the next topic to be presented [2].

The tasks to be performed by the tutor in this system are divided in two agents: the Tutor agent and the Coordinator agent.

The tutor agent is responsible for receiving the education strategy and for selecting the module that starts the education process. At the end of the module presentation the tutor opens space for the student to ask questions based on a questionnaire, where the question to be selected by the student is the one that mostly represents his concern. After the question is selected the Tutor agent asks the Expert agent about the answer and the expert agent performs a search in the knowledge base. This cycle can be repeated until the student signals that he understood the question or until the answer possibilities in the knowledge base are exhausted. For the second case, a notification is sent via e-mail to a human teacher registered in the system, and in the future the teacher contacts the user to clarify his question.

The Coordinator Agent selects the strategy in education (or method) according to the initial student level, which will be obtained through a standard questionnaire to be filled out by the student at the beginning of the course. When the teaching strategy is selected the method will also be available to the Tutor agent. The Agent Coordinator is additionally responsible for selecting the next module to be submitted by the Tutor and must define goals for each course, such as teaching the subject until the end, taking questions from students and others. As these goals are met the Tutor agent goes through the teaching process. The definition of the course modules, the content of the next module to be presented, and the teaching strategy is the student's course plan, while the lecture plan is the course plan added to the goals to be achieved on each topic introduced to the student.

3.3 Expert Model

The expert model is responsible for the base knowledge maintenance, providing any necessary information to all actors of the system. The Tutor and Coordinator agents use the knowledge base in the selection of the teaching strategy and in the selection of a response to a query of the pupil, respectively.

The model represents a major expert system intelligent tutor and the rest of the system depends on its knowledge. Therefore, it should be modeled as an expert in order to address the peculiarities inherited to the field you want to express. For example, in a STI in the geometric area, the modules may have different features than other fields like history or nutrition [12].

In this work we do not propose to model the ontology for a specific domain, instead only the basic features of an expert system are modeled.

This expert model adopts the production rule model, which is the most appropriate for the representation of declarative knowledge. This agent is called Expert and has a memory region or blackboard. The region has a mechanism where the requests of

other agents and the system will be checked to match the rules in the knowledge base. In the case of correspondence, it starts the search process of production, which together with its rule is the knowledge. Thus knowledge is sent to the request agent.

3.4 Interface Model

The interface model is represented as streamlined as possible, since it depends on the displayed area, the teaching method style, presentation and used resources. In this work the interface is represented only as a mechanism for input and output of information between the system and user, where the system represents only the type of student.

3.5 Other System Agents

During the system requirement description the need for auxiliary specific tasks was noticed and two agents were detected: the Management agent and the Database Management agent.

The Management agent is active at the beginning of the system as are all the other agents, but it is the first to perform a task. It's first task is to detect the presence of the user in the system by monitoring the system. The agent will invite the user to register, which will have questions concerning the system domain area. This determines the student's initial level, which will be recorded in the database. Consequentially, the Management agent must create the student registration. The initial student level is recorded by the Coordinator agent that will select an appropriate strategy for the student education.

The Database Management agent is responsible for both meeting the other system agent requirements of databases, for maintaining the information validation and formatting the correct information required by other agents.

4 Educational Multi-Agent System MaSE Modeling

The Educational Multi-Agent System was named Educ-MAS and was developed based on MaSE methodology and on the issues discussed and defined in the previous section. The tool used to draw the diagrams in MaSE was AgentTool [1] that guides the methodology application [5], [7], [8].

In the Analysis first step, the main goal and its sub-goals were identified from the initial set of the system. They were then structured in a goal tree as shown in Figure 1.

The goal *Define Tuition Plan* was structured based on the four models described in section 2. Our examination of the importance and inter-relationships of the goals was also taken into consideration. The *Promoting individual learning* is partitioned into four goals: *Explore Student*, *Plan Tuition*, *Manage Knowledge* and *Manage Tuition*. The goals are also decomposed into other goals. For example the goal *Plan Tuition* was partitioned into two sub-goals (*Consult Defined Goals* and *Define Tuition Plan*) and the sub-goal *Define Tuition Plan* has two sub-goals named *Defining Content of the Module* and *Defining Plan Presentation of the Modules*.

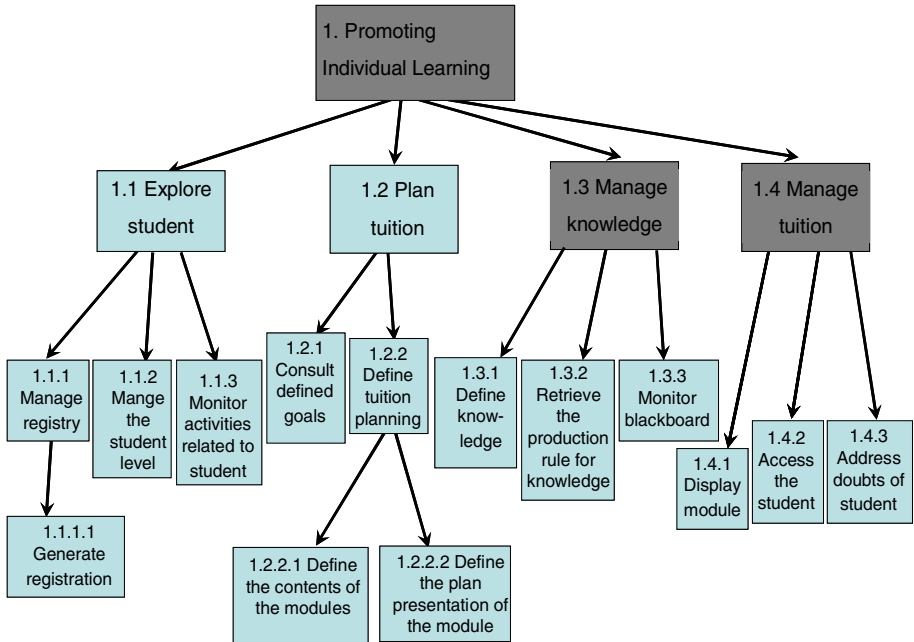


Fig. 1. Educ-MAS Goals

Use Cases

Manage register
Define student's level
Define lesson plan
Retrieve content of module
Plan the course
Provide class
Assess the student
Monitor activities of student

Sequence Diagrams

- Student's class
- Question resolved
- Question not resolved

Description

Use case: Provide class

Agents: Tutor, Expert, Administrator

-Pre-conditions:

- 1) A course plan must exist assigned to the student

-Normal flow:

- 1) The Tutor asks the Administrator agent to retrieve the course plan and with it the module content
- 2) The Tutor selects and show the next topic to the student
- 3) The student says that the topic is finalized, the Tutor goes to step 2
- 4) If the topics ended up, the Tutor opens a question session
- 5) If the student says that has a question, the Tutor search for a bookmark with the Expert agent that contains questions that represents the default questions
 - 5.1) The student selects a question, the Tutor agent try to solve
 - 5.2) If the student does not satisfied the Tutor agent selects another answer with the Expert agent
- 6) The student says that understood the answer and the Tutor agent ends the session after registering the doubt

Fig. 2. Use Case Provide Class

After defining the goals, eight primary *use cases* were generated based on the goals *Explore Student* and *Plan Tuition*. The Figure 2 shows the *use case Provide Class*, its description and also the name of *Sequence Diagrams* that retracts the scenarios of this

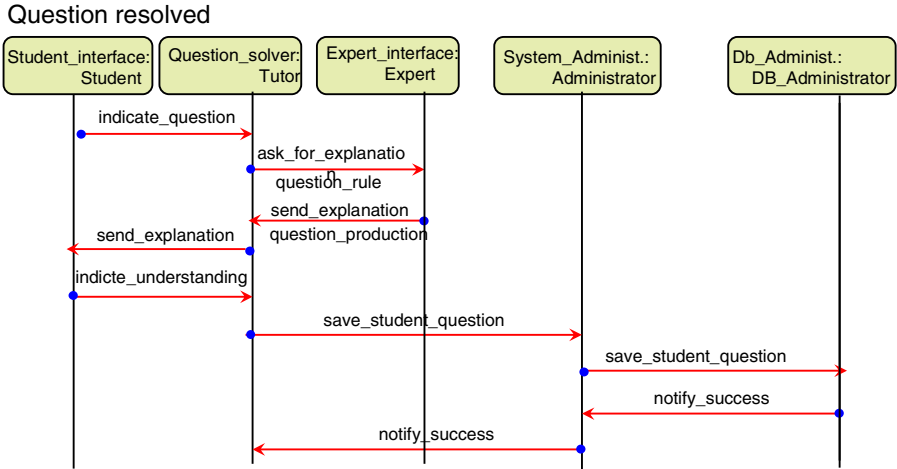


Fig. 3. Question Resolved Sequence Diagram

use case. The scenarios are Student’s Class, Questions Resolved and Questions Not Resolved. For each one we built a Sequence Diagram that shows how the system behaves. Figure 3 demonstrates the agent behavior with the second scenario of the case *Provide Class*. We defined for each use cases the description and their scenario Sequence Diagram. The whole specification of Educ-MAS can be found in [11].

We developed a set of roles and tasks to show how the goals are reached based on the goals, the use cases (diagrams and descriptions) and the Sequence Diagrams. The Figure 4 represents the Preliminary Role Diagram where the goals were mapped to system roles. For example, the *Course Administrator* role achieves the goals *Plan Tuition*, *Consult Defined Goals*, *Define Tuition Plan*, *Defining Content of the Module* and *Defining Plan Presentation of the Modules*. In the complete Role Diagram we introduced the tasks responsible for the roles and the associations among themselves to reach the responsible goal’s roles.

Once the Role Diagram was developed, we defined a *Concurrent Task* Diagram for each task as presented in Figure 5 for the task Monitor Blackboard. This task is associated to the Expert interface role, which is responsible for the 1.3.3 goal that monitors the blackboard in order to interface the knowledge base introducing the contents and answering questions.

The Analysis Models were built showing the behavior of the system by deriving and establishing the goals that were used to create the use cases and the sequence Diagrams. Those goals were mapped into roles and tasks defined in the Role Diagram and detailed in the Concurrency Task diagram. In the Design phase we use those diagrams and we constructed the individual components of the agent classes as presented in Figure 6. We chose a simple agent architecture and an example of a Tutor agent class partial structure component as shown in Figure 7, where the attributes and methods were derived from the *Concurrent Task* diagram.

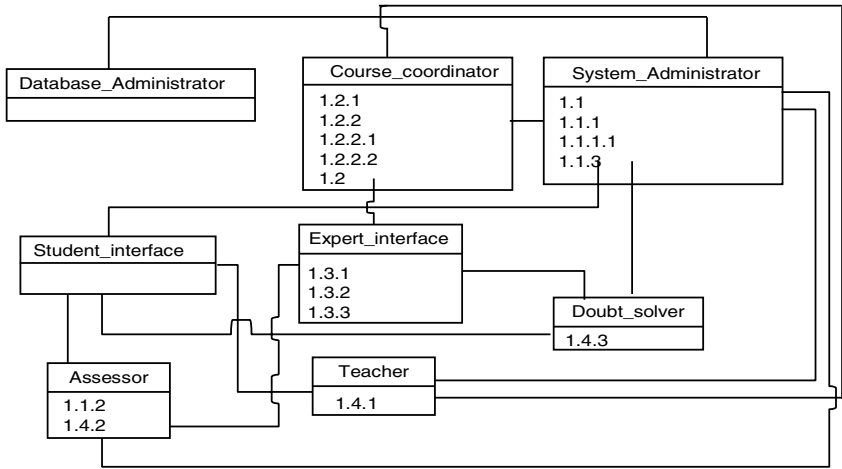


Fig. 4. The Educ-MAS Partial Role Diagram

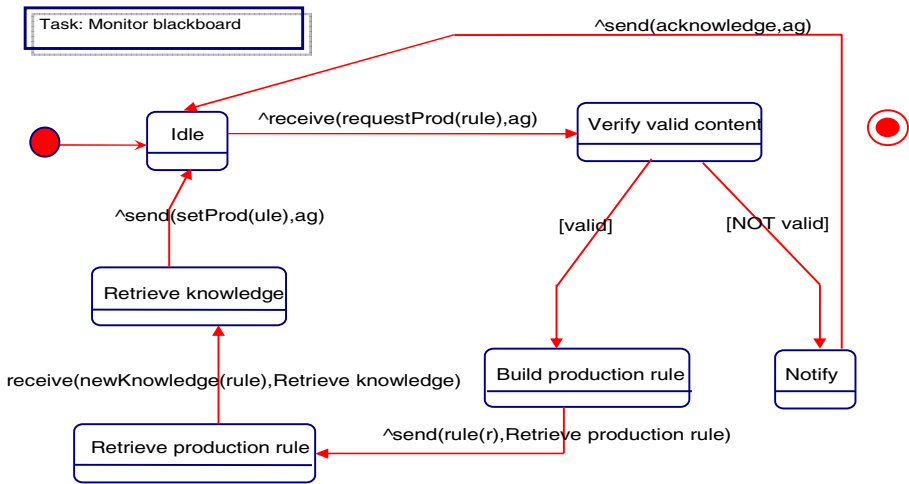


Fig. 5. Concurrent Task Diagram for the task Monitor blackboard

The last step in the Design phase was to develop an overall operational design. The Educ-MAS is presented in Figure 8. The Tutor, Administrator, Coordinator and Expert Agents are defined in an environment as a system. The Interface (Student Model) starts at the student’s computer where the Database Management and the other part of the system are in network computers.

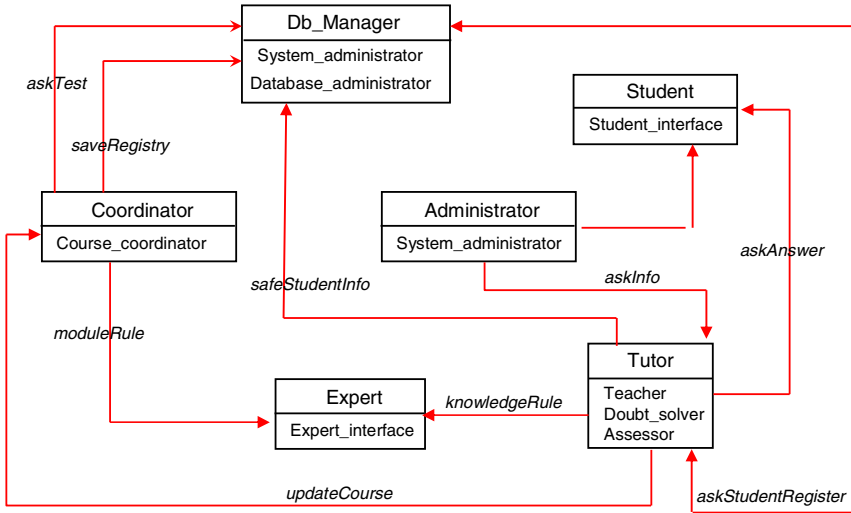


Fig. 6. The Educ-MAS Agent Class Diagram

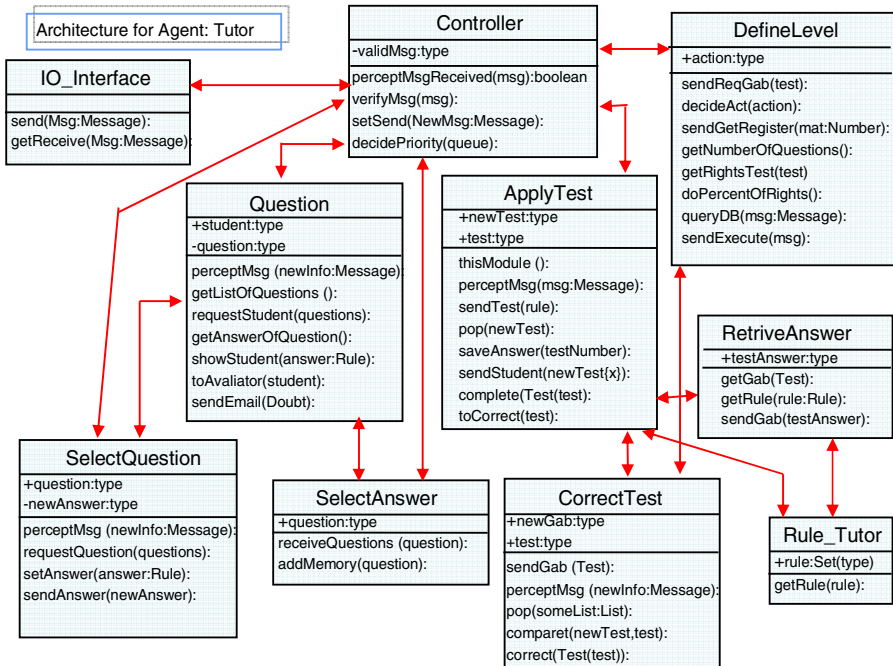


Fig. 7. Tutor Agent Class Partial Structure Component

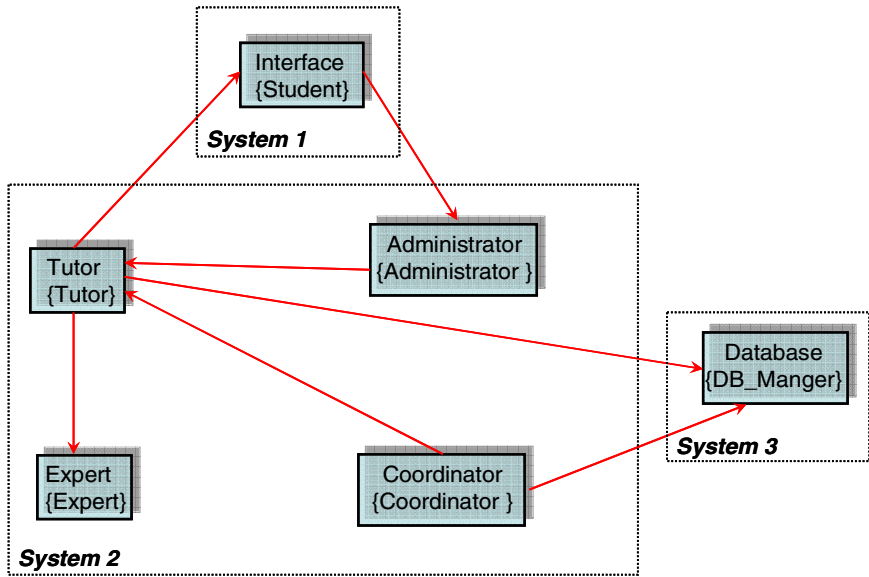


Fig. 8. The Educ-MAS Deployment Diagram

5 Conclusion

During the development of this work several concepts related to the paradigm of the oriented agents were explored. The MAS has a wide application both in industry and in science. Since the MAS is an area of artificial intelligence, agents can be considered capable to make decisions, to learn and thus to be able to update their knowledge base. Therefore, the ability of MAS as a system to optimally solve problems to achieve the system goals becomes an attraction to the pedagogy area.

Although the development of artificial intelligence applied to MAS has evolved over the years, the intelligent tutoring systems still have certain limitations when compared to human professors and teachers. The reason for why the student does not learn certain topics, for example, is still not fully resolved in intelligent tutoring systems. This is an important feature to develop further knowledge especially on the student system.

Currently, the development of intelligent tutors for distance education has gained importance, and has been implemented to be accessible over the Internet, enabling teaching at any time or place.

The ITS modeled in this work was developed considering only the common features of ITS, disregarding the knowledge tutor field. However the MaSE modeling in its original definition does not mention any particular method of knowledge base modeling. The developers of the methodology understood this need and changed the life cycle of modeling MASE in order to include these types of systems. For the STI knowledge base construction we will use an ontology that contributed with the organization of the expert knowledge that seemed to be a tool to clarify the expert reasoning. The ontology represents the world in terms of an explicit conceptualization

that can be understood by different people and can be used for different approaches for computer systems development.

The MASE methodology introduced a further step in the process of analysis, specifying that the definition of ontology should happen within the modeling of the system [6]. This step will be applied during this phase when we build two specific EducMAS: the Geometry learning for high school and the Nutrition Bio-availability in a Nutrition Graduating program.

References

1. Agent Tool, <http://macr.cis.ksu.edu/projects/agentTool>
2. Bercht, M., Vicari, R.M.: Pedagogical agents with affective and cognitive dimensions. In: Sanchez, J. (ed.) Proceedings of the V Congreso Iberoamericano de Informática Educativa, 2000, Vinãdel Mar. RIBIE 2000. V Congreso Iberoamericano de Informática Educativa, RIBIE 2000, CD-ROM. Universidade do Chile (2000)
3. Bordini, R.H., Lesser, V., Viccari, R.M., Bazzan, A.L., Janone, R.O., Basso, D.M.: AgentSpeak(XL): efficient intention selection in BDI agents via decision-theoretic task scheduling. In: Castelfranchi, C., Johnson, W.L. (eds.) Proceedings of the AAMAS 2002—autonomous agents and multi-agents systems 2002, pp. 1294–1302. ACM Press, Bologna (1999)
4. Bryson, J., Stein, L.: Modularity and design in reactive intelligence. In: Int. Joint Conf. on Artificial Intelligence IJCAI 2001, Seattle (USA), pp. 1115–1120 (2001)
5. Deloach, S.A.: Analysis and Design using MaSE and agentTool. In: 2th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2001), Miami University, Oxford (2001)
6. Dileo, J., Jacobs, T., Deloach, S.: Integrating Ontologies into Multiagent Systems Engineering. In: Fourth International Bi-Conference Workshop on Agent-Oriented Information Systems AOIS 2002 (2002)
7. Goulart, R.R.V., Giraffa, L.M.M.: Arquiteturas de Sistemas Tutores Inteligentes. In: TECHNICAL REPORT SERIES – Faculdade de Informática – PUCRS – Brasil, http://www.pucrs.br/inf/pos/mestdout/rel_tec/tr011.pdf
8. wHenderson-Sellers, B., Giorgini, P. (eds.): Chapter XI – Multi-Agent Systems Engineering: An Overview and Case Study. In: Agent Oriented Methodologies. 1ed: Igi Global, pp. 317–340 (2005)
9. O'Malley, S.A., DeLoach, S.A.: Determining when to use an agent-oriented software engineering paradigm. In: Wooldridge, M.J., Weiß, G., Ciancarini, P. (eds.) AOSE 2001. LNCS, vol. 2222, p. 188. Springer, Heidelberg (2002)
10. Russell, S., Norving, P.: Intelligent Agents. In: Artificial Intelligence: A Modern Approach, 3rd edn. (2002), <http://www.cs.berkeley.edu/~russell/aima1e/chapter02.pdf>
11. suppressed
12. Vicaria, R.M., Flores, C.D., Silvestrea, A.M., Seixas, L.J., Ladeirac, M., Coelho, H.: A multi-agent intelligent environment for medical knowledge. Artificial Intelligence in Medicine 27, 335–366 (2003)
13. Wooldridge, M., Jennings, N.: Intelligent Agents: Theory and Practice (1997), <http://www.doc.mmu.ac.uk/STAFF/mike/ker95/ker95-html.html>

Enhancing Decentralized MAS-Based Framework for Composite Web Services Orchestration and Exception Handling by Means of Mobile Agents Technology

Mounira Ilahi¹, Zaki Brahmi², and Mohamed Mohsen Gammoudi²

¹ LI3 Member

² URSIIVA Member

{ilahi_mounira, zakibrahmi, gammoudimomo}@yahoo.fr

Abstract. Decentralized orchestration offers performance improvements in terms of increased throughput and scalability and lower response time. However, decentralized orchestration also brings additional complexity to the system, mainly, in terms of exception handling. The research presented in this paper is carried out on the basis of some previous work of the authors, including: decentralizing orchestration of composite Web services and exception handling. We focus also on current works expanding the previous one, exhibiting thus a higher performance degree which the integration of mobile agents performs by moving the application's functionality through the network.

Keywords: BPEL, Exception Handling, MAS, Web Service Orchestration, Mobile Agents.

1 Introduction

Web services enable the new generation of Internet based applications and adjust the way business applications are developed. Although there can be some value in accessing a single web service, the greater value is derived from assembling web services into more powerful applications. Business processes are typically complex operations, comprising of numerous individual stages, and in the context of SOA each such stage is realized as a web service. Web Services Business Process Execution Language (WSBPEL) is the current industry standard frequently used to specify the composition of these steps (control flow, data flow, etc), and executed by a Web Services Orchestration (WSO) platform. Nevertheless, a main inadequacy is that BPEL relies on the centralized coordinator on which the whole specification of a business process is executed. Thus, all the messages amongst the business partners are transferred and processed by the coordinator. Accordingly, the communication overload is relatively high and in some cases, the centralized coordinator becomes a bottleneck limiting so the scalability of the system. There are also arguments that this model is not flexible enough for scenarios where data flow has to be transported in a given path due to certain business constraints [6]. Starting from the intention to overcome these drawbacks, some decentralized solutions [7, 8, 9, 10, 12, 13] have been proposed.

A second key problem that we can identify in process-oriented composition languages like BPEL concerns support for dynamic adaptation of the composition logic. Such languages assume that the composition logic is predefined and static, an assumption that does not hold in the highly dynamic context of web services. In fact, new web services are offered and others disappear quite often. In addition, the organizations involved in a web service composition may change their business rules, partners, and collaboration conditions. Furthermore, software, machine or communication link failures may render certain sub-process of composite services unavailable, precluding thus the successful execution of the business process. Therefore, services-based systems are inherently vulnerable to exceptions. In these cases, a replacement component should be identified and substituted for the failed one. The replacement component should have the same skills to the later i.e. to have same functionality and QoS [2].

Note that due to the static nature of BPEL, which dictates that service bindings should be hard-coded in the scenario, it is not feasible to include calls to all replacement services within a fault handler (typically 2 or 3 alternates will be specified [3]) and not possible at all to dynamically introduce new bindings or remove outdated ones, to align with the changes in service availability. Each such change should generate a maintenance activity that will lead to modifications in the BPEL scenario code. This motivates the need for more flexible web orchestration engine of BPEL process, which supports a dynamic orchestration.

To deal with these limitations, we have proposed in [14] a decentralized framework for WSO and exception handling¹ based on Multi-Agent System technology. This framework allows to a BPEL process to be partitioned in a set of sub-processes. These sub-processes will be executed in a distributed fashion by a set of agents. Our approach provided a mechanism for dealing with system exceptions which is based on agent technology. Nevertheless, similarly to all other discussed proposals, our last framework is based on the assumption that each site is skilled and has the required infrastructure to do its work. An assumption that is not forever affordable. Here, we are focusing on a direction pro enhancing our previous framework. We swap, in the exception handling mechanism, the agent technology by the mobile agent technology for a purpose of resolving the above limitation, exhibiting thus a higher performance degree.

The rest of this paper is organized as follows: Major related work is discussed in Section 2. Then, the authors' previous work which builds the foundation of this research is introduced in Section 3, as well as a brief introduction to mobile agent technology. Section 4 is dedicated to outline our current work expanding the previous one by means of incorporating the mobile agent technology. Finally, Section 5 concludes this paper and sketches out our future work.

2 Related Work

There are two primary areas of research that are related to this work: decentralized composite web services orchestration and exception handling. This section stands

¹ Exception handling defines how exceptional situations that can arise during execution of the composite model should be handled.

some approaches dealing with these two issues. We present each proposal with respect to our two concerns (decentralizing composite web services orchestration of exception handling).

In [8], the authors present a technique to partition a composite web service written as a single BPEL program into an equivalent set of decentralized processes based on the PDG (Program Dependence Graph). This technique creates one partition per component web service in the flow specification. Each partition is deployed in network proximity with the component web service that it invokes, and acts like a proxy that processes and manages incoming and outgoing data at that component related to this flow. Nevertheless, a major boundary of this work is that errors do not propagate correctly in the decentralized setup. Thus, handling errors at build time remains centralized and all errors are propagated back to the client. Further, this approach assumes that each site should have the infrastructure supporting this execution model.

To handle exceptions, the authors in [7] extend their work in [8] to overcome the shortcoming listed above. The exception handling mechanism relies on correct partitioning of the input service specification. As well, the partitions are increased with extra code that aids in the overall exception handling mechanism. However, the proposed scheme is still centralized since fault handlers and compensation handlers for a scope reside in the root partition.

The author in [10] develops A Web Service Composition Tool (WSCT) for specifying a composite service. The WSCT has only to activate the first task in the composite service. The rest of the tasks will be activated by the service providers of the tasks' predecessors when these ones are executed. However, we discern that factual decentralizing mechanism is unclear and details on how to partition the different tasks among service providers are not provided.

To deal with exceptions, the author uses a Non-blocking mechanism. Candidate providers need to monitor the failure of the primary provider. In that case, they run an election protocol to opt for the next highest ranked candidate provider as the new primary provider. Accordingly, the failure of a service provider will not block the progress of a composite service. Though, this approach suffers from (1) a static aspect like BPEL when arranging the candidate providers, (2) a high communication overhead when ensuring the exception handling mechanism, and (3) uncertainty of failure detection: During its execution, a primary provider might be wrongly suspected of having failed. As a result, a new primary provider is elected and carries out the task. Thus, the task is executed twice.

In [12], the author proposes a peer-to-peer approach which is of continuation²-passing style. Knowing the continuation of the current execution, the control can be passed to the appropriate subsequent processing entities without any involvement of a central engine. The approach is based on the assumption that a site can figure out the execution plan that follows up if a message also contains a continuation. Thus, conducting the execution of a process is the sequences of sending and interpreting messages that contain continuations. Nevertheless, the proposed approach is also based on the assumption that each site is apt to figure out the execution plan that follows up, which might be not always affordable. Once more, certain scenarios businesses might want to impose restrictions on access to the data they provide. Hence, this approach

² A continuation is the rest of an execution at a certain point of the execution.

can lead to violation of these data flow constraints since each site has access to all the input data.

If some fault event occurs, a partially executed process must be rolled back. To make possible process rollbacks, two continuations are allied with any particular execution point. The success continuation represents the path of execution towards the successful completion of the process. The failure continuation represents the path of execution towards the proper compensation of committed effects after certain failure events. Further details on how the approach supports process recovery by automatically generating recovery plans into failure continuations are lacking.

As well, software agents have been recognized as a promising technology for managing workflows. For example, in [9], SwinDeW-A (Swinburne Decentralized Workflow with Agents), a service workflow management framework which is based on Peer-to-Peer and agent technologies have been introduced. The multi-agent system consists of a collection of distributed software agents that work conjunctly with each other to afford core workflow services. Each agent is connected with and acts on behalf of a consumer or provider service. The service requestor agent has only to subscribe monitoring data to keep itself up to date about the state of workflow enactment. However, intermediate application data and control are transmitted among relevant service providers agents openly to bring together the execution of the workflow. Therefore, a main drawback is that Monitoring data remains centralized since the agent of the service requestor must keep itself informed about the state of workflow enactment.

Upon a service failure or an unavailable provider exception, the service requestor agent must negotiate with the other available service providers agents which can afford the required service. The involved agents collaborate and negotiate with each other to handle the exceptions and recover automatically. Then, the service requestor agent dynamically contracts a new supplier for the failed service. Accordingly, exception handling mechanism is yet again centralized as monitoring of service enactment refers to nonstop scanning of the state of service by the service requestor agent. As well, we can worry from a scalability issue.

Going over, a key limitation of most of these technologies is that they treat exception handling mechanism in a centralized way. In fact, the disadvantages of this trend are:

- Centralized exception handling clearly represents a performance bottle neck and a single point of failure.
- High communication overhead.

And for the most part, the assumption on which all these proposals are based; to run composite services in a decentralized manner, they assume of course that each site is skilled with the required infrastructure to support this execution model, which might be not for all time affordable.

To deal with these limitations, and based on [8], we propose a decentralized framework for both Web services orchestration and exception handling. Our framework relies on MAS technology. Indeed, software agents are a powerful high-level abstraction for modeling complex software systems through the interactions among the autonomous agents to achieve specific tasks. They had been used in Grid management, mobile computing environment, network monitoring and fault-tolerance in

process automation systems, etc. In this work, we use mobile agents to invoke services, detect occurrence of exceptions and achieve intelligent mechanism to discovery and substitute the failed service by another which has same skills.

3 Background Work

3.1 Decentralized MAS-Based Framework for Composite Web Services Orchestration and Exception Handling

Based on the philosophy that Web services technology and software agent technology have complementary strengths, and that the arrangement of these two skills might build an interoperability environment provided with advanced capabilities, we have conducted careful research on a decentralized Multi-Agent System-based architecture of orchestration Web services and exception handling [14]. The system is composed of four types of agent: User agent (*UsrAg*), BPEL Orchestration Engine Agent (*BPELOEAg*), Web Service Agent (*AgWS*) and User Profile Agent (*UsrPAg*). The following figure (Figure 1.) illustrates the different modules of the system.

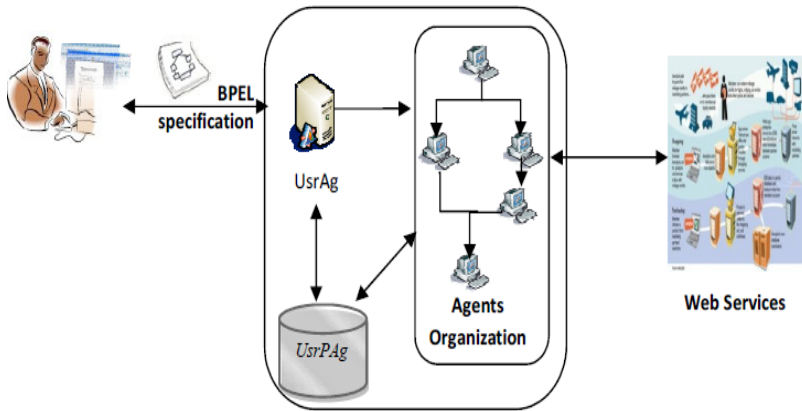


Fig. 1. Architecture of the MAS-based Framework for Decentralized Orchestration [14]

The functional architecture of the system is based on interactions between the agents to execute the BPEL process. The initiator of the conversation is the *UsrAg*. Upon receipt of a BPEL specification from the user, it proceeds to the organization of *BPELOEAg*. For this purpose, we use the algorithm proposed in [8] to partition the BPEL code into a set of BPEL sub-scenarios. The agents' organization means to assign to each code partition a *BPELOEAg*. Each *BPELOEAg* sends to *UsrPAg* a query for information on user's profile and history of contracts. This information is, then, communicated to each *WSAg*. Each *WSAg* performs the service invocation. If the execution succeeds, the result is communicated to the *BPELOEAg*. Otherwise, *WSAg* will analyze its eventual causes. If it is a *BLE (Business Logic Exception)*, it is communicated to the *BPELOEAg*. Else (*System Exception*), *WSAg* proceeds to seek out

for a set of similar services for the failed WS_j . Then, the best service is selected to replace the failed WS_j . In contrast to other works, we do not replace the WS_i in the BPEL specification. Our idea is based on separating the business logic from invoking services and therefore, the system exception management. Thus, the selected service will be invoked by the same agent transparently to the *BPELOEAg*. Each *BPELOEAg* communicates the result of its BPEL sub-scenario execution to the agents that build its network of acquaintances in the organization. The *BPELOEAg* which is the root of the organization communicates the final result to the *UstrAg* which communicates it, in turn, to the user. The user evaluates the result provided by each service, and sends its evaluation to the *UstrAg* which communicates it to the *UstrPAg* in order to bring up to date its base. Further details have been reported in [14].

3.2 Mobile Agents Technology

A mobile agent is a program that represents a user in a computer network and can migrate autonomously from node to node, to perform some computation on behalf of the user. Applications can inject mobile agents into a network, allowing them to roam the network, either on a predetermined path or one that the agents themselves determine based on dynamically gathered information. Having accomplished their goals, the agents can return to their home site to report their results to the user [11]. The goal is to improve application performance which mobile agents do by letting the application move its functionality dynamically between machines. Mobile agents offer a uniform approach to handling code and data in a distributed system [5]. There are at least seven main benefits, or good reasons, to start using mobile agents [4]:

- They reduce the network load.
- They overcome network latency.
- They encapsulate protocols.
- They execute asynchronously and autonomously.
- They adapt dynamically.
- They are naturally heterogeneous.
- They are robust and fault-tolerant.

Thus, the mobile agent technology has drawn further attention as a new distributed paradigm.

4 Overall Mobile Agents-Based Framework

Starting from our claim that, without being based on the assumption that each site is skilled to accomplish its sub-process, high level environments providing the support to execute decentralized BPEL processes are needed and detecting, for the best of our knowledge, that there is no unifying approach today allowing to do this in a proper manner, we think that mobile agents are as much as skilled to do this.

For that purpose, we extend our previous framework presented in Section 3 by the use of mobile agents *WSMAg* as alternatives for agents used to invoke Web services (*WSAg*). Figure 2. depicts the framework architecture extending the one proposed in [14] by supporting mobile agents to execute the invoking activities.

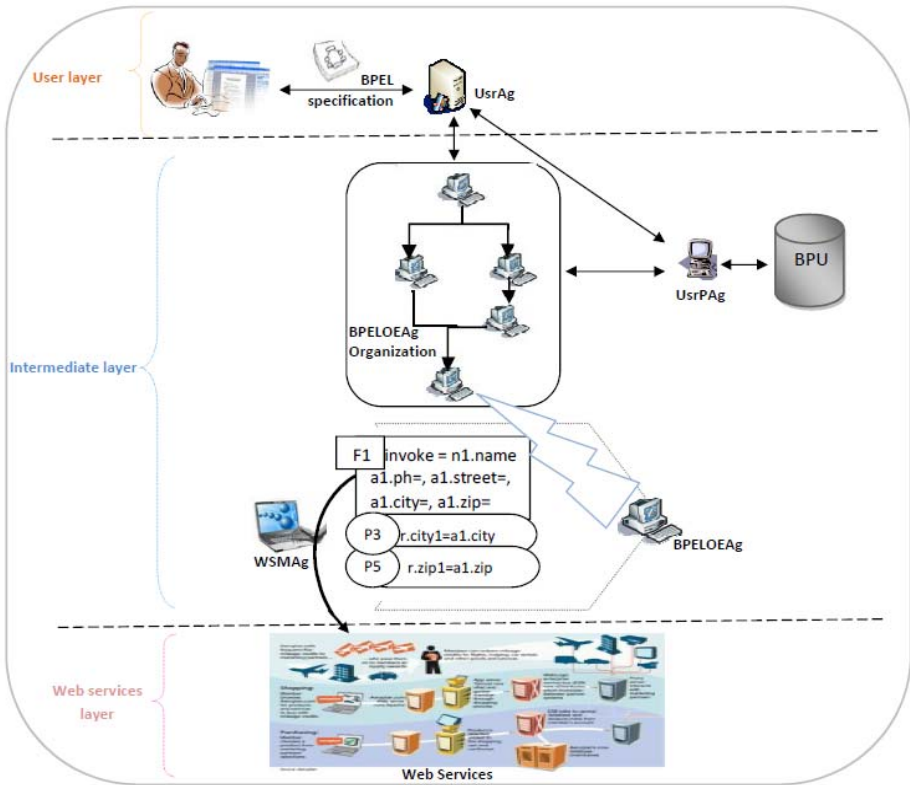


Fig. 2. Architecture of the Enhanced MAS-based Framework for Decentralized Orchestration and Exception handling by means of the Mobile Agents Technology

For better readability, the architecture is organized on three layers: (i) the User layer, (ii) the Intermediate layer and (iii) the Web services layer. This architecture gives also a zoom on what a partition could envelop (being based on [8] with our *BPELAg* support), such that each partition has exactly one fixed node (*receive*, *reply* and *invoke* nodes are designated as fixed nodes) and zero or more portable nodes (all other activities are considered as portable nodes).

We walk now layer by layer through the architecture of Figure 2 to illustrate how the whole process execution is applied. Similarly to our previous system, the current proposal is composed of four agent types: User agent (*UsrAg*), BPEL Orchestration Engine Agent (*BPELOEAg*), Web Service Mobile Agent (*WSMAG*) and User Profile Agent (*UsrPAg*). Below, we describe separately each entity:

User Layer: With it a process or service is started. It is the interface between the user and the system. It encloses the *UsrAg* which offers to the user all the necessary functionalities so that he specifies his BPEL process *P*, and turns to him the result of the execution of *P*. Indeed, it organizes the *BPELOEAg* agents while being based on the business logic described in *P*. This organization relies on a process of partitioning of

the BPEL scenario into a set of BPEL sub-scenarios. Each sub-process is associated with a *BPELOEAg* agent.

Intermediate Layer: It surrounds the *BPELOEAg* Organization, the *UstrPAG* and the *WSMAG*. The *BPELOEAg* checks the role of an execution engine of a BPEL sub-process. In fact, it distributes and monitors the implementation of its BPEL sub-scenario. It associates to each service invocation an agent *WSMAG*. For each *invoke* activity, a plug-in is carried out in order to create a mobile agent of service *WSMAG*. The plug-in allows launching a *WSMAG* with all the necessary skills and knowledge using a predefined skeleton code. The skills include the discovery of web services (*Discovery* ()), invoking a service (*Invoke* ()), exception (system and business logic) handling (*CheckException* ()) and the substitution of a failed service by another (*Selection* ()). Knowledge is represented by the user profile, the constraints of QoS and history of contracts with Web services providers.

The *UstrPAG*'s role consists on the user profile base management. It gets the information related to the user as its geographical location and his interests. It also maintains the history of services whose execution fails and those whose execution succeeds. These informations will be used by *WSMAG* for more secure and reliable discovery and selection of Web services.

The *WSMAG*, the crucial agent in our system, has to invoke the service. Figure 3. introduces modules that make up the internal architecture of a *WSMAG* (Note that its architecture preserves the same modules as the *WSAg* in [14]):

- *WS invocation module*: This module is responsible of the service. The result could be:
 1. An answer of the successful execution of the invoked service. Then, it will be communicated to the *BPELOEAg* so that it achieves the whole execution of its BPEL sub- scenario.
 2. A System Exception (*SE*) or a Business Logic Exception (*BLE*). If this is a system exception, the invoked service parameters will be communicated to the *discovery module*. Otherwise (*BLE*), it will be communicated to the *BPELOEAg*.
- *Discovery module*: If a *SE* is raised, this module is triggered with the service parameters communicated since the invocation module. It proceeds to generate a list *L* of the equivalent services to the invoked WS_i , as well as their eventual QoS.
- *Selection module*: This module is activated on the coming of a message from the discovery module. This message holds the list *L*. The selection module picks out the best service by ordering *L* according to the following informations : i) QoS of equivalent services, measured by WSLA, ii) informations related to providers of the global BPEL process and iii) Constraints of QoS integrated into the BPEL specification. These informations may be associated to services providers, as the number of contracts with the user and a rate of success for each one, the confidence degree associated by the user to the provider, the provider's geographical location, etc. This information is usually drawn from the history of interactions between the user and the providers. The selected service will be transmitted to the *WS invocation module* for running. While migrating or execution on a site, if the *WSMAG* is lost, a clone of it is automatically relaunched by the *BPELOEAg*.

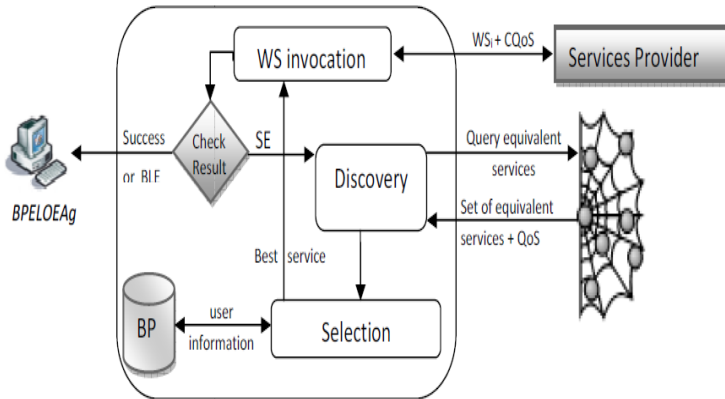


Fig. 3. WSMAg architecture

Web Services Layer: It provides Web Services to be accessed by our framework on the internet.

5 Conclusion

In this paper, we introduced our research efforts to address exception handling issues in decentralized orchestration of composite Web services. These issues are critical for generating straight decentralized Web service business processes. We elucidated why the use of mobile agents could be a powerful technique to address the problems of executing BPEL sub-scenarios among a decentralized environment. As a future work, a first prototype implementation is projected to sustain. As well, further research is needed to consolidate the conceptual foundations of this approach.

References

1. Suna, A.: CLAIM et SyMPA: Un environnement pour la programmation d'agents intelligents et mobiles. Thesis, Université Paris 6 - Pierre et Marie Curie Laboratoire d'Informatique de Paris 6 (décembre 2005)
2. Dellarocas, C., Klein, M.: A knowledge-based approach for handling exceptions in business processes. *Information Technology and Management* 1(2), 155–169 (2000)
3. Karelitis, C., Vassilakis, C., Rouvas, E., Georgiadis, P.: Exception Resolution for BPEL Processes: a Middleware-based Framework and Performance Evaluation. In: *Proceedings of the tenth International Conference on Information Integration and Web-based Applications & Services 2008 iiWAS 2008* (2008)
4. Lange, D.B., Oshima, M.: Seven Good Reasons for Mobile Agents. *Communications of the ACM* 42(3), 88–89 (1999)
5. Milojicic, D.: Trend Wars: Mobile Agent Applications. *IEEE Concurrency* 7(3), 80–90 (1999)

6. Chafle, G.B., Chandra, S., Karnik, N., Mann, V., Nanda, M.G.: Improving Performance of Composite Web Services Over a Wide Area Network. In: 2007 IEEE Congress on Services SERVICES 2007 (2007)
7. Chafle, G., Chandra, S., Kankar, P., Mann, V.: Handling Faults in Decentralized Orchestration of Composite Web Services. In: Proceedings of the 3rd International Conference on Service Oriented Computing (ICSOC 2005), Amsterdam, Netherlands (December 2005)
8. Chafle, G.B., Chandra, S., Mann, V., Nanda, M.G.: Decentralized orchestration of composite web services. In: Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & WW Alt. 2004 Posters, New York, USA, May 19 - 21, pp. 134–143. ACM, New York (2004)
9. Yan, J., Yang, Y., Kowalczyk, R., Nguyen, X.T.: A Service Workflow Management Framework Based on Peer-to-Peer and Agent Technologies. In: Proceedings of the Fifth International Conference on Quality Software, QSIC 2005 (2005)
10. Ye, X.: Towards a Reliable Distributed Web Service Execution Engine. In: IEEE International Conference on Web Services, ICWS 2006 (2006)
11. Qu, W., Shen, H., Defago, X.: A Survey of Mobile Agent-Based Fault-Tolerant Technology. In: Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, PDCAT 2005 (2005)
12. Yu, W.: Peer-to-Peer Execution of BPEL Processes. In: The 19th International Conference on Advanced Information Systems Engineering, CAiSE 2007 (2007)
13. Yu, W., Yang, J.: Continuation-Passing Enactment of Distributed Recoverable Workflows. In: 22nd Annual ACM Symposium on Applied Computing (SAC 2007), Seoul, Korea, March 11 - 15 (2007)
14. Brahmi, Z., Ilahi, M., Gammoudi, M.M.: Decentralized MAS-based Framework for Orchestration of Composite Web Services and Exception Handling. In: 2nd International Conference on Information Systems and Economic Intelligence, SIIE 2009 (2009)

Multi-objective Analysis on Optimization of Negotiation Support

Yu-Teng Chang^{1,2}, Chih-Yao Lo^{1,2}, Ping-Chang Chen², and Shu-Huei Han²

¹ School of management, Huazhong University of Science & Technology

² Department of Information Management, Yu Da University,
Miaoli County, Taiwan 361, R.O.C.

{cyt, jacklo, pcchen, 96310024}@ydu.edu.tw

Abstract. With the emergence of global marketing, the development of sales channels becomes critical in terms of competitive advantages. In fact, the co-competition between companies evolves to be increasingly complex. A co-operation project may involve more than just two companies, and sometimes three or more companies are engaged in the decisions in association with co-competition. Meanwhile, each company may have a large number of strategic options. Therefore, this paper applies the game theory and combines it with the model designed with multiple agents to identify the optimal strategic option acceptable by every participant in the game. A renegotiation system with multiple agents is designed based on the method proposed in this paper so that companies can quickly and effectively identify an outcome acceptable by all the competitors. The agent mechanism not only enhances the efficiency of solutions for equilibrium solutions, but also resolves the problems that no solutions can be derived against the endless loops in the strategic games. It is also able to derive solutions in the most efficient way for the multiple-player, multiple-strategy not presentable on a two-dimension matrix. This approach greatly shortens the time required to compute solutions in the game and enhances negotiation efficiency, in order to generate the strategic combinations with maximum returns for all the competing companies.

Keywords: Strategic Games, Conflict Resolution, Multiple Agents.

1 Introduction

1.1 Research Background and Motives

Decisions over marketing channels are one of the key decisions for any company. Distribution strategies are not only highly relevant to other marketing strategies, but also represent the long-term commitments a company has with its distributors. The agency relationship between manufacturers and distributors is based on the different perspectives and roles assumed by both parties. The power issues surrounding channels lead to constant conflicts and co-competitions. Manufacturers often question distributors on the ground that distributors keep market information to themselves, are lacking in brand loyalty, delay payments and return goods at large quantities. On the other

hand, distributors are dissatisfied with manufacturers who constantly raise prices, threaten to terminate distribution relationships, or do not provide goods on a timely basis. There are mistrust and conflicts between both parties. Both parties have conflicts on the expected prices of targeted goods and therefore, they bargain and negotiate. The process of renegotiations and communication is fraught with uncertainties. It is very difficult to reach optimal decisions in such a complex situation. The relationship between manufacturers and distributors is both cooperative and competitive. Companies often have to choose between collaboration and competition with other companies in order to improve profits. In face of other companies with which cooperative ties are sought, a company should know how to select the strategy that best benefits itself. When it comes to details in cooperation, a company should know how to maximize the total profits. Therefore, this paper examines the problems association with conflict resolutions for corporate strategies by establishing multiple-player, multiple-strategy game renegotiation system. It is hoped that this model can serve as a reference for archetype for similar problems.

2 Literature Review

The core element of this paper consists of two parts. The first part is a discussion of bargaining games and strategic moves games in game theory. The second part is a conflict resolution model in the environment of multiple agents. Below are separate discussions of these two parts.

2.1 Bargain Game Theory

Bargain theory, as part of game theory, examines how interests are allocated via renegotiations when there are only two participants. When it comes to economical analysis, the focus on bargaining process to determine the allocations of interests for both economic entities is a long-standing practice. Generally speaking, there are two approaches to these problems in addition to classical bargaining theory. One is Nash Axiomatic approach and the other is Sequential bargaining game [1]. The characteristics of basic bargaining process are as follows:

1. With two or more participants involved: Markets are composed of a group of buyers and sellers. When the transactions are ongoing, neither side is aware of the demand and supply from the other side. Therefore, both sides gather and bargain in order to identify the most appropriate transaction price.
2. Existence of obvious or potential interests: The premise of any renegotiation is that both the buyer and the seller have their own stance and expectations. They bargain and renegotiate in order to meet their respective goals. Therefore, compromise from one party means benefits for the other.
3. Dependence upon each other: In the process of bargaining and renegotiating, even if participants cannot reach their targets, they will endeavor to achieve 100% satisfaction. Other parties think the same. However, the overly focus on own satisfaction will cause complaints from the other party. No matter what the trading system is, it is necessary to explore the bottom line from the other party. Therefore, the bargaining parties are in fact inter-dependent.

4. Willingness to reach solutions by working together: Bargains and renegotiations are processed with the aim to reach transactions for both buyers and sellers. The shared purpose of bargains and renegotiations is to reach a balanced distribution of interests. This is the agenda, as well as common intention for both parties to reach solutions. Without this, it cannot be a successful bargain.

2.2 Strategic Moves Game

Strategic moves mean the use of tactics to change others' convictions, thoughts or behaviour. The restriction of own actions purposefully can enhance own benefits. Dixit & Skeath [2] believe that unless one party is authoritative or advantageous in a game, each participant has the desire to control the rules in order to fight for their own interests. This type of controlling actions, no matter on or under the table, can be called "strategic moves".

Credibility is the key to strategic moves, which can only work when they are established on the basis of mutual trust. Meanwhile, strategic moves aim to control the game rules. Priorities of actions and the returns originally expected may be different from the previous assumptions due to the occurrence of strategic moves [3]. Strategic moves can be classified into three types: commitment, threat and promise [4]. The use of these three types of strategic moves is to turn the situation from unfavorable to favorable to the action taker. The premise is to convince other parties into believing the terms, conditions or proposals suggested by the action taker [2].

In addition to credibility, strategic moves have to be observable and irreversible. If manufacturers are not sufficiently sensitive to detect the decisions by distributors or simply refuse to discover these decisions, they of course cannot take responsive actions. In such an instance, the actions by both parties become unrelated. If manufacturers have to secure certain distributions (irreversible), they have to take actions in response to distributors and adjust their own actions accordingly for maximum profits. There are two types of strategic actions [2]: (1) unconditional; (2) conditional.

2.3 Conflict Resolution Model and Structure

Conflict resolution models can be classified as calculation confliction resolution model, human conflict resolution model, game theory [5], CEF and collaborative design system [6]. There are a wide range of strategies and models in literature. However, there are few discussions on linear planning in quantitative analysis. Also, bargains and multiple agents can be used to develop solutions to conflict resolution models.

1. Bargains

Sycara proposes a distributed system, PERSUADER, to generate solutions to renegotiations via bargains between agents. The bargaining procedures include the identification of purposes of other agents and change purposes accordingly in order to avoid conflicts, creation of some techniques to gain trust from agents, inference and modification of trust of other agents in order to reach bargains. PERSUADER is based on CEF AND CDE.

2. Multiple agents

A multiple agents system is an effective method to resolve complex systems [7-8]. It uses parallel distribution processing techniques and modular designs to

divide a complex system into independent agent sub-systems. Solutions to complex problems are derived via co-competition between agents [9]. Each agent only possesses incomplete information and ability for the tasks they are required to complete, given that the statistics and resources of a multiple agent system are distributed. In fact, the concept of their tasks is local, not an across-the-board control system.

3 Research Design

3.1 Research Methods and Procedures

This paper first delves into the respective business environments of individual companies and examines the strengths and weaknesses of these companies. The results can serve as an important reference for the design of an agent knowledge base for respective companies. Afterwards, the satisfactions and expectations from both manufacturers and distributors are the major elements in the design of the strategic moves game in compliance with the purposes of this paper. A matrix of returns to strategic moves from companies is then established before the introduction of the concept of multiple agents and conflict resolutions to renegotiations. The strategic renegotiation system of multiple agents for multiple-players, multiple-strategies is constructed as Figure 1 shows.

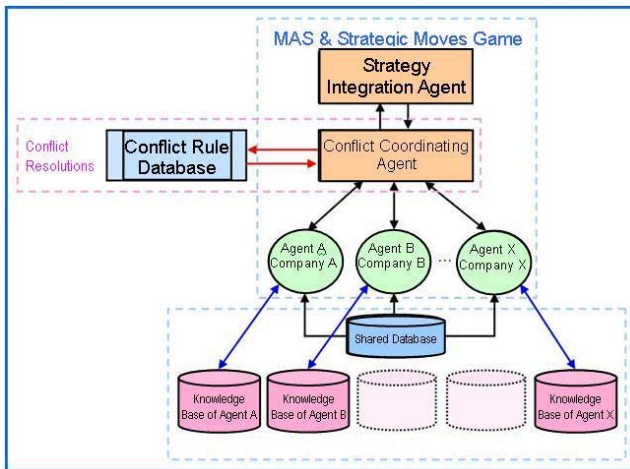


Fig. 1. Research Structure of Multiple-players and Multiple-strategies Renegotiation System

3.2 Formulation of Renegotiation Strategies

The level of satisfactions and expectations is the most important factor to the success of any renegotiations. Therefore, only by understanding the preferences and expectations of participants can only accurately bargain and renegotiate on the behalf of the participants.

After each company discusses with each other, they can seek the final integration by using the conflict resolution agent system designed by this paper, in order to derive the optimal strategic combinations with maximum returns and acceptable by all the companies working together for a final goal. Figure 2 illustrates the procedural structure.

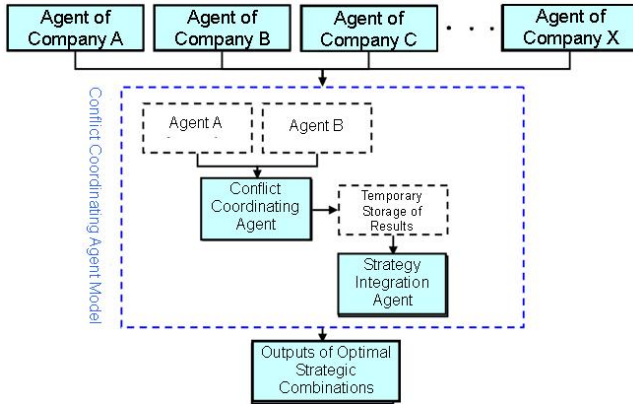


Fig. 2. Actual Processes of Multiple Agents

3.3 Conflict Resolutions of Multiple Agents

This paper assumes that each of the participating company takes up two roles in a zero-sum game of two parties, in order to facilitate analysis. First of all, this paper selects a participating company and defines it as Agent A and assumes Agent A has two strategies, to collaborate and not to. Similarly, a participating company is defined as

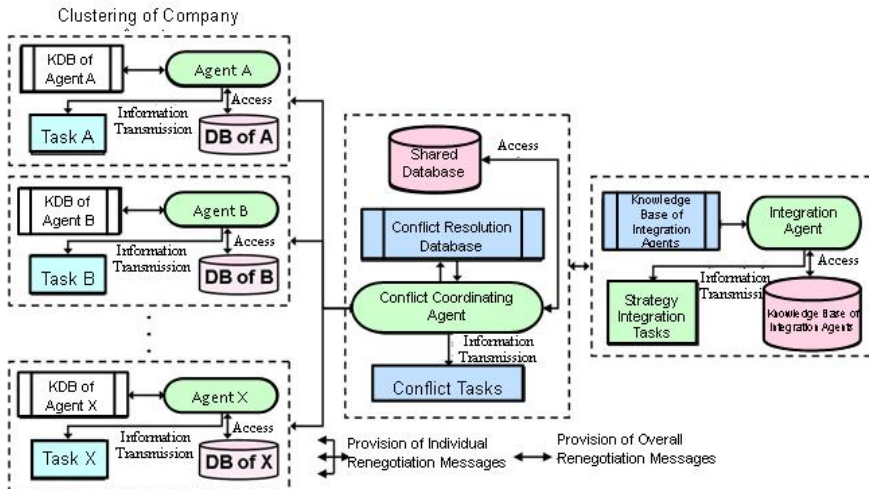


Fig. 3. Structure of Multiple Agents

Agent B, who also has two strategies, to collaborate and not to. The renegotiations from both parties are subject to internal and external environmental factors. Although participants are highly willing to discuss, the terms and conditions proposed by one party will be very different from what is expected by the other party due to considerations of interests. However, given the premise of consent for renegotiations, it is an optimal game equilibrium. Therefore, another agent is required as a coordinator (the third party), a.k.a. conflict coordinating agent to resolve conflicts. Figure 3 depicts all kinds of ideas.

The steps are as follows:

- (1) Design of a multiple-agent mechanism—Corporate agent, conflict coordinating agent and strategic integrator agent represent the company and third party coordinator.
- (2) Solutions to respective roles taken by companies given the strategic returns (rankings) under the demand restrictions concerning.
- (3) Company agents convey the value of strategic returns to the conflict coordinating agent at the same time.
- (4) Whether strategic returns values received by the conflict coordinating agent conflict with each other.
- (5) In case of any conflict, resolution rules are selected from the decision rule databank. Resolutions are implemented and results are feedback to the company agent mechanism so that conflict coordinating agents can determine whether individual needs are met.
- (6) If the results implemented by conflict coordinating agents cannot satisfy the respective needs of any agent, another resolution rule will be selected from the decision rule databank for conflict resolutions for a new result. The procedure will continue until a certain result is able to satisfy the needs of all agents. The system will derive a solution from probabilities for the actions implemented for a certain number of times.

The pairing of all the company agents and the integration actions taken by the strategic integration agents will produce results, as the equilibrium strategic combinations of this multiple-players, multiple-strategy game.

3.4 System Development and Result Assessment

This paper derives solutions with the agent mechanism for the multiple-players, multiple-strategy game by following the abovementioned steps. To allow each participating company to be engaged in strategic renegotiations more easily, this model adopts a virtual network to stimulate the scenarios. In the case where there are two manufacturers and one distributor, this paper uses Visual Basic 6.0 to write and develop the system program for case stimulations. The steps and flows of the system are as follows.

4 Experiment Design

4.1 Steps of Realistic and Strategic Renegotiations

This paper defines more than one strategic coordinator. The initial action selects two companies and seeks the optimal strategic combinations with results integrated by the

agents. The strategies adopted are based on the actual situations of companies. Figure 4 illustrates the steps.

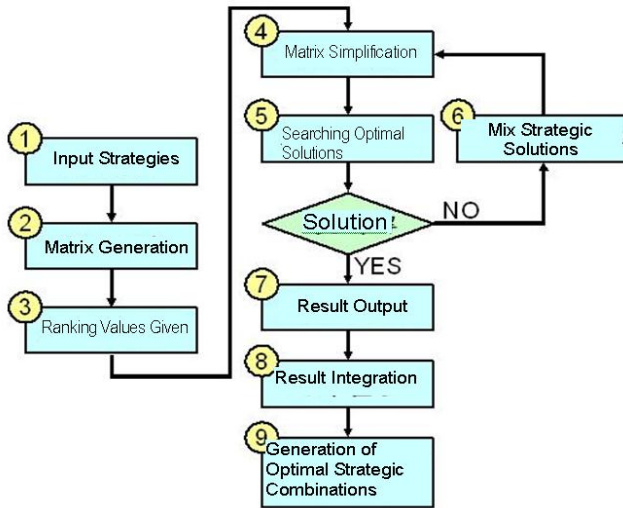


Fig. 4. Flows to Derive Solution

4.2 Description of Scenarios for Case Stimulations

According to the abovementioned multiple-agents strategic renegotiation model, this paper performs scenario analysis to address the issues brought forward. The scenario assumes the following:

In a given supply chain, there is no possibility to increase any profits with the current terms and conditions of cooperation. Faced with competition from foreign manufacturers, the distributor wishes to adjust the purchase ration and reduces the expenses associated with non-promotional products. Manufacturers hope to enhance their respective profits via different cooperation proposals. This paper assumes there are two manufacturers and one distributor. They are examining different (i.e. more advantageous) strategic cooperation options. Before the proposals, both the distributor and manufacturers have to make relevant assessments. In this scenario, the manufacturers are at a disadvantage. Therefore, manufacturers can work toward two directions. One is to collaborate with the other manufacturer. Since their products are not highly substitutable but interdependent daily goods, the two manufacturers can work together in certain ways. It is hoped that such cooperation can strengthen their market position to compete with foreign manufacturers and increase overall profits. The other alternative is to compete with each other. However, purchasing power has declined due to an economic recession and virtually everybody is concerned about expenses cuts (in the form of bargains and reduction in the purchase of unnecessary goods). Although both manufacturers have certain branding power, working with each other may achieve higher profits via promotional campaigns or appropriate price cuts. Given the space of shelves is limited, the distributor is concerned with how to ration allocations of

procurements from the two manufacturers or whether it is a good idea to give away more benefits to one of the two manufactures in order to achieve higher returns. Below is a simple description.

Distributor:

- A. Maintenance of the original cooperation pattern and allocation of procurements based on the proposals from respective manufacturers.
- B. Granting one of the two manufacturers with a higher ratio (if that manufacturer offers better terms).

Manufacturer:

- A. Working with the other manufacturer for joint promotion campaigns and reductions in distribution costs (This paper focuses on competition due to a large number of considerations required for this possibility.)

Therefore, this paper infers the following cooperation scenarios. Firstly, both manufacturers and the distributor in question all work together. This is because Manufacturer A and Manufacturer B believe their cooperation cannot only benefit their own incomes, but also enhance their bargaining power with the distributor. The distributor may also gain from better profits with the integration of manufacturers. It is believed that this scenario is a feasible collaboration. The distributor who originally had a dominant position will perhaps see their bargaining power equal to that of manufacturers. Alternatively, manufacturers may be in an advantageous position to renegotiate with the distributor, as shown in Figure 5.

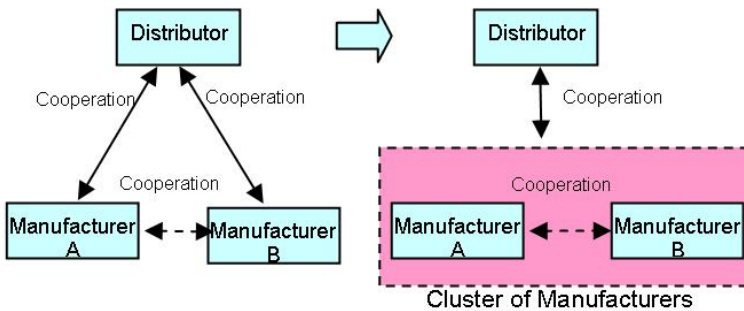


Fig. 5. Stimulated Scenario -1

Secondly, if manufacturers do not think it is possible for them to cooperate but also expect distributors to offer better terms, they have to come up with a more attractive proposal to distributors than that offered by competing manufacturers given limited resources. It is to obtain better shelf spaces or more purchases. However, without knowing all the proposals from competing manufacturers, they have to list out all their feasible cooperation proposals, as outlined in Figure 6. Given the limited time and resources, this paper examines this scenario.

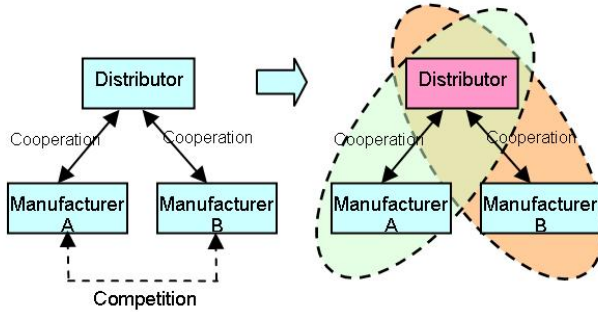


Fig. 6. Stimulated Scenario-2

5 Validation and Results

5.1 Results of Case Stimulations

This paper examines the issues surround the relationship between manufacturers and distributors so that companies can maintain certain advantages in renegotiations. The advantages include private information that companies do not wish to disclose. Without considering the possibility of costs in association with shortage of delivery, this paper derives the maximum expected profits for distributors and manufacturers based on the following formula.

Maximum expected profits of Distributor

$$= \text{Max} \left\{ \left((P_i - C) \times Q - (C_{Ti} + K) \right) \times \frac{H}{R_i} - C_{\text{other}} \right\}$$

Let P_i = Sales price

Q = Quantity of products (This sample assumes the expected sales equals to procurement amount.)

C = Procurement cost (i.e. out-of-factory unit price from Manufacturer)

C_{Ti} = Transportation cost

K = Ordering cost

$\frac{H}{R_i}$ = No. of Delivery/Good Replenishments

C_{other} = Other costs (This example assumes it to be private information of Manufacturer.)

Maximum expected profits of Manufacturer

$$= \text{Max} \left\{ \left(P_j \times Q + K \right) \times \frac{H}{R_i} \right\} - TC_m$$

Let P_j = Out-of-factory unit price

Q = Quantity of products

TC_m = Total fixed costs of Manufacturer (This example assumes it to be private information of Manufacturer.)

K = Cost of ordering for Distributor

$\frac{H}{R_i}$ = No. of Delivery/Good Replenishments

This paper lists all the feasible strategies for manufacturers and distributors based on the variables in the above formulae.

Manufacturer A suggests that if distributors can prioritize their products on shelf, they can give up to 3% discounts (only applicable to the unit price of NT\$250 and above). Therefore, the strategic proposal from Manufacturer B is as follows (Table 1):

Table 1. Feasible Variable Combinations for Manufacturer A

Out-of-factory unit price (NT\$/unit)	Quantity (units)	Deliveries (times/p.a.)	Ordering cost (NT\$/times)
250	100	10, 12	150, 100
230	200	4, 6	200, 180
200	300	4, 9	200, 150

Similarly, Manufacturer B suggests that if distributors can prioritize their products on shelf, they can give up to 5% discounts (only applicable to the unit price of NT\$350 and above). If the quantity of procurement from a distributor exceeds 350 units for 8 deliveries, ordering cost is waved. Therefore, the strategic proposal from Manufacturer B is as follows (Table 2):

Table 2. Feasible Variable Combinations for Manufacturer B

Out-of-factory unit price (NT\$/unit)	Quantity (units)	Deliveries (times/p.a.)	Ordering cost (NT\$/times)
350	150	10, 12	120, 100
310	200	6, 9	150, 110
240	250	4, 6	200, 150

The considerations for Distributor C consist of two parts. Firstly, Distributor C prices the product made by Manufacturer A at NT\$335 per piece. The previous policy for the procurement of the product made by Manufacturer A is as follows: out-of-factory unit price NT\$230, each procurement for 200 units at six deliveries and the procurement cost of NT\$180. Secondly, Distributor C prices the product made by Manufacturer B at

NT\$400 per piece. The previous policy for the procurement of the product made by Manufacturer B is as follows: out-of-factory unit price NT\$240, each procurement for 250 units at four deliveries, procurement cost at NT\$200 and the transportation cost at NT\$200/time. The transportation cycle is 4 times per year (every three months) and 6 times per year (every two months) for different products. Based on the above details, Distributor C comes up with the following strategic proposals (Table 3):

Table 3. Feasible Variable Combinations for Distributor C

Product of Manufacturer A (No. of units)	Product of Manufacturer B (No. of units)	No. of deliveries for good replenishments (No. of times/p.a.)	Total transportation costs (NT\$/p.a.)
100, 200	150, 200, 250	4	800
100, 200	150, 200, 250	6	1000 (discount NT\$200)
300	150	4	800
300	150	6	1000 (discount NT\$200)

Therefore, it is possible to calculate the previous maximum expected profits for both manufacturers and Distributor C.

- For products of Manufacturer A (product a):
Maximum expected profits of distributors (product a) are as follows:

$$\begin{aligned} & \text{Max} \left\{ \left((P_i - C) \times Q - (C_{Ti} + K) \right) \times \frac{H}{R_i} - C_{\text{other}} \right\} \\ & = \text{Max} \left\{ \left[\left((335 - 230) \times 200 - (200 + 180) \right) \times 6 + 200 \right] - C_{\text{other}} \right\} \\ & = \text{Max} \{ 123920 - C_{\text{other}} \} \end{aligned}$$

Maximum expected profits of Manufacturer A

$$\begin{aligned} & \text{Max} \left\{ \left(P_j \times Q + K \right) \times \frac{H}{R_i} \right\} - TC_m \\ & = \left[(230 \times 200 + 180) \times 6 \right] - TC_m \\ & = 277080 - TC_m \end{aligned}$$

- For products of Manufacturer B (product b):
Maximum expected profits of distributors (product b) are as follows:

$$\begin{aligned} & \text{Max} \left\{ \left((P_i - C) \times Q - (C_{Ti} + K) \right) \times \frac{H}{R_i} - C_{\text{other}} \right\} \\ & = \text{Max} \left\{ \left[\left((400 - 240) \times 250 - (200 + 200) \right) \times 4 \right] - C_{\text{other}} \right\} \\ & = \text{Max} \{ 158400 - C_{\text{other}} \} \end{aligned}$$

Maximum expected profits for Manufacturer B

$$\begin{aligned} & \text{Max} \left\{ \left(P_j \times Q + K \right) \times \frac{H}{R_i} \right\} - TC_m \\ & = \left[(240 \times 250 + 200) \times 4 \right] - TC_m \\ & = 240800 - TC_m \end{aligned}$$

This paper collates all the previously mentioned strategic proposes and summarizes them into a list of strategic discussions for companies. When ranking these priorities, companies will have different emphases due to their different roles. The key concerns for distributors are the historical sales percentages for the manufacturers and relative profitability compared to other international brands. On the other hand, the key considerations for manufacturers are their own capacity and manageable delivery and turnovers.

6 Conclusions and Contributions

This paper analyzes strategic applications and observes bargains and renegotiations among multiple companies. This approach allows objective attitudes for observation, but may also lead to personal and subjective interpretation of behaviour bias of participants. In order to reduce such errors and enhance the accuracy of final results, this paper assigns companies to input strategies and assess importance themselves. In the past, the research approach applying game theory has to specifically indicate the returns of corresponding strategies. However, the method used by this paper only describes the strengths and weaknesses of respective strategic combinations. The concept of sequencing is combined with the mechanism of multiple agents so that the games can proceed with better efficiency.

The case studies modelled by this paper can serve as a reference to multiple-player strategic move games. However, considering that the variables are private information companies do not wish to disclose, this paper adopts a semi-automatic model with partial involvement from these companies in strategic renegotiations. In fact, this model provides a recommendation for a meaningful strategic combination as assistance to multiple companies in bargains and renegotiations. Therefore, it is suggested that future studies can examine the relationship between manufacturers and distributors, or even multi-stage supply chain, with a set of more comprehensive variables, in order to produce renegotiation results closer to the thoughts of each participant and the improve the feasibility of optimal and equilibrium strategic combinations.

References

- [1] Agrawal, D.: Effect of Brand Loyalty on Advertising and Trade Promotions: A Game Theoretic Analysis Empirical Evidence. *Marketing Science* 1, 86–108 (1996)
- [2] Dixit, A., Skeath, S.: *Games of Strategy*. W W Norton & Co Inc., Newyork (1999)

- [3] Bergen, M., Dutta, S., Shugan, S.M.: Branded Variants: A Retail Perspective. *Journal of Marketing Research* 33, 9–19 (1996)
- [4] Roger Myerson, B.: *Game Theory: Analysis of Conflict*. Harvard University (1991)
- [5] Feldman, D.: A Taxonomy of Intergroup Conflict Resolution Strategies. In: *The 1985 Annual Conference on Developing Human Resources* (1985)
- [6] Durfee, E.H., Lesser, V.R.: Using Partial Global Plans to Coordinate Distributed Problem Solvers. In: *Proceedings of IJCAI 1987*, pp. 875–883. IJCAI (1983)
- [7] Gasser, L., Braganze, Herman, N.: MACE: A Flexible Testbed for DAI Research. In: Huhns (ed.) *Distributed Artificial Intelligence*, Morgan Kaufmann, Pub Inc., Los Altos (1987)
- [8] Georgeff, M.: Communication & Interaction in Multi-agent Planning. In: *Proceedings of AAAI 1983*, pp. 125–129. AAAI, Menlo Park (1983)
- [9] Gevins, A.S.: Overview of the Human Brain as a Distributed Computing Network. In: *IEEE International Conf. Computer Design: VLSI in Computers* (1983)

Rough Set Based Personalized Recommendation in Mobile Commerce

Lei Shi, Li Zhang, Xinming Ma, and Xiaohong Hu

College of Information and Management Science, Henan Agricultural University,
Zhengzhou, 450002, China

sleicn@126.com, zhangli200312@eyou.com,
xinmingma@126.com, wd9702@163.com

Abstract. Mobile commerce (M-commerce) combines the advantages of electronic commerce with the mobility and freedom of wireless devices such as cellular telephones and PDAs. As M-Commerce become more and more prevalent to people, it is very critical to provide the right information to the right customers. In this paper, a novel personalized recommendation method based on tolerance rough set is proposed to help customers purchase needed products conveniently. Tolerance rough set is used to deal with latent information and capture customer preference according to both the customer's interests and his current context.

1 Introduction

Mobile commerce (M-commerce) is the buying and selling of products and services anytime from anywhere through wireless handheld devices such as mobile phone and personal digital assistant (PDAs). Compared with electronic commerce, M-commerce presents many advantages, e.g. ubiquity, accessibility, personalization and convenience [1], [2]. With the rapid development of M-commerce, it provides more and more choices for users now. However, users usually get lost in the vast space of product information and can not find the products they really want. Thus, it is very critical to provide the right information to the right customers as M-commerce become more and more prevalent to people [3], [4].

In this paper, a novel personalized recommendation method based on tolerance rough set is proposed to help customers purchase needed products conveniently. Tolerance rough set is used to deal with latent information and capture customer preference according to both the customer's interests and his current context. The tolerance rough classes of products currently browsing by customer are constructed to discovery the relationship among different products and obtain latent associated products in the transaction database.

The rest of the paper is organized as follows. Section 2 introduces the rough set theory briefly. Section 3 presents the novel personalized recommendation model based on rough set in detail. Finally, section 4 concludes the paper.

2 Rough Set Theory

2.1 Rough Set

The rough set theory is a powerful mathematical tool for modeling inexact, uncertain or vague knowledge [5]. Since it was introduced by Pawlak in the early 1980s, rough set has extracted a lot of attention from theory and been applied in many fields [6]. In rough set theory, an information system, which is also called a decision table, is defined as $S = (U, A \cup D, V, f)$, where U is the finite set of objects, A a collection of condition attributes, D a collection of decision attributes, V a set of values of attributes in A and $f : A \rightarrow V$ a description function. For any $R \subseteq A$, there is an equivalence relation $I(R)$ as follows [7]:

$$I(R) = \{(x, y) \in U^2 \mid \forall a \in R \ a(x) = a(y)\} \tag{1}$$

If $(x, y) \in I(R)$, then x and y are indiscernible by attributes from R . The equivalence classes of the R -indiscernibility equivalence relation $I(R)$ are denoted $[x]_R$. For any concept $X \subseteq U$ and attribute subset $R \subseteq A$, X could be approximated by the R -lower approximation and R -upper approximation as Figure 1.

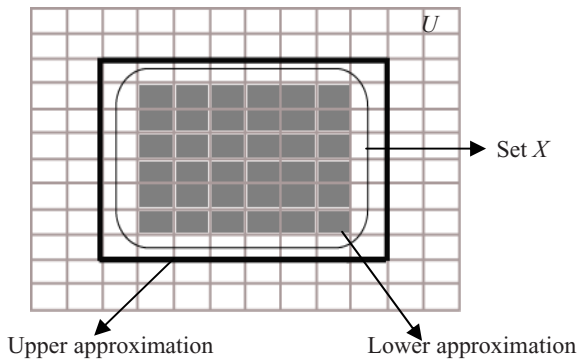


Fig. 1. Lower and upper approximations of a rough set

The R -lower approximation of X is the set of objects of U that are surely in X , defined as:

$$\underline{R}X = \{x \in U \mid [x]_R \subseteq X\} \tag{2}$$

The R -upper approximation of X is the set of objects of U that are possibly in X , defined as:

$$\overline{R}X = \{x \in U \mid [x]_R \cap X \neq \emptyset\} \tag{3}$$

The C -positive region of D is the set of all objects from the universe U which can be classified with certainty into classes of U/D employing attributes from C , that is:

$$POS_C(D) = \bigcup_{x \in U/D} \underline{C}X \tag{4}$$

2.2 Tolerance Rough Set

The rough set theory introduced by Pawlak is based on equivalence relation. However, the requirement for equivalent relation of rough set prohibits it from being applied to many real applications. Skowron introduced a generalized tolerance space by relaxing the equivalence relation to a tolerance relation, for which transitivity property is not required. For the further discussion in the paper, a generalized tolerance space is described below [8], [9], [10], [11].

$P(U)$ is used to denote sets of all subsets of U , $I:U \rightarrow P(U)$ is used to denote a tolerance relation, if and only if $x \in I(x)$ for $x \in U$ and $y \in I(x) \Leftrightarrow x \in I(y)$ for any $x, y \in U$. Thus the relation $xIy \Leftrightarrow y \in I(x)$ is a tolerance relation which satisfies reflexive condition and symmetric condition and $I(x)$ is a tolerance class of x . The tolerance rough membership function $\mu_{I,v}$ is defined as follow [9]:

$$\mu_{I,v}(x, X) = v(I(x), X) = \frac{|I(x) \cap X|}{|I(x)|} \tag{5}$$

The tolerance rough set for any $X \subseteq U$ are then defined as [9]:

$$L_R(X) = \{x \in U \mid v(I(x), X) = 1\} \tag{6}$$

$$U_R(X) = \{x \in U \mid v(I(x), X) > 0\} \tag{7}$$

3 Personalized Recommendation Model Based on Rough Set

The model of the proposed personalized recommendation model base on rough set is described in Fig.2.

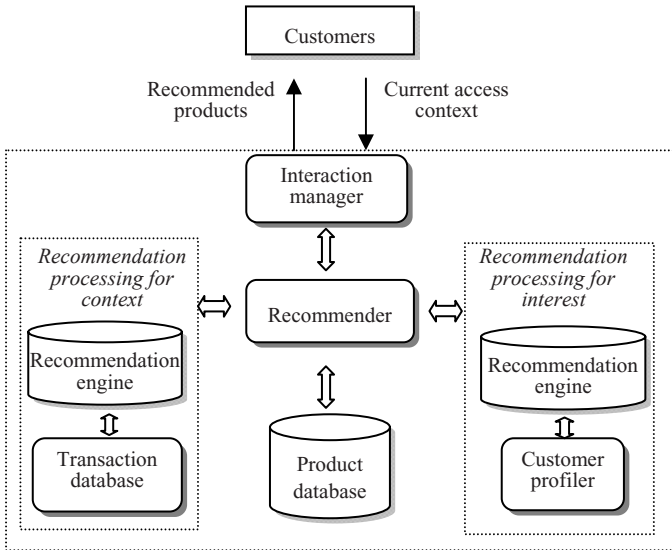


Fig. 2. Schematic view of personalized recommendation model

The model is mainly composed of four components, i.e., interaction manager, recommender, recommendation engine for context, and recommendation engine for interest.

3.1 Interaction Manager

Once the customer logs in the system and browses information of products, the interaction manager accepts the customer information and his current access context, and sends it to recommender, which responses for information processing and obtain the recommended results. Then, the information of the recommended products are fetched and presented to the customer.

3.2 Recommender

This component implements the function of organizing recommend processing. Firstly, it calls the recommendation engine for context to mine the associated products according transaction database based on tolerance rough set theory. Then, it calls the Recommendation engine for interest to analysis the preferences of customer according customer profiler. Further, before been sent and presented to the customer, the recommended products are check whether has storage or not by inquiring product database.

3.3 Recommendation Engine for Context

In this paper, $U = \{d_1, \dots, d_M\}$ is used to represent the set of purchase records and $T = \{t_1, \dots, t_N\}$ is used to represent the set of items for U . Then, we define the tolerance

space over a universe of all items for U . Furthermore, we determine the tolerance relation I as the co-occurrence of items in all purchase records from U .

We use $f_U(t_i, t_j)$ to denote the number of record that both item t_i and t_j occurs in U . Then, we define the uncertainty function I with regards to co-occurrence threshold θ as follow [9], [11]:

$$I_\theta(t_i) = \{t_j \mid f_U(t_i, t_j) \geq \theta\} \cup \{t_i\} \tag{8}$$

The above function satisfies reflexive condition and symmetric condition, i.e., $t_i \in I_\theta(t_j)$ and $t_j \in I_\theta(t_i) \Leftrightarrow t_i \in I_\theta(t_j)$ for any $t_i, t_j \in T$. According to the concepts of tolerance rough set introduced in section 2.2, $I_\theta(t_i)$ is the tolerance class of item t_i .

Further, we can define the membership function μ for $t_i \in T, X \subseteq T$ as follow:

$$\mu(t_i, X) = v(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|} \tag{9}$$

Then, the lower approximation of the subset $X \subseteq T$ can be determined as follow:

$$L_R(X) = \{t_i \in T \mid v(I_\theta(t_i), X) = 1\} \tag{10}$$

And, the upper approximation of the subset $X \subseteq T$ can be determined as follow:

$$U_R(X) = \{t_i \in T \mid v(I_\theta(t_i), X) > 0\} \tag{11}$$

3.4 Recommendation Engine for Interest

This component implements the function of discovery the interested products for customer according to customer profiler. The customer profiler maintains the profiles of all customers who registered the server of M-commerce. It is used by the recommendation engine to retrieve and store knowledge about customer, such as products purchased by him and latent interested products. Based on tolerance rough set, tolerance class for purchased products by customer can be obtained. The products in the tolerance class are considered latent interested products for customer and thus recorded in the profiles.

4 Conclusion

In this paper, a novel personalized recommendation method based on tolerance rough set is proposed. Tolerance rough set is used to capture customer preference according to both the customer’s interests and his current context. Based on the proposed method, a model is described in detail. Our future effort is to incorporate rough set with other soft computing theory (such as fuzzy set and granular computing) to design more effective personalized recommendation approach and model in mobile commerce.

References

1. Clarke, I.: Emerging value propositions for m-commerce. *Journal of Business Strategies* 18(2), 133–148 (2001)
2. Varshney, U., Vetter, R.J., Kalakota, R.: Mobile commerce: A new frontier. *IEEE Computer* 33(10), 32–38 (2000)
3. Lee, E., Jin, J.: A Next Generation Intelligent Mobile Commerce System. In: Ramamoorthy, C.V., Lee, R., Lee, K.W. (eds.) *SERA 2003*. LNCS, vol. 3026, pp. 320–331. Springer, Heidelberg (2004)
4. Tsalgaidou, A., Pitoura, E.: Business Models and Transactions in Mobile Electronic Commerce: Requirements and Properties. *Computer Networks* 37, 221–236 (2001)
5. Pawlak, Z.: Rough Sets. *Int. Journal of Computer and Information Sciences* 11(5), 341–356 (1982)
6. Shi, L., Xi, L., Duan, Q.: Personalized Web Search System Based on Rough Sets. *Computer Science* 34(10), 162–164 (2007)
7. Pawlak, Z.: *Rough sets: Theoretical aspects of reasoning about data*. Kluwer Academic Publishers, Dordrecht (1991)
8. Järvinen, J.: Approximations and Rough Sets Based on Tolerances. *Rough Sets and Current Trends in Computing*, 182–189 (2000)
9. Ho, T.B., Nguyen, N.B.: Nonhierarchical document clustering based on a tolerance rough set model. *International Journal of Intelligent Systems* 17(2), 199–212 (2002)
10. Yao, Y.Y.: Information granulation and rough set approximation. *International Journal of Intelligent Systems* 16, 87–104 (2001)
11. Ngo, C.L., Nguyen, H.S.: A tolerance rough set approach to clustering web search results. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) *PKDD 2004*. LNCS (LNAI), vol. 3202, pp. 515–517. Springer, Heidelberg (2004)

SpamTerminator: A Personal Anti-spam Add-In for Outlook

Wenbin Li^{1,2,4}, Yiying Cheng¹, Ning Zhong^{2,3}, TaiFeng Liu¹,
and Xindong Zhang¹

¹ Shijiazhuang University of Economics, Shijiazhuang 050031, Hebei, China

² Beijing University of Technology, Beijing 100022, China

³ Dept. of Information Engineering, Maebashi Institute of Technology,
460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan

⁴ Hebei Normal University, Shijiazhuang 050000, Hebei, China

Abstract. Spam filtering has witnessed a booming interest in the recent years, due to the increased abuse of email. This paper presents SpamTerminator, a personal anti-spam filtering add-in of Outlook. Outstanding characteristics of SpamTerminator are as follows. First, it provides eleven filters including rule-based, white lists, black lists, four single filters, and four ensemble filters. As a result, SpamTerminator can automatically work for users in different stages even if they do not train machine learning-based filters. Secondly, by using our proposed method named TPL (Two-Phases Learning) to combine multiple disparate classifiers, ensemble filters can achieve excellent discrimination between spam and legitimate mail.

1 Introduction

E-mail is one of the most successful computer applications yet devised. As email becomes a more prevalent salesmanship of e-business, dealing with spam is becoming more and more costly and time consuming. There is now a great deal of interest in email clients that allow a technically naive user to easily filtering spam. In fact, to solve the problem of spam, many email clients such as Outlook, foxmail provide a filtering function based on keyword-spotting rules. In a “keyword-spotting” rule, the primitive conditions test to see if a word appears (or does not appear) in a certain field of an email message. However, this method has two disadvantages at least. First, constructing and maintaining rules for filtering is a burdensome task. Second, any changes to the keywords will result in restructuring the filtering rules.

These scenarios provide ample ground for developing more powerful and practical functions to block spam in email clients. This work presents such an add-in named SpamTerminator, toward a very prevalent client, i.e., Outlook. Richard Segal, et al. pointed out that “... We believe that *no one anti-spam solution is the right answer*, and that *the best approach is a multifaceted one*, combining various forms of filtering ...” [1]. Motivated by this, SpamTerminator provides eleven filters including a rule-based one, the second based on white list, the one

based on black list, Naive Bayes (NB) [2], k -NN [3], C4.5 [4], ANN [5], and four ensemble filters. The first three filters should be constructed manually by users without training datasets, nevertheless other ones are built automatically once the count of new spam and legitimate email exceeds a threshold. As a result, SpamTerminator can automatically work for users in different stage even if they do not train machine learning-based filters. It is worth while to note that, Two-Phases Learning (TPL) algorithm is proposed to combining multiple filters. The goal of combining learned filters is to obtain a more accurate prediction than can be obtained from any single source alone.

The remainder of this paper is organized as follows. In Section 2, the architecture and some snapshots of SpamTerminator are given. We then present the design and core algorithm TPL in Section 3. In Section 4, Some experimental results about TPL are shown. Finally, we give conclusions in Section 5.

2 SpamTerminator Overview

2.1 Architecture

Figure 1 shows the architecture of SpamTerminator. SpamTerminator provides following functions, 1) categorizing emails according to senders' addresses; 2) filtering emails according to rules constructed by hand; 3) filtering emails according to white-/black lists built automatically from selected samples; 4) filtering emails via 8 machine learning-based filters. To support these functions, as shown in Fig. 1, some databases such as rules, white-/black lists, feature table, feature subset, sample vectors are needed. The rules database is used for storing filtering rules. The white-/black lists are used for storing white-/black lists. The feature table is used for storing the vocabulary including features, i.e., words or phases appeared in all training examples. The features useful for filtering are stored into the feature subset database. The sample vectors database stores represented vectors of training instances in feature subset space. In addition, Chinese word segmentation is a necessary component for a content-based filter oriented both of English and Chinese emails. As a result of VSTO (Visual Studio 2005 Tools for Office) [6] based developing, SpamTerminator should be supported by .net framework. VSTO 2005 is a component of Visual Studio. It provides C# and VB.NET developers with tools for building applications that leverage standard Office programs like Outlook, Word, and Excel, all in an environment that matches the development, security, and deployment of existing .NET applications. VSTO is another example of how Microsoft is extending classic Office products through the use of XML and Web services.

SpamTerminator provides eleven filters. Three of them (rules, white-/black lists) categorize an new email into "spam" or "legitimate" according to header's information of the message. They do not need to be trained in advance. However, other filers are learned from a given dataset. In SpamTerminator, in order to construct such a filter, following steps are executed. After that, a filter can be learned from email vectors.

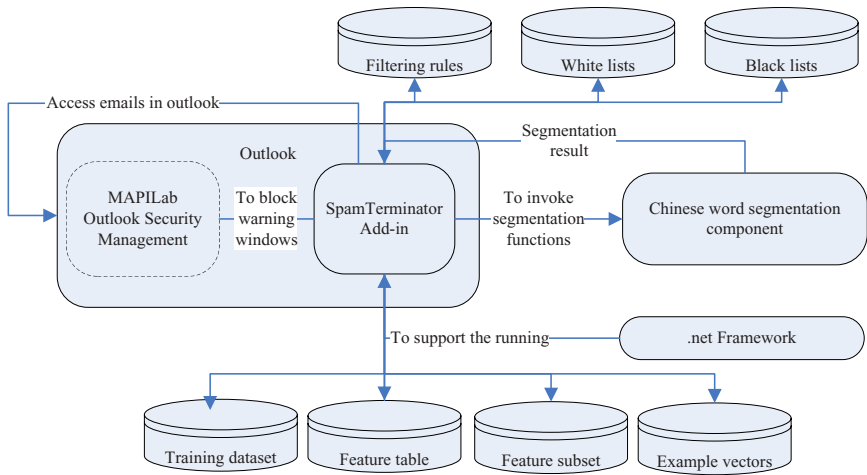


Fig. 1. The architecture of SpamTerminator

- Dataset preparing. SpamTerminator provides an interface for users to specify some messages in inbox or spam folder as “spam” and “legitimate” samples. Then, those samples are stored as files in the disk.
- Word segmentation. This step is conducted on the files got in the above step. The word segmentation component is invoked to add a space between two separated words/phases.
- Feature table generation. Words/phases are collected from the files processed in the second step. All words/phases and their relative information are stored in the feature table database.
- Feature selection. This step is used to select “best” features for classification from the feature table. Information Gain (IG) [13] method is adopted in SpamTerminator to finish this task. The selected features are stored in the feature subset databases. Suppose that f_1, \dots, f_M are features in the subset. Then, (f_1, \dots, f_M) is called as feature space.
- Email representation. In order to train a base filter, each of training example should be represented as a vector in the feature space (f_1, \dots, f_M) . Absolute word frequency and relative word frequency are candidates for our email representation. In SpamTerminator, 0–1 word frequency is used. All training vectors are stored into the sample representation database.

2.2 Snapshots

Figure 2 shows the menu and filtering snapshots of SpamTerminator. Its main function items are shown in dashed ellipses. The above toolbox shown in left dashed ellipse provides the function for a user to specify an email as spam or legitimate example. The below one shown in the same ellipse offers a user the functions, 1) extract white/black lists from some selected messages; 2) filtering

new emails that satisfy the setting conditions (corresponding window is the top window shown in Fig. 2). In the right dashed ellipse, as shown in Fig. 2, the main menu of SpamTerminator is given. In English, the functions shown in the menu are: 1) initialize SpamTerminator; 2) classify email according to sender's address; 3) manage rules; 3) manage white lists; 4) manage black lists; 5) manage training dataset; 6) select feature subset; 7) train filters; 8) about us. The steps word segmentation, feature table generation, feature selection and email representation mentioned in above subsection are implemented in menu item "select feature subset".

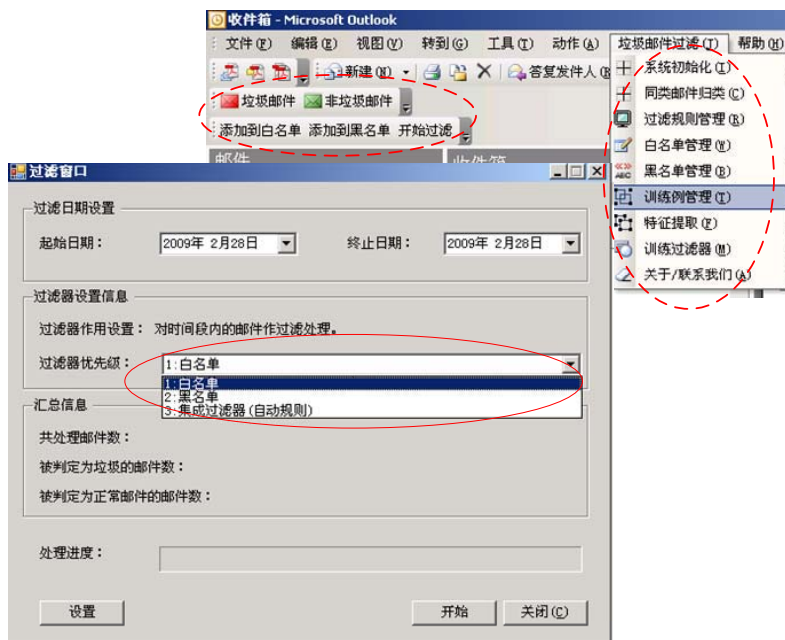


Fig. 2. The menu and filtering snapshots of SpamTerminator

When a new email arrives at inbox, SpamTerminator will start to category it according to user's setting. The main setting parameters include time range, filters's priority. That is, a user can let SpamTerminator filter unlabeled emails in a time range. A new email may be categorized by multiple filters. As shown in the real line ellipse in Fig. 2, the user choose three filters: white lists, black lists, and a ensemble filter. At the same time, the user set their priority as white lists \succ black lists \succ the ensemble filter. Therefore, SpamTerminator will filter the message according to white lists. If the process is successful, other filters will be ignored. Otherwise the black lists and the ensemble filter will be activated in turn.

SpamTerminator provides 11 filters. We categorized them into three types. First type is manual one including rules, white-/black lists. Second type is base

filter including NB, k -NN, C4.5, ANN. The last type is ensemble kind including K -NB, K -kNN, K -C45, K -ANN. Each kind of them has valid contributions towards SpamTerminator. Specifically speaking, the first type offers ideal accuracy and speed. However, it needs frequency maintenance by user. The second kind release users from manual burden. Nevertheless, users worry about its precision. To bridge the gap, Two-Phases Learning (TPL) algorithm is proposed to combining multiple filters. The goal of combining learned filters is to obtain a more accurate prediction accepted by users. Below, TPL algorithm will be described in detail.

3 Combining Filters with TPL

Mentioned as above, spam filtering is a task classifying emails into two categories essentially [78]. In such a task, there are two key steps [79]. The first one is selecting the best features from a vocabulary V . The second step is to train a best filter. In order to get the best performance, we need the best features as well as the best filter. However, it is difficult for us to reach the two “best” at the same time [10]. One way addressing this problem is to combine multiple classifiers, in which the classifier can supply a gap for each other. Hence, even each classifier is not the best one, maybe their combining “wisdom” can attain an ideal effect. Combining multiple homogeneous or heterogeneous filters to improve the performance of a single filter is the first motivation of TPL. In fact, there are considerable research work on combining classifiers. However, most of them do not adapt to be used in email client, because they do not consider the limitation of client’s memory. In other words, higher performance, fast training speed and limited memory requirements are practical goals of an email filter. Therefore, in our opinion, many classifying algorithms with ideal accuracy such as BP-ANN, SVM, AdaCost are not candidates in SpamTerminator. Because they perform slowly in training process on the one hand. On the other hand, they need large memory, especially when the feature space is large. Hence, how to guarantee acceptable accuracy and train a filter fast in limited memory are challenging problems. This is the second motivation of TPL.

TPL consists of two main learning phases. The first one is called as the phase of *direct learning*, while another one is *indirect learning*. In the first phase, multiple homogeneous or heterogeneous filters are directly learned from the training dataset. After this, to some extent, each of these filters holds the knowledge that distinguishes spam and legitimate. When all these filters “tell” their knowledge to a committee, the committee also can distinguish spam and legitimate. That is, the committee is not learned from a training dataset but from the knowledge of the voting filters. Thus, we call this phase as the *indirect learning*. Figure 3 shows the working flow of TPL.

As shown in Fig. 3, TPL first trains multiple classifiers. In spamTerminator, we train four heterogeneous filters with a same dataset. The learning algorithms are Naive Bayes (NB), k -NN, ANN, C4.5 respectively. Then, the process of constructing filters’ knowledge is conducted.

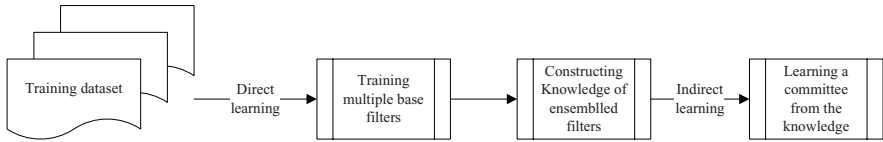


Fig. 3. The working flow of TPL

The next step is to construct knowledge base of combined filters. For the sake of convenience, we define following symbols. N denotes the count of ensemble filters. D is the count of instances in the training dataset. \tilde{h}_i ($i = 1, \dots, N$) knows the label or class probability of each training example. For a hard filter \tilde{h}_i , we use a vector $\vec{k}_i \equiv \langle c_{i1}, c_{i2}, \dots, c_{i|D|} \rangle$ to denote its knowledge, where c_{ij} is the class outputted by \tilde{h}_i for the j th training instance. If \tilde{h}_i is a soft filter, there are two possible representation methods. We may use a vector $\vec{k}_i \equiv \langle p_{i1}, p_{i2}, \dots, p_{i|D|} \rangle$ to denote the knowledge of \tilde{h}_i , where p_{ij} is output of \tilde{h}_i , i.e., the class probability (or function value) of the true class of the j th training instance. Alternatively, we can convert a soft filter to a hard one. For example, suppose that x_j is the j th training example, its true class is c_1 , and \tilde{h}_i outputs a class probability $p_{ij}(c_1|x_j)$ for x_j . If $p_{ij}(c_1|x_j) > (p_{ij}(c_0|x_j) + \alpha)$, then \tilde{h}_i labels x_j as c_1 , else it classifies into c_0 , where α is a threshold set by users. If we want to adopt a cost-sensitive method to represent the knowledge of \tilde{h}_i , we should set $\alpha > 0$. Otherwise, $\alpha = 0$. Formally, we can represent a set of filters as in matrix form as $K \equiv [\vec{k}_1^T, \dots, \vec{k}_N^T]$. In order to apply a supervised learning method to learn the $E(x)$ from K , we add another column so that K is given as $[\vec{k}_1^T, \dots, \vec{k}_N^T, \vec{k}_{(N+1)}^T]$, where $\vec{k}_{(N+1)}$ stores the corresponding labels of training examples.

After that, a committee is learned from K with a learning algorithm. In SpamTerminator, we respectively use NB, k -NN, ANN, C4.5 to learn a voter. Therefore, four types combined filters are provided by SpamTerminator. They are individually depicted as K -NB, K -kNN, K -ANN, K -C45. “K” denotes the knowledge. We call the process learning from “K” as indirect learning.

4 Experimental Results of TPL

To validate TPL, a series of experiments are conducted before SpamTerminator is designed and implemented. Main conclusions drawn from these experiments are that firstly, the performance of TPL is little affected by the number of combined filters. The second feature of TPL is that more selected relevant features will result in more powerful prediction abilities of TPL. Thirdly, ensembling heterogeneous filters in TPL was better than ensembling homogeneous ones. Because of the space limitation, part of results are given here.

Our experiments have been performed on spambase, PU1 corpus [11]. The PU1 corpus consists of 1099 messages, 481 of which are marked as spam and 618 are labelled as legitimate, with a spam rate of 43.77%. The messages in

PU1 corpus have header fields and html tags removed, leaving only subject line and mail body text. To address privacy, each token was mapped to a unique integer. The corpus comes in four versions: with or without stemming and with or without stop word removal. Spambase only distribute information about each message rather than the messages themselves for avoiding privacy issues. With 57 pre-selected features, each real email was represented by a vector. This corpus contains 4601 vectors about emails. Among these vectors, 2788 ones are about legitimate messages and 1813 are about spam.

We use *precision* [12] and *recall* [12] as evaluation criteria. Test method is 5-cross validation. k in k -NN is set to be 5. 150 features are selected by Information Gain method [13] for PU1 dataset. Respectively, hidden units, output units, weight decay, and number of training cycles of ANN are 10, 1, 0.2, 15. All experiments are conducted on Acer TravelMate 3273 laptop. With support of MATLABArsenal [14] toolbox (a machine learning toolbox for MatLab), experimental programs run in Matlab7.1.

Figures 4 and 5 give the results of NB, k -NN, ANN, C4.5, K -NB, K -kNN, K -ANN, K -C45.

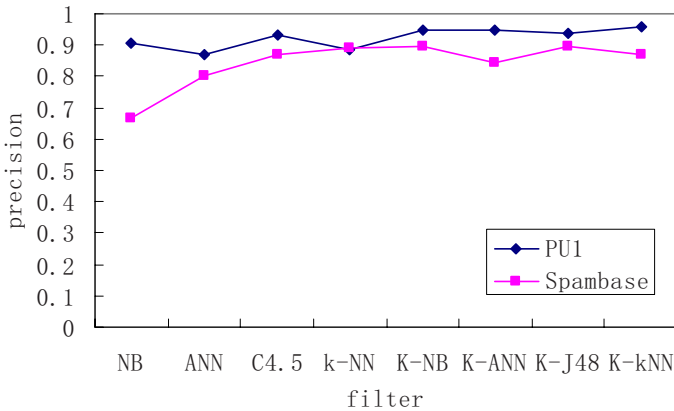


Fig. 4. Comparative results of *precision* on two datasets

Figure 4 shows those filters' results of *precision* on the datasets. From this figure, we can see that all the tested filters are roughly sorted in descending order by *precision* on PU1 as $(K-NB, K-ANN, K-J48, K-kNN) \succeq C4.5 \succeq NB \succeq (ANN, k-NN)$. That is, the ensemble filters learned by TPL are better than base filters. On Spambase, the sorted list is $(K-NB, k-NN, K-J48, K-kNN) \succeq (C4.5, K-ANN) \succeq ANN \succeq NB$. Figure 5 shows those filters' results of *recall* on the datasets. This figure shows that firstly $K-C4.5, K-ANN$ reach top performance on PU1. Secondly, on Spambase, $K-NB, K-ANN$ are better than other filters. These two figures demonstrate a same conclusion that TPL is an effective method. A ensemble filter combined with TPL can get better performance than a single

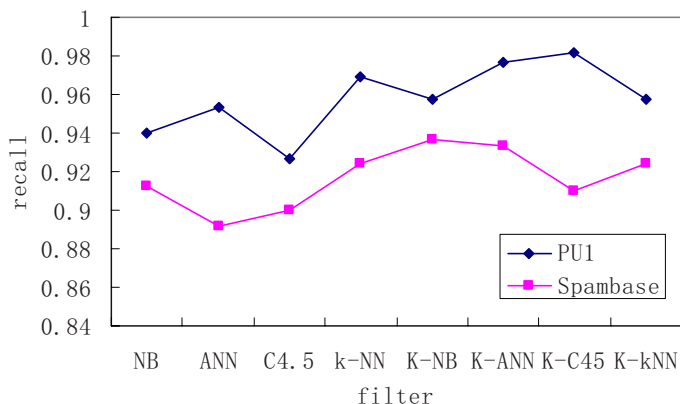


Fig. 5. Comparative results of *recall* on two datasets

one. SpamTerminator adopt K-NB, K-ANN, K-kNN, K-C45 just because TPL has such an advantage.

5 Conclusions

To prevent email spam, both end users and administrators of email systems use various anti-spam techniques. Some of these techniques have been embedded in products, services and software to ease the burden on users and administrators. This paper presents SpamTerminator, a personal anti-spam filtering add-in embedded in Outlook. This work introduces its architecture, main functions, and key techniques. SpamTerminator provides eleven filters including rule-based, white lists, black lists, four base filters, four ensemble filters. On the one hand, it can automatically work for users in different stage even if they do not train machine learning-based filters. On the other hand, by using ensemble filters combined by our proposed method TPL, it can achieve excellent discrimination between spam and legitimate mail.

Acknowledgements

The work is partially supported by NSFC (NO. 60673015), and Project (NO. 07213507D, NO. 2001BA201A12, NO. 072435158D) of Dept. of Science and Technology of Hebei Province, Doctor Fund of Shijiazhuang University of Economics.

References

1. Segal, R., Crawford, J., Kephart, J., et al.: SpamGuru: An enterprise anti-spam filtering system. In: Proceedings of the First Conference on Email and Anti-Spam (2004)

2. Domingos, P., Michael, P.: On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29, 103–137 (1997)
3. Belur, V.D.: Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques (1991) ISBN 0-8186-8930-7
4. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco (1993)
5. Raul, R.: *Neural Networks: A Systematic Introduction* (1996) ISBN 978-3540605058
6. VSTO, <http://msdn.microsoft.com/en-us/office/aa905543.aspx>
7. Li, W.B., Liu, C.N., Chen, Y.Y.: Combining multiple email filters of Naive Bayes based on GMM. *ACTA ELECTRONICA SINICA* 34(2), 247–251 (2006)
8. Li, W.B., Zhong, N., Liu, C.N.: Design and implementation of an email classifier. In: *Proc. of International Conference on Active Media Technology, AMT 2003*, Chongqing, China, pp. 423–430 (2003)
9. Li, W.B., Zhong, N., Yao, Y.Y., Liu, J.M., Liu, C.N.: Filtering and Email-Mediated Applications. In: Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S., Li, K. (eds.) *Web Intelligence Meets Brain Informatics. LNCS (LNAI)*, vol. 4845, pp. 382–405. Springer, Heidelberg (2007)
10. Li, W.B., Zhong, N., Liu, C.N.: Combining multiple email filters based on multivariate statistical analysis. In: *The 15th International Symposium on Methodologies for Intelligent Systems, Bari, Italy. LNCS(LNAI)*. Springer, Heidelberg (2006)
11. PU1 and Ling-Spam dataset:
<http://iit.demokritos.gr/skel/i-config/downloads/>
12. Makhoul, J., Francis, K., Richard, S., Ralph, W.: Performance measures for information extraction. In: *Proc. of DARPA Broadcast News Workshop*, Herndon, VA (February 1999)
13. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: *Proc. of 14th International Conference on Machine Learning, ICML 1997*, Nashville, TN, US, pp. 412–420 (1997)
14. MATLABArsenal,
<http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenal.htm>

Classifying Images with Image and Text Search Clickthrough Data

Gavin Smith, Michael Antunovic, and Helen Ashman

WebTech and Security Lab, University of South Australia
{gavin.smith,helen.ashman}@unisa.edu.au

Abstract. Clickthrough data from search logs has been postulated as a form of relevance feedback, which can potentially be used for content classification. However there are doubts about the reliability of clickthrough data for this or other purposes. The experiment described in this paper gives further insights into the accuracy of clickthrough data as content judgement indicators for both HTML pages and images. Transitive clickthrough data based classification of images contained in HTML pages has been found to be inferior to direct classification of images via image search clickthrough data. This experiment aimed to determine to what extent this is due to the inferior accuracy of clickthrough-based classification accuracy in HTML. The better classifications resulting from clickthroughs on image searches is confirmed.

1 Introduction

1.1 Background

Clickthrough data is increasingly seen as a potential indicator of relevance feedback on search results, not just for search ranking purposes but also for other applications such as classification of content.

Clickthrough data is generated when a user selects (clicks on) results from a page returned from a search. The implication is that the user will primarily click on results of most relevance to their information requirement as expressed in the query term submitted to the search engine. Thus clickthrough is a form of relevance feedback that rates the relevance of the tendered results to the query.

While simple in principle there are a number of known issues (e.g. trust bias, see section 2) with the reliance on clickthrough data as implicit judgment. Such issues and the severity of their impact then vary depending on the type of search (document vs. image) providing the clickthrough data. They introduce noise, and the nature of click-through logs [13] introduces low coverage. However, the existence of such cheap, large and continuously generated logs inspires attempts to resolve these issues of noise and increase coverage. This paper evaluates the potential of web search clickthrough data to contribute to image classification via a transitive labeling method [1]. Such a method seeks to alleviate the sparsity problem by using an alternative resource

to the click-through data from the resource's own search (as in [4]) and one for which a potentially larger source exists¹.

1.2 Motivation – Transitive Classification of Images

It seems that reliable clickthrough data should be a primary source of relevance feedback on searches especially in terms of ranking of image results, as has been proposed for web document search results [3] [6]. So far this has been seldom reported in the open literature and this may be because of the difficulty in procuring web log data that includes image searches. However, clickthrough data can also be used to extend the applicability of established labels and classifications, such as in the *transitive* method for image labelling which inherits a classification or label for an HTML or text page onto potentially every image contained in that document [1]. This method additionally requires filtering out "non-content" images which were identified by inspection as being frequently generated by advertisements, formatting images (such as bullet point icons and lines), and banners.

One purpose of the experiment described in this paper is to determine whether this transitive labelling method can accurately supplement existing image classification technologies. We experimentally validated the image labels generated by this transitive method (along with the direct method and 4 other widely-used methods) with ground-truthing by over 100 human assessors. We found that these transitive labels were inferior to both the Google Image Search and the derivative direct labelling method (which consists of applying a search term as a label to an image if it has 2 or more clickthroughs from an image search on that term which confirm the image's relevance to the term) [1]. However it was not immediately clear whether the errors in the transitive method arose from a poor filtering algorithm or whether it was an artifact of "garbage in, garbage out", that is, whether the image classifications were poor because they inherited poor labels from their containing web pages. Given the doubt about text search clickthrough validity (see section 2.1), the page labels themselves could have created a significant proportion of the errors which were inherited onto the images they contained.

This paper thus reports on an experiment that establishes whether a set of Web pages, containing images which were labelled using the transitive method, was accurately labelled. This first result demonstrates that the accuracy of images labelled with clickthroughs using the direct labelling method is significantly higher than the accuracy of HTML pages labelled in the same way. This indicates how much "garbage" goes into the transitive labelling algorithm. We then go on to correlate the accuracy of transitively-generated image labels with the accuracy of HTML page labels, finding that the filtering algorithms that purportedly screened out "non-content" images and left in "content" images were at least partly responsible for the inaccuracies of the transitive labelling method.

We next go on to section 2, which considers related work by others evaluating the use of clickthrough data. Section 3 then describes the experiment that ground-truths

¹ In the three years of collected Web logs from the University of Teesside used in this experiment, we found that image searches amounted to only around 5% of the total amount of web image and text searches.

the accuracy of HTML search clickthrough data, while section 4 details the results of the experiment. Section 5 discusses the results, and section 6 concludes.

2 Related Work

In this section we look at other assessments of the relevance of clickthrough data. We find that most such assessments to date focus on text searches.

2.1 Clickthroughs to Web Pages

Clickthrough data from traditional web page search has been the subject of much recent work. Proposed for a wide range of uses, in 2005 the usefulness of this data began to be questioned. From the research it became clear that it was not correct to rely on the assumption that clickthrough data could be directly used as an absolute judgment of relevance [5] [8]. Others [4] [7] found that users of the search systems were biased in a number of ways when it came to clicking results, causing them to click on results that may not be the most relevant. Specifically quality-of-context, trust and position biases were identified. Despite this drawback, the prolific and cheap nature (e.g. compared to explicit human labeling) of such data has seen its continued use, with research looking at ways of re-weighting judgments to counter the identified biases, most recently [3] [6]. As such the notion that clickthrough data can provide relative relevance is more commonly accepted.

In particular it is proposed that clickthrough data from image searches is more accurate than that of text searches, as discussed in the next section.

2.2 Clickthroughs to Images

In contrast to clickthrough data from traditional web page search, clickthrough data from web image search has seen little evaluation. While a recent study [9] indicated the higher accuracy of image search clickthrough than web search clickthrough data, the exact presence and impact of the biases identified in web search clickthrough data is not clear. It is also notable that proposed research for re-weighting clicks cannot be directly applied due to the different presentation of search results - image search results are often presented in a grid with no obvious "best" or "first" result.

However, when looking to create highly accurate labels of images (as opposed to re-ranking) the presence of various forms of bias is of less concern. Such a goal has applications to techniques such as query and image clustering [2] [10] and image concept learning [11]. In prior work, [1] the ability to accurately label images by a number of different methods was evaluated, including two based on different applications of clickthrough data. In the first of these two methods, the *direct* method, image search clickthrough data was used to effectively filter Google image search results. Using this method, high levels of accuracy were reported, however, the method returned a low number of images due to the sparsity problem inherent in clickthrough data [13]. In contrast a *transitive* method, based on mining images from clicked web pages from a web search clickthrough log and associating the clicked page's query with the image, was evaluated. Such an approach suffered less from the sparsity problem as significantly more text searches are performed than image searches (only 5%

as noted earlier) and hence more pages are clicked than images. However, the accuracy of the classifications generated by the transitive was somewhat less, with completely relevant image/label pairs from the transitive method occurring around 63% of the time, compared to the Google image search's 80%. This paper extends this evaluation work, identifying the cause of this loss in annotation accuracy in order to develop larger sets of images for a given concept (search keyword(s)).

3 Experiment to Ground-Truth Clickthrough Accuracy

3.1 Experiment Description and Motivation

The main comparison to make is between the precision of images classifications/labels created with clickthroughs and the precision of their containing websites, also labelled with clickthroughs. Having established in the earlier experiment [1] that the accuracy of image/label pairs generated with the direct method (around 84%) is greater than the image/label pairs generated by the transitive method (around 63%), it is necessary to identify the cause of this discrepancy.

There are two potential causes of the discrepancy:

i) The labels belonging to the Web pages themselves were inaccurate. This seems feasible given the doubts in the literature about the reliability of clickthrough data for classification purposes (as discussed in section 2);

ii) The filtering mechanisms employed to remove "non-content" images, such as advertisements and formatting, were too coarse. This also seems feasible as the filtering algorithms were derived prior to any analysis of the relevance of contained images, and were largely speculative. The assumption is that images are by default relevant to their containing pages, unless they are "non-content" as defined here.

It may be that both of these causes affected the accuracy of the transitive method, so the experiment set out to ground-truth the accuracy of clickthroughs on the selected set of Web pages from which the images were extracted and labelled.

In summary, the experiment sets out to measure the accuracy of clickthrough as content classification judgement on the set of Web pages from which images were extracted and likewise classified, and to measure the impact of this accuracy.

3.2 Experimental Set Up

There have been two experiments set up to ground-truth the accuracy of labels applied as a result of clickthrough data. In the earlier image label evaluation experiment [1], six different image labelling methods were evaluated for their precision, these being the direct labelling method (a derivate of Google's Image Search as described above), the transitive image labelling method (also described above with implemented filters summarized in Table 1), the human-provided image labels from the Google Image Labeller game² (based on the ESP game [12]), the flickr image hosting site³ and Getty

² <http://images.google.com/imagelabeler/>

³ <http://www.flickr.com/>

Images⁴, plus the Google Image Search facility⁵. Using a mostly canonical set of query terms, a collection of image/query term pairs was generated, selecting the top 11 or 12 for each query as tendered to Google Image Search, flickr, Getty Images, the direct method and the transitive method, along with a random selection of around 1000 Google Image Labeller images, screen-scraped from the game interface.

These six methods were evaluated by comparing the implied label or classification of each over a total of 4693 images, and we found that the direct labelling method was 4% more accurate than Google Image Search from which it was derived, while the transitive method was incrementally more accurate than the remaining methods, although significantly behind the direct method and Google Image Search [1].

The poor showing of the transitive method needed further analysis, as it was not clear whether it was due to inaccuracy in the filtering process, or due to the Web pages being poorly-labelled themselves. To analyse this we set up the second experiment to firstly assess the validity of the web pages from which the transitive images were drawn, and secondly to compare the accuracy of the labels of the images versus their containing web pages. This second experiment uses the same approach as the original image-assessing experiment, and combining results from the two allows both an assessment of accuracy of web page labelling based on clickthrough, and a measurement of the liability of poor page labelling in the image labels.

Some variables were fixed as far as possible:

- fixed: we used the original clickthrough data from the same dataset for both images and websites;
- fixed: we used the same set of 71 queries for both the image search clickthroughs (from which the direct method created image labels) and for the text-based search (from which the web page labels were derived using the direct method and subsequently the image labels were derived using the transitive method);
- semi-fixed: ground truthing of both images and webpages was done by many of same people. 12 people have ground-truthed all of the websites compared to 10 have ground-truthed all of the images, and 7 of these did both complete sets.

Table 1. The transitive filtering functions as implement in [1]

Filter	Reject condition	Filter	Reject condition
repeated	Img repeated in page	tooSmall	Img height or width < 50 pixels
aspectRatio	Img aspect ratio > 3:1	logos	Img path contains text 'logo'
		advertisement	Img in blacklist 'Rick752's EasyList' ⁶

The clickthrough data was extracted from around three years worth of anonymised Web logs provided by the University of Teesside School of Computer Science.

The original image/label ranking experiment showed each image in a frame with the associated label above it, below a set of 4 options for the evaluator to select:

⁴ <http://www.gettyimages.com/>

⁵ <http://images.google.com/>

⁶ <http://easylist.adblockplus.org/>

- 0 - image did not load or user did not recognise it;
- 1 - image was not relevant to the label;
- 2 - image was partially relevant to the label, e.g. a car tyre, labelled 'car'
- 3 - image was completely relevant to the label.

Each evaluator was presented all images from all methods in a random order, preventing identification of the exact method used for an individual image.

The web page ranking experiment was then set up allowing evaluators to evaluate the relevance of the source website against the same search terms.

The second experiment was done similarly to the original image rating web application, but where the website was loaded within the page itself via an inline frame, and the search term used to find images from that website placed above the frame. The same Likert rating scale was used, with the following interpretations of the ratings made:

- 0 - Website did not load, or the user did not know what the website was about;
- 1 - Website not relevant to the search terms used to find images on the site;
- 2 - Website partially relevant to the search terms, i.e. some content but not all;
- 3 - Website completely relevant to the search terms used to find images on the site.

This latter experiment relevance-ranked the 184 sites from which the transitively-labelled images were drawn, and hence the labels being evaluated were the same as those being evaluated in the image experiment.

4 Experimental Results

In this section, the data collected is described, then results are given on firstly the relevance of the web pages to their search term from which the transitive images were drawn, and secondly on the accuracy of the labels of the images versus their containing web pages.

4.1 The Data Generated

The data generated by the two experiments now yields a set of relevance ranking data for each of the three sets, the clickthrough-classified web pages, the clickthrough-classified images and the transitively-classified images, as follows:

- a set of images and a set of web pages, both of which were classified using exactly the same clickthrough method (the direct labelling method). These objects were chosen from the same main set of web log data, so as to have coincident classification labels. It is not certain but it is believed likely that the participants generating the original clickthroughs overlapped, since the clickthrough data is from a relatively small population of staff and students in a university school, with strong association of search activity with undergraduate assignments and similar research tasks.
- A third set of objects (the transitively-labelled images) has classification labels derived from its parent objects (the set of web pages).

There are two subcategories of this third set, the first subcategory being the "non-content" images which were those that the filtering process filtered out as being not relevant (due to a wrong aspect ratio, being too small or from a known advertisement source), while the second subcategory were the "content" images, which were those not filtered out. Importantly, both the content and non-content images were relevance ranked by the human assessors, as this gave information about both the false positives of the filtering process (filtered out but should not have been, i.e. those identified as non-content but which were in fact genuinely relevant to their labels) and the false negatives (those not filtered out but which should have been).

4.2 Relevance of Web Sites to Search Terms (Clickthrough Accuracy)

A sample of the data showing the raw data generated from the website relevance ranking experiment is given in table 1. The *precision* is the number of people ranking the Website as 3 (fully relevant) divided by the total number of non-zero rankings (zero rankings are excluded as the evaluator was unable to make a judgement on their relevance). The *partial precision* was the number of rankings of either 3 or 2, divided by the total number of non-zero rankings.

Table 2. Sample data from Website ground-truthing experiment

<u>Site</u>	<u>Precision</u>	<u>Partial Precision</u>	<u>3s</u>	<u>2s</u>	<u>1s</u>	<u>0s</u>	<u>Average Ranking</u>	<u>TotalNo. Rankings</u>
Site 1	0.64	0.73	7	1	3	1	2.36	12
Site 2	0.92	1.00	11	1	0	0	2.92	12
Site 4	0.54	1.00	7	6	0	0	2.54	13
Site 5	1.00	1.00	12	0	0	0	3.00	12
Site 6	1.00	1.00	12	0	0	0	3.00	12
Site 7	0.58	0.92	7	4	1	0	2.50	12
Site 8	1.00	1.00	13	0	0	1	3.00	14
Site 9	0.92	1.00	11	1	0	0	2.92	12
Site 10	1.00	1.00	12	0	0	0	3.00	12
Site 11	1.00	1.00	12	0	0	0	3.00	12

We furthermore found that there were 25 sites that were unanimously ranked 3 by all users, giving them a 100% precision. Another 97 sites had all rankings 2 or 3, i.e. a partial precision of 100%. There were no sites with unanimous 2 or 1 rankings. 57 of the sites (including the 25 unanimously-ranked ones) showed a very consistent ranking with a relative standard deviation (RSD) across all evaluators of not more than 10%. A further 77, making up to 134 sites, showed an RSD of less than 25% - examples include a number of sites with 8 3-rankings and 4 2-rankings whose precision was 2.67, well within the "relevant" range. A small number, 5 sites, showed highly variable rankings where the RSD was over 50%, where presumably the evaluators found the label contentious.

We grouped sites according to their relevance to the search term associated with the clickthrough as follows:

- i) relevant to the search term: any site with an average ranking of 2.5 or above, i.e. having at least half of its rankings at 3;
- ii) peripherally relevant to the search term: any site with an average ranking of 1.8 to under 2.5;
- iii) not relevant to the search term: any site with an average ranking below 1.8.

With this grouping, we find that the following results:

Table 3. Table of site relevance to search term

Relevance	number of sites	proportion of sites
relevant	133	76%
peripherally relevant	33	18%
not relevant	11	6%

These figures suggest that the relevance of clicked sites to the search term is not as low as indicated by other studies (see section 2). This could be due to the relatively small number of sites evaluated (184), or that the sites selected were not necessarily typical of all searches in a more general context or population⁷. Alternatively it may be due to the distinct type of relevance assessment. Some prior studies assessed user satisfaction with results returned for a given search term [5] [8] where there is implicitly a meaning assigned by the user to the search term in these circumstances. In contrast, the evaluation here is not of the relevance of the website to a given meaning, but rather its relevance to *any* meaning of the search term, i.e. regardless of any ambiguity in the search term, and any meaning the searcher may have had in mind.

4.3 The Match between HTML Page Labels and Contained Image Labels

Having established that the accuracy of text search clickthrough data is, at least in this context, performing better than the image relevance where classifications were transitively-generated, we now consider the accuracy of each image compared with the page it occurred in.

The following table shows the number of each of the possible pairings of image relevance with containing website relevance. Each such pairing is represented as a pair of numbers, with for example (3, 2) representing the case where the site was ranked 3 and the image was ranked 2. The images are also separated out into their distinct filtered type, with Content referring to images that the filtering algorithm thought were relevant to the page content, Too Small being images under 50x50 pixels, Wrong AR (aspect ratio) being an image too narrow, Logo meaning an image for some company and Advertising being an image from a known advertisement-supply site. Note that all of the transitively-labelled images were ranked, including those that were categorised as non-content. The purpose of this was to assess the impact of both false positives (images wrongly judged to be content, i.e. not filtered out but should

⁷ On inspection, it seemed apparent that a large proportion of searches, especially multiple searches on the same term, from these weblogs were performed by students seeking information for preparing reports and assignments.

have been) and false negatives (images wrongly judged to be non-content) from the filtering process.

The transitive method had been applied to generate 445 unique image/label pairs. Filtering algorithms were then applied to exclude those images that were superficially deemed to be "non-content", resulting in 185 of the 445 images being categorised as not being content relevant to the web page. Of the non-content, 150 were categorised as "too small", 13 as "wrong aspect ratio", 5 as "logo" and 17 as "advertisement".

Table 4. Paired relevance rankings of transitively-labelled images and sites, separated according to image filtering

(site, image)	Content	Too Small	Wrong AR	Logo	Advertising
3, 3	71.02%	8.05%	6.02%	0.00%	2.27%
3, 2	12.22%	3.82%	7.09%	0.00%	4.55%
3, 1	13.00%	80.00%	64.17%	100.00%	48.18%
2, 3	2.00%	0.24%	0.00%	0.00%	0.00%
2, 2	0.38%	0.38%	0.40%	0.00%	0.00%
2, 1	0.99%	5.28%	13.24%	0.00%	15.00%
1, 3	0.01%	0.02%	0.00%	0.00%	0.00%
1, 2	0.17%	0.06%	0.27%	0.00%	0.00%
1, 1	0.20%	2.41%	8.82%	0.00%	30.00%

Table 2 makes it clear that the filtering process leaves much to be desired. The filtering rules were derived by inspection only, and until this experiment had not been evaluated for their accuracy. Also, it was necessary to understand the extent to which the underlying web pages had been accurately classified so that the influence of the filtering algorithms could be understood and this factor considered separately from the filtering algorithm.

We consider now primarily the correlation between relevant sites (ranked 3) and images labelled or excluded from them. In the non-relevant sites (ranked 1), there were almost no relevant sites, as would be expected, as since the site is not relevant to the label, it is unlikely that any contained images would be, except by chance.

Having evaluated the outputs of the filtering, we can make the following interpretations of the results:

- In the Content category, the majority (71.02%) of site and image rankings combined were positive results for both (3,3). However from the relevant sites, there was a further 12% of images only partially relevant, and another 13% of images not at all relevant, and thus should have been excluded but were not blocked by the filtering algorithm. *In total, there was a false-positive rate (image ranked 1 when site was ranked 3 or 2) of almost 14%.*

- In the Too Small category, the majority of images (80%) that were ranked as being not relevant were deemed to come from an appropriate site. This supports the notion that many of the images were graphical artifacts such as bullet points, icons, avatars etc. despite their placement on a relevant site. There are however almost 12% of Too Small images filtered out but which were ground-truthed as being relevant (8.05%) or partially-relevant (3.82%). The filtering algorithm will need to be refined to better detect the potentially relevant but small images.

– In the Wrong Aspect Ratio category, most of the images that were excluded due to a wrong aspect ratio were ranked as not relevant by the evaluators. There is however a small but significant proportion of false negatives, excluded by the filtering algorithm but rated by the evaluators as being relevant (6%) or partially relevant (7%). Any refinement of the filtering algorithm will need to correct this.

– In the Logo category, all of the excluded images were both ranked as being not relevant to the label (100%) while having come from an appropriate site. Note however that in some cases, a logo might be considered relevant to the site, such as if it was the site belonging to the logo holder, however such sites were not among the pages and images assessed in this experiment.

– In the Advertisement category, very few of the images rejected because of their advertising provenance were either relevant (2.27%) or partially-relevant (4.55%) to the search term. Over 93% were not relevant, with over 63% correctly rejected by the filtering algorithm. There is some scope to correct for the wrongly excluded images so more sophisticated rules, such as correlating the image provenance with the content of both the image and site, could refine these results.

There is scope for improving the classification accuracy with better filtering algorithms. These modified filtering algorithms would be easy to assess as we would merely need to assess the new filtered outputs against the already ground-truthed images and websites. Possible improvements for refinement of the filtering algorithm by reducing false positives and negatives include:

- Aspect ratio filtered images were based on dividing the width by height by length and comparing this result to a predetermined ratio, which initially was 3:1. This ratio can be refined incrementally, for example, 3.5:1, working in 0.5 increments to monitor whether the amount of filtered images increases or decreases based on the adjustments and eventually fixing on the ratio with the lowest error rate;
- Logos can be filtered differently, including potential for automated text analysis in the image. The current filtering mechanisms simply look for the text *logo* in the filename. No consideration is made for language specifics, alternatives of the word (eg. banner) or potential for character matching in larger words that may affect accuracy (eg. logoff, logorrhea);
- Too small image sizes can be adjusted to be smaller, resulting in a decrease in the volume of images filtered out, potentially reducing the number of false negatives;
- Advertisements are a difficult category, as many of the sites were relevant to advertisement images due to site-specific tailoring strategies. Up-to-date "adblock" lists may improve this area of the filtering mechanism.

5 Discussion

So how much of the inaccuracy of the transitive method was due to the poor labelling of the original Web pages and how much was due to the filtering algorithms?

In section 4.2, the relevance of the websites to the search term (i.e. the accuracy of the clickthrough data) was measured, and it was found that 76% of the assessed sites were fully relevant, while a further 18% were peripherally relevant, and hence that

94% of the sites evaluated were at least partially relevant to the search term. However this was not reflected in the transitively-labelled images. The precision of the transitive method was much lower, with only 63% of the transitively-labelled images being relevant, or around 80% of images being partially-relevant. This suggests that the problem lay mainly with the filtering algorithm since in both cases, relevant and partially-relevant, the site relevance rankings were well above the image relevance rankings.

When we look at the correlated image/site rankings in table 3, we find both false positives (images that should have been excluded) and false negatives (images that were wrongly excluded). The false positives have a real impact on the precision of the transitive method, wrongly labelling images with the site's label and associating irrelevant images with the label. Less pressing in terms of the precision of the transitive method but still worth further investigation is the level of false negatives, where images are excluded but are still relevant.

From the point of view of the accuracy of the transitively-generated image labels, there is a false positive rate of almost 14%, with the vast majority of these (13% of the total, or nearly 93% of the false positives) being contained in relevant (ranked 3) sites. 14% of the images labelled by the transitive method are not relevant to the label but are not excluded by the filtering algorithm. This false positive rate accounts for the majority of the discrepancy between the precision of the transitive method (over 63%) and the precision of the direct method (over 83%).

However there is a smaller but still sizeable proportion that is not due to the false positives of the filtering algorithm, but can be attributed to the inaccurate labelling of sites containing images, where those sites were labelled themselves based on clickthrough data. If over 30% of the sites are only partly or not at all relevant to the label, this will propagate to a similar proportion of mislabelled images. In table 2, there were 18% of sites partially-relevant to the label, and 6% not relevant at all. Regardless of the relevance of the site to the image, or the accuracy of the filtering algorithms, there will be a proportion of mislabelled images arising from the inaccurate site labelling, the number depending on how many images the site contains.

Dealing with the wrongly-labelled sites depends more on managing the accuracy of clickthrough data. It might be argued that insisting on numerous clickthroughs before labelling a site would assist here, although we found in the earlier experiment [1] that there was very little improvement in click relevance when the click threshold was raised. However this observation was based on image clickthrough data and should be investigated within text searches to confirm.

6 In Conclusion

In this paper we have considered how the relevance of clickthrough data to text searches differs from that of image searches, and whether this affects the accuracy of the transitive classification method. It appears that the main hurdle for the transitive method lies not in any problem with the accuracy of clickthroughs on Web pages, but rather on the filtering mechanisms that exclude images that are not pertinent to the web page itself. There is undoubtedly a level of inaccuracy in the clickthrough data that classifies the underlying Web page but the filtering algorithms present the greatest potential for improvement at this stage.

Acknowledgements

We would like to extend our thanks to everyone who helped with programming, evaluating or anything else, most especially Mark Truran for the data that underlies this work, and to Christoph Donner, Rebecca Frith, Eric Rebelos and Jan-Felix Schmakeit for their programming work.

References

1. Ashman, H., Antunovic, M., Donner, C., Frith, R., Rebelos, E., Schmakeit, J.-F., Smith, G., Truran, M.: Are clickthroughs useful for image labelling? In: Proc Web Intelligence (September 2009)
2. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proceedings of ACM SIGKDD, Boston, Massachusetts, US, pp. 407–416 (2000)
3. Chapelle, O., Zhang, Y.: A Dynamic Bayesian Network Click Model for Web Search Ranking. In: Proceedings of ACM WWW, pp. 1–10 (2009)
4. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An Experimental Comparison of Click Position-Bias Models. In: Proceedings of ACM WSDM (2008)
5. Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. ACM TOIS 23, 147–168 (2005)
6. Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y., Faloutsos, C.: Click Chain Model in Web Search. In: Proceedings of ACM WWW, pp. 11–20 (2009)
7. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting click-through data as implicit feedback. In: Proc. SIGIR, ACM, Brazil, pp. 154–161. ACM, New York (2005)
8. Scholer, F., Shokouhi, M., Billerbeck, B., Turpin, A.: Using clicks as implicit judgments: Expectations versus observations. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 28–39. Springer, Heidelberg (2008)
9. Smith, G., Ashman, H.: Evaluating implicit judgements from Web search interactions. In: Proceedings of the 1st Web Science Conference, Athens (2009)
10. Truran, M., Goulding, J., Ashman, H.L.: Co-active Intelligence for Information Retrieval. In: Proc. Multimedia 2005, pp. 547–550. ACM, New York (2005)
<http://doi.acm.org/10.1145/1101149.1101273>
11. Tsirikia, T., Diou, C., de Vries, A.P., Delopoulos, A.: Image annotation using click-through data. In: Proceedings of the 8th ACM International Conference on Image and Video Retrieval, July, Santorini, Greece, July 8-10 (2009) (to appear)
12. von Ahn, L., Dabbish, L.: Labelling Images with a Computer Game. In: Proc. SIGCHI CHI 2004, vol. 1, pp. 319–326. ACM, New York (2004)
13. Xue, G., Zeng, H., Chen, Z., Yu, Y., Ma, W., Xi, W., Fan, W.: Optimizing web search using web clickthrough data. In: Proceedings of ACM ICIKM, pp. 118–126 (2004)

A Novel Fast Inter Mode Decision Algorithm in H.264/AVC for Forest Fire Prevention Surveillance

Chen Chen¹, Ning Han¹, Chunlian Yao^{2,*}, and Yuan Li¹

¹ Department of Automation, School of Technology, Beijing Forestry University, 100083 Beijing, P.R. China

² School of communication and art, Beijing Technology and Business University, 100048 Beijing, P.R. China

chenchen870713@gmail.com, hn217@bjfu.edu.cn,
yaocl@btbu.edu.cn, bjfu.yuan.li@gmail.cn

Abstract. The new generation video coding standard H.264/AVC has been developed to achieve higher coding efficiency than previous video coding standards. Its variable block size motion compensation requires huge computation, which is a big problem for real time application. To reduce the complexity, we propose a fast inter mode decision algorithm with early mode prediction rules and early block size classification by using RD cost of the mode. Only parts of the inter prediction modes are chosen for motion estimation and rate distortion optimization. An adaptive threshold makes the proposed algorithm works smooth for different video sequences with a wide variety of activity. The experimental results indicate that the proposed algorithm can achieve 53.9% time saving in inter mode decision on average without any noticeable performance degradation, and the increased of bit rate is less than 1%.

Keywords: H.264/AVC, inter mode decision, spatial-temporal correlation, surveillance.

1 Introduction

The H.264/MPEG-4 Part-10 AVC [1] is the latest video coding standard developed by the Joint Video Team (JVT), which is an organization of the ISO Moving Picture Experts Group (MPEG) and the ITU-T Video Coding Experts Group (VCEG) [2]. It can achieve higher coding efficiency and less bit rate than other standards due to variable block size motion compensation, multiple reference frames, quarter-pixel motion vector accuracy, and so on. Compared with previous video coding standards such as H.263/MPEG-4, H.264/AVC incorporates seven block sizes from 16×16 to 4×4 for motion compensation [2,4]. It evaluates the distortion and rate of each candidate mode prior to selecting the mode for the current macroblock to achieve good rate-distortion performance. However, such exhaustively evaluating each mode entails high computational complexity which is a major problem for real-time application such as video surveillance. In forest video surveillance system, as in Fig. 1, video

* Corresponding author.

cameras are deployed in the forest, and each of them transmits the video to the terminal sever. Computational complexity reduction is desirable for low power consumption of the video sensor nodes as well as real-time video display on the screen in order to take measures as soon as a forest fire is found.

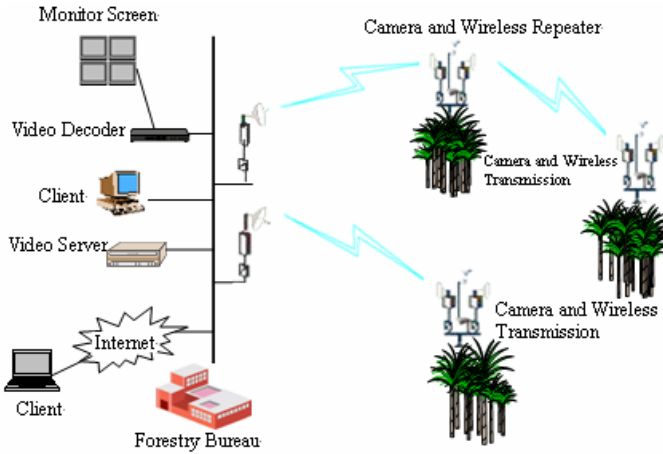


Fig. 1. Forest video surveillance system

To achieve this goal, many research have been done focusing on fast inter-frame mode decision. In [3], Kannangara et al. propose an algorithm for skip prediction in a H.264 encoder. However, this algorithm works well only for the low-motion sequences, and it still has to evaluate all the rest modes after a macroblock is chosen coding. Kim in [4] presents a fast macroblock mode prediction and decision algorithm based on contextual mode information for P-Slices and the probability characteristic of the inter mode. The concept of Region of Interest (ROI) is introduced in [5], and based on which different conditions for skip and large block mode decision are applied for the fast mode decision. Early skip mode detection, fast inter macroblock mode decision and intra prediction skip are three parts of the algorithm to reduce the computational complexity. In [7], a new fast mode decision algorithm is presented based on analysis of the histogram of the difference image between frames, which classifies the areas of each frame as active or non-active by means of an adaptive threshold technique, and it shows promising results with a time saving of over 43% for the test sequences. Common characteristic of these approaches is that the correlation among the modes of the macroblocks haven't been fully explored, which leaves a large place for further exploration.

In our proposed algorithm, we observe the modes of the macroblocks in a set of sequences encompass a wide variety of activity. After careful observation, we introduce five prediction rules by exploring the correlations among the modes of the macroblocks, and combined with an adaptive RD cost threshold TH_n to reduce the candidate modes.

The rest of the paper is organized as follows. In Section 2, we overview the rate-distortion optimized mode decision of H.264/AVC. Detailed statistical analysis is

presented in section 3. In Section 4, we propose our fast inter mode decision algorithm. In Section 5, we give our experimental results of the proposed algorithm. Finally, we draw some conclusions in Section 6.

2 Overview of Mode Decision in H.264/AVC

In H.264/AVC, in order to choose the best block size for a macroblock, RD cost of every candidate mode is calculated and the mode having the minimum RD value is chosen as the optimal mode. Fig. 2 depicts all available modes of a macroblock. The Lagrangian minimization cost function as follows,

$$J_{mode} = D + \lambda_{mode} * R$$

Where J_{mode} is the rate-distortion cost and λ_{mode} is the Lagrangian multiplier decided by QP value. D is the distortion measured as the sum of squared difference between original and reconstructed macroblock, and R reflects the total number of bits, including the macroblock header, motion vectors and the residual signal with regard to the coding mode [8].

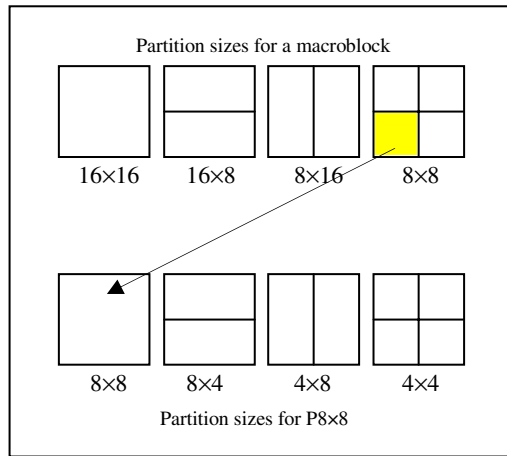


Fig. 2. Macroblock partition for inter-frame prediction

From Fig.2, we can see that the computation complexity for deciding the best inter mode is very high. Therefore, in the next section, we will develop a fast algorithm for inter mode decision to accelerate the coding process.

3 Observation and Statistical Analysis

In order to understand the percentage of each mode in the natural sequences, some experiments have been done on ten different sequences by using the H.264 reference software JM8.6[8]. Five test sequences are QCIF format and the other five sequences

are CIF format. All sequences are coding in IPPP structure with QP=28. We tested 250 frames for each sequence. The percentages of Skip mode and Inter16x16 mode are shown respectively in Fig. 3.

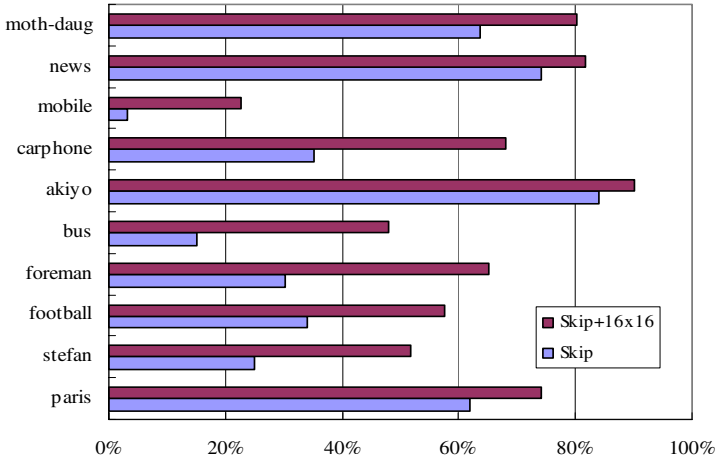


Fig. 3. Percentage of mode distribution

From Fig.3 we can easily know that Skip mode and Inter16x16 mode account for a large proportion of the mode distribution. Thus, if we can select the Skip mode and large partition size modes without going through all the candidate modes, a large quantity of coding time can be saved to meet the requirement of real time application.

3.1 Prediction Rules for Macroblock's Partition Mode

X is a macroblock in the n^{th} frame; X' is the collocated macroblock in the previous frame ($(n-1)^{th}$ frame); A is the up macroblock of X; B is the left macroblock of X. The positions of the three neighbor macroblocks are showed in Fig. 4. Statistical data suggest that the modes of A, B and X' have strong correlation with the mode of current macroblock X.

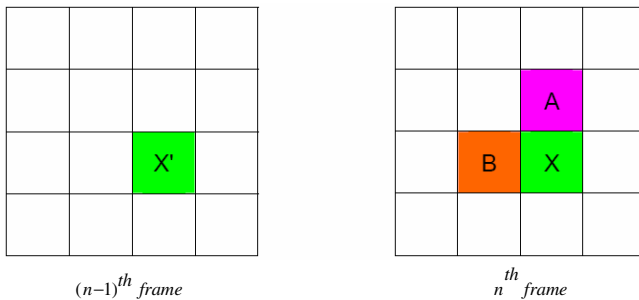


Fig. 4. Spatial and temporal neighbor macroblocks

According to the correlations between the modes of macroblock A, B and X' and the mode of current macroblock X, we develop five rules to predict the mode of macroblock X prior to motion estimation. The rules are illustrated as follows:

- Case 1: If the modes of A, B and X' are all Skip mode, the mode of X is Skip mode.
- Case 2: If the modes of A, B and X' all lie in subset {Skip, Inter16×16}, the mode of X lies in subset {Skip, Inter16×16, Inter16×8, Inter8×16}, but with one exception. That is when the modes of the three macroblocks are two Skip modes and an Inter16×16 mode, and then the mode of X lies in subset {Skip, Inter16×16}.
- Case 3: If the modes of A, B and X' all lie in subset {Inter8×8, Inter8×4, Inter4×8, Inter4×4}, the mode of X lies in subset {Inter8×8, Inter8×4, Inter4×8, Inter4×4}.
- Case 4: If at least one of A, B and X' has the mode in subset {Intra16×16, Intra4×4}, the mode of X lies in subset {Inter8×8, Inter8×4, Inter4×8, Inter4×4, Intra16×16, Intra4×4}.
- Case 5: If the modes of A, B and X' don't match any conditions of the four cases above, then full search algorithm will be processed to find the best mode.

Since there are no top macroblock or left macroblock for the macroblocks in the first line or first column, the prediction rules for macroblocks in the first line and first column are different. We use the modes of macroblocks in the previous two frames, as shown in Fig. 5, to predict the mode of current macroblock.

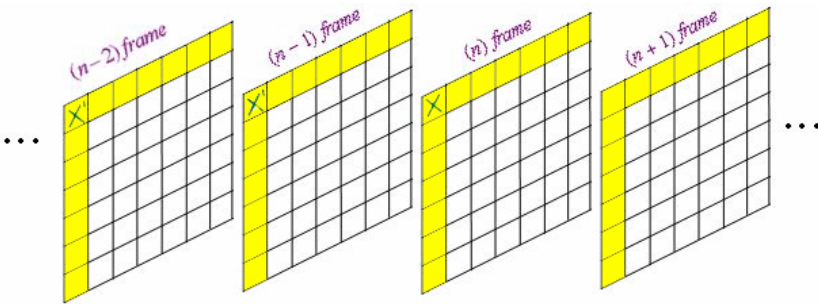


Fig. 5. The positions of macroblocks in the previous frames

The prediction rules for macroblocks in first line and first column are presented as follows:

- Case 1: If the modes of the collocated macroblocks in previous two frames are Skip modes, the mode of the current macroblock is Skip mode.
- Case 2: If the modes of the collocated macroblocks in previous two frames lie in subset {Skip, Inter16×16}, the mode of the current macroblock lies in subset {Skip, Inter16×16, Inter16×8, Inter8×16}.
- Case 3: If the modes of the collocated macroblocks in previous two frames lie in subset {Inter8×8, Inter8×4, Inter4×8, Inter4×4, Intra16×16, Intra4×4}, the mode of the current lies in the same subset.

Case 4: If the modes of the collocated macroblocks in previous two frames don't match any conditions of the three cases, then full search algorithm will be processed to find the best mode.

Once we predicted which subset the macroblock belongs to, we code the macroblock with all the modes in the subset, and choose the mode with minimum RD cost as the best mode.

With the above prediction rules, we test the classification accuracy by using a few sequences. Table 1 reflects some statistical results of 6 sequences with 200 frames. In the table, Err Ratio is the percentage of the optimal encoding modes of macroblocks which are different from that obtained from exhaustive mode selection in all the macroblocks. Full Search Ratio is the percentage of macroblocks should be coded by exhaustive mode selection in all the macroblocks. Direct Skip Ratio is the percentage of Skip mode macroblocks that can be directly predicted by Case 1 in all Skip mode macroblocks.

Table 1. Statistical results of some sequences by using prediction rules

Sequence	Format	Err Ratio %	Direct Skip Ratio %	Full Search Ratio %
Akiyo.qcif	IPPPP	0.8	88.9	6.2
Coastguard.qcif	IPPPP	2.8	10.5	59.6
Carphone.qcif	IPPPP	3.7	35.7	58.9
News.qcif	IPPPP	3.0	65.4	20.2
Foreman.cif	IPPPP	3.6	24.6	44.5
Stefan.cif	IPPPP	3.5	29.4	53.6

As we can see from the table that sequences with high motion and rich details, there are still a large proportion of macroblocks needed to be coded with full search after performing the prediction rules. In addition, although some macroblocks can be predicted to be in certain subset, all the modes in the subset are tried in order to find the optimal encoding mode. Thus for those sequences with high motion and rich details, the computational complexity is only reduced a little. In this case, fast mode decision algorithm should go further.

3.2 Mode Classification by RD Cost

While encoding different video sequences with various QP values, we find that the RD cost of different modes vary a lot. We test a few sequences with different QP values and Table 2 lists the means of RD cost for several modes in different sequences. From the table, it is found that in inter modes, the RD cost for Skip mode has minimum mean, the RD cost of the large size macroblock modes (16×16, 16×8 and 8×16) have approximate means, but they are very different from those of small size macroblock modes (8×8, 8×4, 4×8, 4×4). So, the fact can also be used to decide a Skip mode macroblock at early stage and decide the possible modes subset for current macroblock.

After performing the prediction rules, we calculate the RD cost of Skip mode $J(Skip)$ for current macroblock using formula (1) and means of RD cost for

each mode in the previous frame, then compares $J(Skip)$ with the mean of Skip macroblocks' RD cost $Mean(Skip)$ to decide whether the current macroblock is a Skip macroblock, which is shown in formula (2).

$$J(Skip) = SSD(S, C, Skip | QP) . \tag{1}$$

$$MODE \begin{cases} = Skip, & J(Skip) \leq Mean(Skip) \\ \neq Skip, & J(Skip) > Mean(Skip) \end{cases} . \tag{2}$$

There still might be some Skip macroblocks haven't been decided after performing formula (2), so we compare $J(Skip)$ with $J(16 \times 16)$ in formula (3) as supplement for Skip mode detection. Since this step is performed after predicting possible modes and motion estimation for Inter16x16, it will only save the computation of mode 16x8 and 8x16, if the macroblock is decided as a Skip macroblock in this step.

$$MODE \begin{cases} = Skip, & J(Skip) \leq J(16 \times 16) \\ \neq Skip, & J(Skip) > J(16 \times 16) \end{cases} . \tag{3}$$

Table 2. Mean values of RD cost for several modes

Sequence	QP	Skip	Inter16x16	Inter16x8	Inter8x16	Inter8x8
<i>Stefan</i> (CIF)	28	11948.5	14222.4	15281.8	14163.5	19849.3
	32	8820.4	12022.1	17047.1	17710.3	26407.8
	36	9173.2	18551.4	24578.6	27155.2	412371.7
<i>Bus</i> (CIF)	28	4914.9	6848.8	8043.6	7507.2	10890.3
	32	6143.2	8606.5	12302.9	12004.0	20109.8
	36	6954.4	22069.1	25448.4	29616.3	36184.6
<i>Paris</i> (CIF)	28	4408.6	7078.5	6753.3	7277.6	11673.8
	32	6199.3	12255.0	13847.2	18649.4	26335.2
	36	6843.6	14580.7	21023.8	24180.2	35779.0
<i>Coastguard</i> (QCIF)	28	5342	7446	8138	8214	9569
	32	7019.3	8150.2	11046.4	13282.0	15286.3
	36	9276.0	11885.3	15341.4	14766.9	22368.0
<i>News</i> (QCIF)	28	2896.4	4141.5	7279.7	6516.5	10124.8
	32	3628.5	6323.0	8598.6	9121.2	15517.6
	36	3926.0	8519.2	12405.7	15614.2	24479.4
<i>Mobile</i> (QCIF)	28	8561.6	11422.1	11797.0	13016.9	14561.7
	32	10642.7	14312.6	17253.1	17982.2	23939.7
	36	13123.7	18282.3	25501.2	24084.0	34415.8

Most of the Skip macroblocks can be predicted at early stage with the above two steps. For the remaining macroblocks in the frame, we compare the RD cost of Inter16x16 and Inter8x8 for the macroblock by formula (4) to decide which mode subset current macroblock belongs to.

$$MODE \begin{cases} \in \text{SubsetA}(\text{Skip}, \text{Inter16} \times 16, \text{Inter16} \times 8, \text{Inter8} \times 16), J \leq TH_n \\ \in \text{SubsetB}(\text{Inter8} \times 8, \text{Inter8} \times 4, \text{Inter4} \times 8, \text{Inter4} \times 4, \text{Intra16}, \text{Intra4}), J > TH_n \end{cases} \cdot (4)$$

Due to the similarities between successive video frames, the coding cost of a macroblock in the temporally-preceding frame can be used to predict the coding cost of collocated macroblock in current frame. So we set the TH_n as an adaptive threshold, which can be calculated by using the RD cost of Inter16×16 and Inter8×8 modes in the previous frame. In formula (5), parameter α and β are set as 0.4 and 0.6 respectively according to our statistical analysis.

$$TH_n = \alpha \cdot \text{Mean}(\text{Inter16} \times 16)_{n-1} + \beta \cdot \text{Mean}(\text{Inter8} \times 8)_{n-1} \cdot (5)$$

Furthermore, in the two mode subsets, the occurring probabilities of the modes are not uniform. Usually, Skip and Inter16×16 have a large proportion in Subset A, Inter8×8 is the higher probability mode in Subset B, Intra4 and Intra16 occupy a very small proportion in inter frame coding. Therefore, in the proposed algorithm, we prioritize the modes so that mode with the highest probability will be tried first, followed by the second highest probable mode, and so on. During this process, the computed RD cost will be checked against a content adaptive RD cost threshold, which is obtained from the mean of the already coded macroblock in the previous frame, to decide if the mode decision process should be terminated before trying the remaining modes in the subset. In this way we can avoid trying many unlikely coding modes.

4 Fast Inter Mode Decision

The proposed algorithm is summarized as follows:

Step 1: Calculate the mean of RDcost for every modes ($\text{Mean}(\text{mode}_i)$) in the previous frame, and obtain the frequency of every modes ($\text{Freq}(\text{mode}_i)$). Calculate TH_n by using formula (5).

Step 2: Using the prediction rules to predict which subset the macroblock belongs to. If match Case 1, go to Step 8. If match Case 2, go to Step 5. If match Case 3, go to Step 7. If match Case 4, go to Step 7. If match Case 5, go to Step 3.

Step 3: Perform motion estimation for Inter16×16 mode and calculate RD cost $J(16 \times 16)$.

Step 4: If $J(16 \times 16) > TH_n$, set $MODE \in \text{Subset B}$, then go to Step 7; Otherwise, if $J(16 \times 16) \leq TH_n$, set $MODE \in \text{Subset A}$, then go to Step 6.

Step 5: Calculate the RD cost of Skip mode $J(\text{Skip})$. If $J(\text{Skip}) \leq \text{Mean}(\text{Skip})$, go to Step 8. Otherwise, try remaining modes in the subset until all the modes are tried, go to Step 9.

Step 6: Calculate the RD cost of Skip mode $J(\text{Skip})$. If $J(\text{Skip}) \leq \text{Mean}(\text{Skip})$, go to Step 8. Otherwise, if $J(\text{Skip}) \leq J(16 \times 16)$, go to Step 8.

Step 7: Try the modes in the subset in descend order of $\text{Freq}(\text{mode}_i)$, if $J(\text{mode}_i) \leq \text{Mean}(\text{mode}_i)$, set $MODE = \text{mode}_i$, go to Step 9; Otherwise, try remaining modes in the subset until all the modes are tried, go to Step 9.

Step 8: Set MODE= Skip. (MODE is the optimal coding mode)

Step 9: Encode current macroblock with MODE.

This algorithm is for macroblocks which not lie in first line or first column. The macroblocks in the first line or first column are coded as normal after performing prediction rules. The flowchart of the algorithm is as follows:

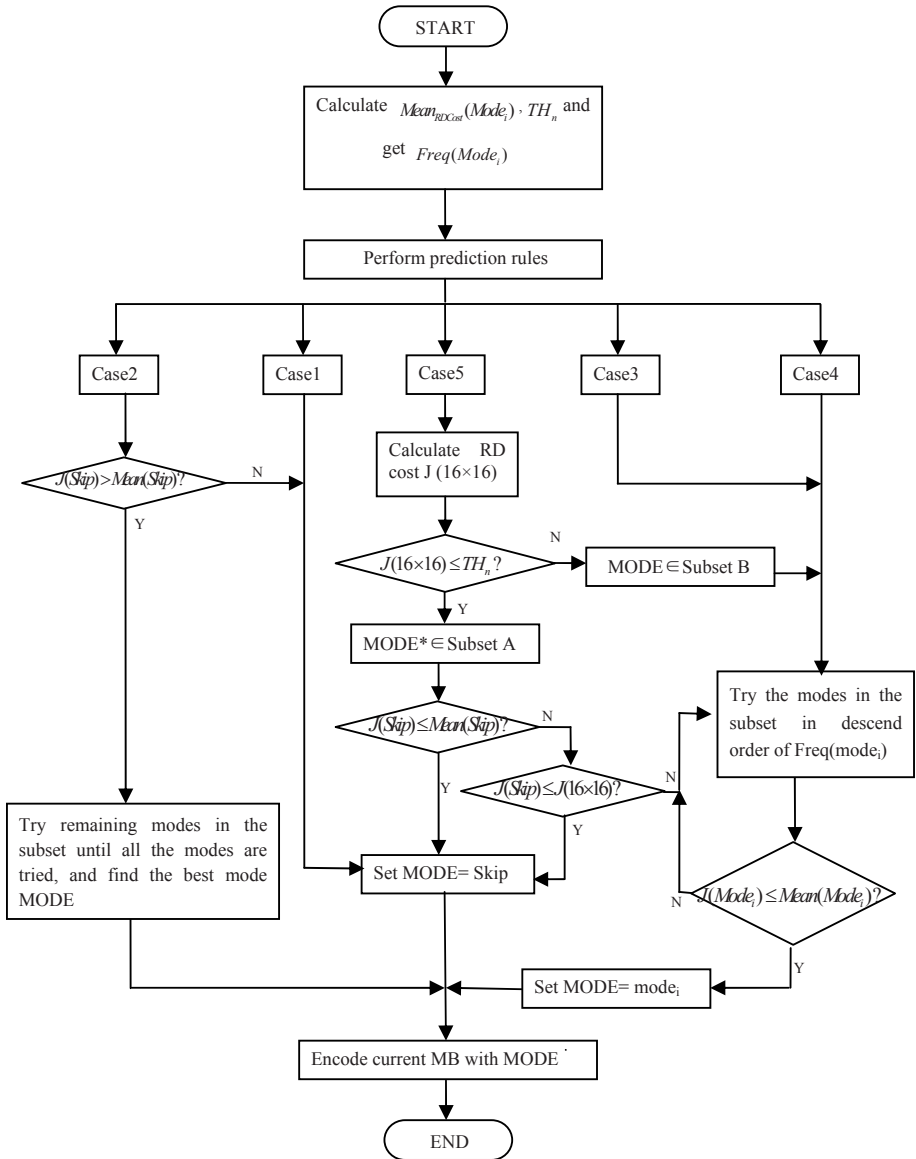


Fig. 6. Flowchart of the algorithm

In order to enhance the robustness of the algorithm, P frames with the original full search mode decision is conducted periodically. The user can set the period themselves according to their requirement.

5 Experimental Results

Our proposed algorithm was implemented into the reference software JM8.6 provided by JVT [6], tested on three standard QCIF sequences (Akiyo, Carphone and Mobile) and three standard CIF sequences (Foreman, Paris and Stefan). Each sequence has 150 frames. And all experimental conditions are described as follows: IPPP sequences, Fast_Motion_Estimation disabled, CAVLC, search range is ± 16 , RD optimization is off, frame rate is 30 and the platform is Intel Pentium IV1.73GHz PC with 1024MB memory. The test results are list in Table 3.

$$\Delta SNR = SNR_{proposed} - SNR_{JM8.6}$$

$$\Delta BitR = (BitRate_{proposed} - BitRate_{JM8.6}) / BitRate_{JM8.6} \times 100\%$$

$$\Delta Time = (Time_{proposed} - Time_{JM8.6}) / Time_{JM8.6} \times 100\%$$

Table 3. Experimental results with several different sequences

Sequence	QP	SNR(Y) (dB)	Δ SNR(Y) (dB)	BitR (kbit/s)	Δ BitR (%)	Time (sec)	Δ Time (%)
Foreman (cif)	28	36.91	-0.07	362.98	+0.21	2035.6	-45.2
	32	34.83	-0.04	216.45	-0.03	2053.3	-46.6
	36	32.81	-0.04	140.06	+0.36	2130.1	-43.7
	40	30.77	-0.03	95.28	+0.12	1864.5	-48.3
Paris (cif)	28	35.37	-0.06	529.31	+0.48	1554.3	-56.2
	32	33.06	-0.07	303.85	+0.83	1593.9	-57.1
	36	30.92	-0.05	176.44	-0.16	1424.5	-60.3
	40	28.84	-0.03	111.24	+0.67	1399.1	-58.9
Stefan (cif)	28	35.41	-0.09	1071.44	+0.53	2110.2	-44.5
	32	32.21	-0.13	541.62	-0.32	2008.5	-45.0
	36	29.43	-0.07	299.75	-0.24	2116.7	-47.4
	40	26.67	-0.11	190.28	+1.13	2089.3	-47.1
Akiyo (qcif)	28	38.25	-0.01	26.09	+0.23	198.6	-83.6
	32	36.57	-0.01	16.28	+0.30	207.3	-84.3
	36	35.08	-0.00	11.67	-0.45	173.6	-86.8
	40	33.73	+0.01	9.71	+0.36	192.1	-83.4
Carphone (qcif)	28	37.35	-0.03	85.65	+0.37	282.7	-68.6
	32	34.79	-0.04	50.09	+0.74	247.3	-72.2
	36	32.53	-0.03	32.61	+0.82	255.6	-71.5
	40	30.46	-0.05	21.63	+0.69	231.5	-73.0
Mobile (qcif)	28	33.35	-0.05	419.92	+1.08	638.6	-34.2
	32	30.18	-0.09	209.96	+2.48	706.4	-33.7
	36	27.69	-0.08	111.07	+2.32	660.2	-36.2
	40	25.30	-0.07	65.23	+0.25	562.5	-38.8

In addition, we apply the proposed algorithm to process part of the recorded video of a real forest fire scene from Linfen forest farm in Shanxi province of China. Fig. 7 shows a few images of the video sequence, and Fig. 8 compares SNR decoded by the



Fig. 7. Examples of original (top) and decoded (bottom) frames of the forest fire video sequences. Frame number 200, 250, 300 and 350 of a sequence are shown.

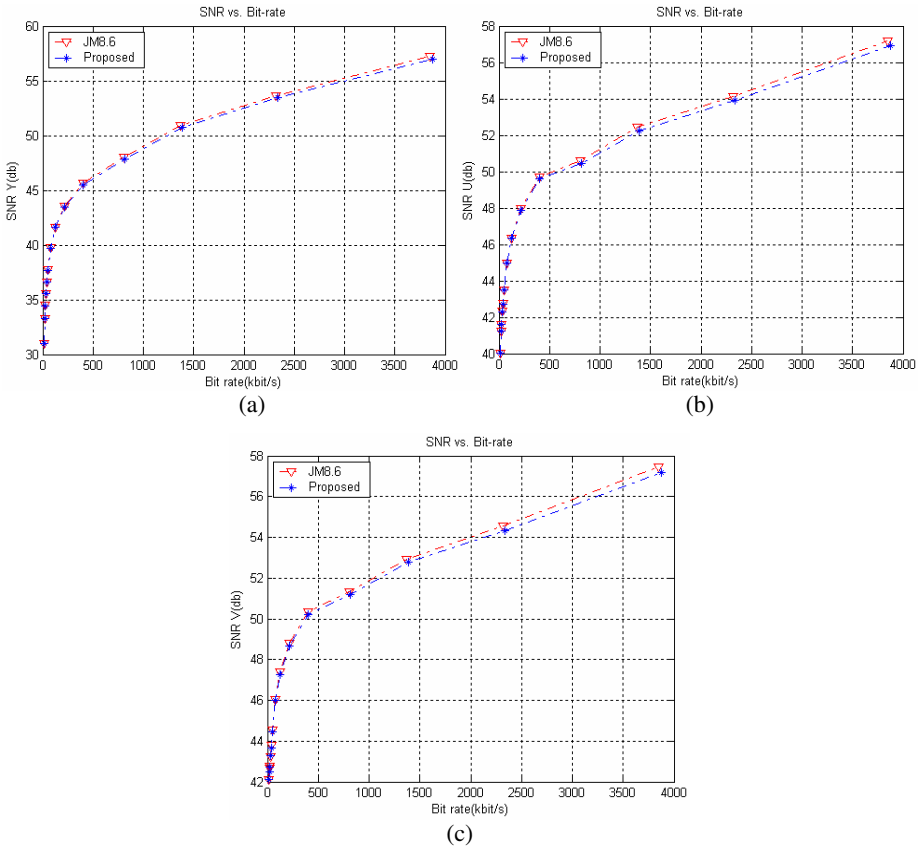


Fig. 8. SNR comparison between JM8.6 and the proposed algorithm (a) SNR Y (b) SNR U (c) SNR V

original JM8.6 baseline profile with that decoded by the proposed algorithm. From the images, we can see that the degradation of the images is hardly noticeable.

6 Conclusion

In this paper, a new fast Inter mode decision algorithm based on statistic and adaptive adjustment has been proposed. By fully exploring the correlation among the modes of the macroblocks and using RD cost to classify the mode category, the algorithm can reduce the computational time significantly with slight loss of PSNR and bit rate increase.

Acknowledgments

This research was supported by the Introduction of Advanced International Forestry Science and Technology Program from the State Forestry Administration, China (Grant No. 2005-4-04) and the National College Students Innovative Research Program from the Ministry of Education of China (Grant No. GCS08016).

References

1. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264|ISO/IEC 14496-10 AVC), JVT-G050r1, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG (May 2003)
2. La, B., Jeong, J., Choe, Y.: Selective intra-prediction Skip Algorithm for inter-frame coding in H.264/AVC. *IEICE Electronics Express* 6(2), 71–76 (2009)
3. Kannangara, C.S., Richardson, I.E.G., Bystrom, M., Solera, J., Zhao, Y., MacLennan, A., Cooney, R.: Low Complexity Skip Prediction for H.264 through Lagrangian Cost Estimation. *IEEE Transactions on Circuits and Systems for Video Technology* 16(2), 202–208 (2006)
4. Kim, B.-G., Song, S.-K.: Enhanced Inter Mode Decision Based on Contextual Prediction for P-Slices in H.264/AVC Video Coding. *ETRI Journal* 28(4), 425–434 (2006)
5. Na, T., Lee, Y., Lim, J., Joo, Y., Kim, K., Byun, J., Kim, M.: A Fast Macroblock Mode Decision Scheme Using ROI-Based Coding of H.264|MPEG-4 Part 10 AVC for Mobile Video Telephony Application. In: *Proc. SPIE*, vol. 7073, pp. 707301 (2008) DOI: 10.1117/12.795614
6. H. 264/AVC reference software JM8.6, <http://iphome.hi.de/suehring/tml/download>
7. Nieto, M., Salgado, L., Cabrera, J., García, N.: Fast mode decision on H.264/AVC baseline profile for real-time performance. *Journal of Real-Time Image Processing* 3(1-2), 61–75 (2008)
8. Grecos, C., Yang, M.: Fast Mode Prediction for the Baseline and Main Profiles in the H.264 Video Coding Standard. *IEEE Trans. Multimedia* 8(6), 1125–1134 (2006)

A Method for Analyzing Software Faults Based on Mining Outliers' Feature Attribute Sets

Jiadong Ren^{1,2}, Changzhen Hu², Kunsheng Wang³, and Di Wu^{1,4}

¹ College of Information Science and Engineering,
YanShan University,
QinHuangDao 066004, China
jdren@ysu.edu.cn

² School of Computer Science and Technology,
Beijing Institute of Technology,
Beijing, 100081, China
chzhoo@bit.edu.cn

³ China Aerospace Engineering Consultation Center,
Beijing, 100037, China
kshwang@126.com

⁴ Department of Information and Electronic Engineering,
Hebei University of Engineering,
Handan 056038, China
bestmoogoo@163.com

Abstract. Faults analysis is a hot topic in the field software security. In this paper, the concepts of the improved Euclidian distance and the feature attribute set are defined. A novel algorithm MOFASIED for mining outliers' feature attribute set based on improved Euclidian distance is presented. The high dimensional space is divided into some subspaces. The outlier set is obtained by using the definition of the improved Euclidian distance in each subspace. Moreover, the corresponding feature attribute sets of the outliers are gained. The outliers are formalized by the attribute sets. According to the idea of the anomaly-based intrusion detection research, a software faults analysis method SFAMOFAS based on mining outliers' feature attribute set is proposed. The outliers' feature attributes can be mined to guide the software faults feature. Experimental results show that MOFASIED is better than the distance-based outlier mining algorithm in performance test and time cost.

1 Introduction

Along with the swift development of network technique, the problem of software security has drawn more and more attention. However, most of the software security problems come from software themselves. Some of the problems are due to the safety function that is still far from completion. There are also some problems arouse by the security flaws which are designed artificially and deliberately. Hence, a new technique is sorely needed to detect and analyze the software or

the abnormalities in software design. Furthermore, the security performance of the software can be improved.

At present, studies on software security analysis are numerous and can be grouped into four general categories: Petri nets based analysis method [1], fault tree based analysis method [2], rough set based analysis method [3] and structure information based analysis method [4]. Deng presented an efficient genetic algorithm for system-level diagnosis [5]. A set of equations that govern the statuses of the units in a system are proposed. Based on this, the authors present a new genetic algorithm for the fault diagnosis of diagnosable systems by designing a novel fitness function. The false negative is decreased validly. Reference [6] discussed quantitative fault tree analysis to software development for space camera. A modeling method for bottom event was presented. It is based on vague set to calculate the relative significance of bottom event and the occurrence probability of top event. Then, analyzed result from the model is used to guide design of software reliability.

The above two algorithms have very good performance to analyze the cause of software faults. However, there are some limitations in the analysis of the dependency between the software faults. The efficiency and accuracy of these algorithms deteriorate sharply. A software reliability growth model considering failure dependency [7] was introduced by Zhao. Based on Non-Homogeneous Poisson Process, a software reliability growth model that considers software failure dependency and differences of testing and operational environment are realized. Furthermore, the reliability of the software can be enhanced greatly.

Recently, much attention has been given to the problem of outlier detection, in contrast to most KDD tasks, such as clustering and classification, outlier detection aims to detect outliers which behave in an unexpected way or have abnormal properties. For many applications, such as fault detection in software, it is more interesting to find the rare events than to find the common ones from a knowledge discovery standpoint. Studying the extraordinary behaviors of outliers can help us uncover the valuable information hidden behind them.

In this article, in order to obtain the meaning of the outliers, the concept of the feature attribute set is presented. According to the outliers' abnormal characters, the outlier mining algorithm is introduced into the software faults analysis. Moreover, to improve the security performance of the software, a novel software faults analysis method based on mining outliers' feature attribute set is discussed.

The remainder of this paper is organized as follows: Section 2 describes the problem definitions. Section 3 gives the MOFASIED algorithm. Section 4 designs the SFAMOFAS method. Section 5 contains experimental results and Section 6 offers our conclusions.

2 Problem Definitions

Assume that $D = \{A_1, A_2, \dots, A_d\}$ be a information system. The name of each dimension is denoted as A_i . We represent $d(A_i)$ as the all values set of A_i . Let

$P_{+D} = d(A_1)d(A_2)d(A_d)$ be the d-th dimensional space that is defined for D. On condition that I, C, O are the divisions of D, Q is the predicate defined for 2^{P+I} , the mark condition is denoted as Q, then quaternion group (I,C,O,Q) is the division of D. P_{+I}, P_{+C}, P_{+O} are called the group mark subspace, state subspace, observation subspace, respectively.

Suppose that R is the subset of P_{+D} . For all $o \in P_{+D}$, we represent $\prod_D(o)$ as the projection that the object o in the P_{+D} , and $\prod_D(R) = \{\prod_D(o) : o \in R\}$. Let $T = \{T_1, T_2, \dots, T_u\}$ be a observation sequence of R. We have $\bigcup_{i=1}^u T_i = \prod_o(R)$. For all $1 \leq i \leq j \leq u, T_i \cap T_j = \emptyset$. The observation window is called the each element of T. If $R = \bigcup_{i=1}^n R_i$, where $\prod_o(R_i) \subseteq T_i$ and $o \in R_i$, and $\prod_C(o)$ is the outlier of the $\prod_C(R_i)$, then o is also the outlier of the observation window T_i .

For the sparseness of the certain areas, the local neighborhood numbers that consist of the near neighbour points are more than the corresponding numbers in the dense areas. Hence, Euclidian distance is modified. The concept of the improved Euclidian distance is proposed.

Definition 2.1. For data set S, the number of the objects is n. $y_i, y_j \in S$, $M(i)$ and $M(j)$ are called the average values between $y_i (i = 1, 2, \dots, n)$, $y_j (j = 1, 2, \dots, n)$ and the other objects, respectively. The improved Euclidian distance between the object y_i and y_j is defined as being:

$$d_{ij}(y_i, y_j) = \frac{|y_i - y_j|}{\sqrt{M(i)M(j)}} \tag{1}$$

It is easy to see, the numerator of $d_{ij}(y_i, y_j)$ is the common Euclidian distance. The denominator is the numerical value. As a result, the improved Euclidian distance that is meet the requirements of the definition of distance can be easily proved. Above all, $d_{ij}(y_i, y_j) \geq 0$, if and only if y_i is equal to y_j , the requirement of the nonnegative property of the distance is satisfied. Secondly, $d_{ij}(y_i, y_j)$ is equals to $d_{ji}(y_j, y_i)$, it can serve the purpose of the symmetric property of the distance. Finally, the sum of $d_{ij}(y_i, y_j)$ and $d_{it}(y_i, y_t)$ is greater than or equal to $d_{jt}(y_j, y_t)$. It is satisfactory for the triangle inequality.

Definition 2.2. For $D_1 \subseteq D$, D_1 is arbitrary non-empty attribute set. U is the subset of P_{+X} , the object o of U is an outlier. If $Y \subset D_1$, $\prod_Y(O)$ is not an outlier in $\prod_Y(U)$, we call o is a nontrivial outlier of the feature attribute set X. Otherwise, o is an ordinary outlier. In case of o is a nontrivial outlier of the feature attribute set D_1 , the outlier o's feature attribute set is defined as D_1 .

For mining the feature attribute sets to the corresponding outliers, the source and meaning of the outliers can be obtained. It is important to note that an outlier may have more than one corresponding feature attribute set. For example, in the process of the software failure, it has two failure descriptions. One is that the variable type is changed in the process of modeling. The other is that the coordinate value is overlapped. Thus the feature attribute sets of the software failure contains the above two reasons. The abnormal behaviors can be determined according to only the attribute that the variable type is changed in the process of modeling or the event that the coordinate value is overlapped.

3 The MOFASIED Algorithm

A novel algorithm MOFASIED for mining outliers' feature attribute set based on improved Euclidian distance is presented. In our approach, the high dimensional space is divided into some subspaces. The outlier set is obtained by using IEDLOMA algorithm in each subspace. Furthermore, the outliers' feature attribute sets in the observation windows are gained. The outliers can be formalized by the attribute sets. MOFASIED is explained in detail in this section.

3.1 The IEDLOMA Algorithm

An improved-Euclidian-distance-based local outlier mining algorithm IEDLOMA is proposed. The outliers in the observation windows can be obtained by applying IEDLOMA algorithm. The distance between the objects is calculated according to the definition of the improved Euclidian distance. The distribution of the objects in data set will be uniform. The distance between the objects in dense area is increased. On the contrary, the corresponding distance in sparse area is decreased. The IEDLOMA algorithm is given as follows:

Algorithm IEDLOMA(S, e, p, q, O)

Input: Data set S; object e; the number of objects p; distance_constraint q

Output: The outlier set O.

- (1) Compute the distance between the objects in data set S by the definition of the improved Euclidian distance;
- (2) For each object e in S
- (3) If exists p objects that the distance to e is greater than the distance_constraint q then
e is judged DB(p,q) abnormality, namely, e is an outlier;
- (4) Add the outlier e to the outlier set O;
- (5) End For

3.2 Mining Outliers' Feature Attribute Sets

In this section, the outliers' feature attribute sets are mined. Assume that each attribute A_i corresponds to the interval V_i . The outlier set $U = \{o_1, o_2, \dots, o_n\}$ and radius r are given. We denote the subinterval that the center is o and the radius is r as $V_i^{o,r}$. The objects in the subspaces $V_1 V_2 \dots V_{i-1} V_i^{o,r} V_{i+1} \dots V_d$ is represented as $L_i^{o,r}$. For any k feature attributes $\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\}$, where $i_1 \leq i_2 \leq \dots \leq i_k$. $L_{i_1, i_2, \dots, i_k}^{o,r}$ is called the all objects in subspaces $V_1 V_2 \dots V_d$, we have:

$$V_j = \begin{cases} V_j & (j \notin \{i_1, i_2, \dots, i_k\}) \\ V_i^{o,r} & (\text{else}) \end{cases} \quad (2)$$

For MOFASIED, the feature attributes are established according to an order of user-defined importance. The feature set is represented by binary. We use array $K[d]$ to record the feature attributes, where $1 \leq i \leq d$. $K[i]$ storages feature set with i feature attributes. The i-th dimensional subspace is recorded by C_i . MOFASIED can be described as follows:

Algorithm MOFASIED(S, e, p, q, O, o, I, d, n, K[d])

Input: Data set S; object e; the number of objects p; distance_constraint q; the outlier set O; the outlier o; the number of feature attributes i; data set dimension d; neighborhoods n

Output: The outliers' corresponding feature attribute sets K[d].

- (1) The outlier set O is obtained by IEDLOMA algorithm;
- (2) For each object o in the outlier set O
- (3) All $L_i^{o,r}$ can be gained by traversing the data set S, where $1 \leq i \leq d$.
 Meanwhile, the number of neighborhoods of the outlier o in $L_i^{o,r}$ is scanned;
- (4) If the corresponding number of neighborhoods is less than n then o is an outlier in, $\{A_i\}$ is recorded as the feature attribute set of o
 Else $L_i^{o,r}$ is incorporated into C_1 ;
- (5) C_k can be generated by C_{k-1} , where $\forall 2 \leq k \leq d$.
 If C_k is not empty, decide the object o whether an outlier in $L_{i_1, i_2 \dots i_k}^{o,r}$
- (6) If the object o is an outlier in $L_{i_1, i_2 \dots i_k}^{o,r}$ then
 Record $\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\}$ is the feature attribute set, delete the $L_{i_1, i_2 \dots i_k}^{o,r}$ form C_k , until C_k is empty;
- (7) The array K[d] records the feature attribute set of outlier o;
- (8) End For

4 The MOFASIED Algorithm

According to the idea of the anomaly-based intrusion detection research, a new software faults analysis method is presented by using MOFASIED algorithm. The outliers' feature attributes can be mined to direct the software faults feature. The flowchart of SFAMOFAS is shown as Fig. 1.

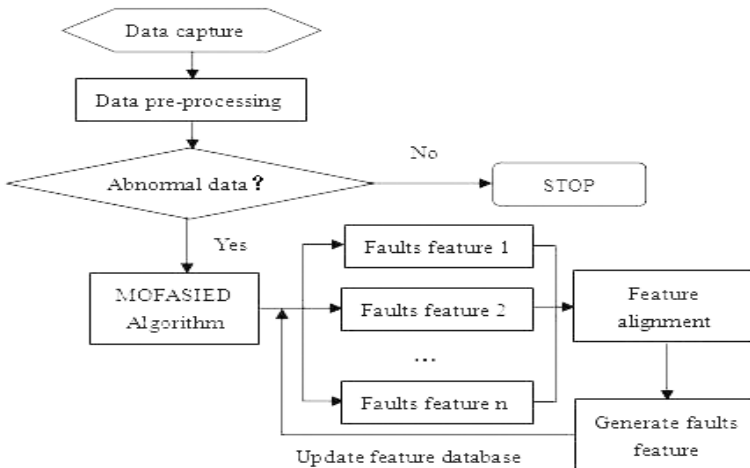


Fig. 1. MOFASIED based software faults analysis method flowchart

The function of each sub-module is described below:

- (1) Data capture: It is responsible to monitor and control the data in software section. Moreover, the data is collected and raced to the next sub-module.
- (2) Data pre-processing: According to the filter criteria of abnormal identification, delete the noise points, and the abnormal data can be gained.
- (3) Feature alignment: The faults features can be obtained by using MOFASIED algorithm. For the same kind of faults, the results of the feature alignment are gained by comparing the given feature with the feature of the data that after data pre-processing.
- (4) Product faults feature: Converse the result of the faults feature alignment to the rules that meet the standards of software faults feature.
- (5) Update feature database: The new software faults features that are extracted are used to update the feature database.

5 Experimental Results and Analysis

In order to evaluate the performance of our algorithm MOFASIED, we compare the MOFASIED with the presented distance-based outlier mining algorithm DOMA. The MOFASIED is written in VC++6.0 and our experiments are performed on the 2.4GHz CPU, 2GB main memory and Microsoft XP Professional.

In this section, we conduct our experiments on both synthetic DataSet1 and real datasets KDDCUP'99. There are 30 dimensions and 800K data set size in Synthetic DataSet1. However, the data set size and the dimension can be controlled by inputting parameters. It is about five million connection records in datasets KDDCUP'99. Attacks fall into four main categories, such as denial-of-service DOS, unauthorized access from a remote machine R2L, unauthorized access to local super-user privileges U2R and Probing.

In our experiments, we select two mixed attack data sets and four single differences types attack data sets. The proportion of normal records to intrusion records is 1001. The descriptions of the above data sets can be shown as Table 1.

Table 1. Test data set detail account

Data set	Normal record	Intrusion record	Total	Type
Mix1	9900	100	10000	Mixed
Mix2	19800	200	20000	Mixed
PROBE	44241	447	44688	PROBE
R2L	19765	200	19965	R2L
U2R	28597	289	28886	U2R
DOS	48899	495	49394	DOS

Essential parameters are described as follows: the number of objects p is 12 and the distance constraint q is 0.31. For testing the performance of MOFASIED, we compare MOFASIED with DOMA in two aspects, running time and performance analysis.

5.1 Running Time Analysis

In order to test the running time with data set size in MOFASIED, the synthetic data sets are generated, the sizes of them are 100K, 200K, 300K and 400K, respectively. The dimension of each data set is 30. The experimental result is shown as Fig. 2.

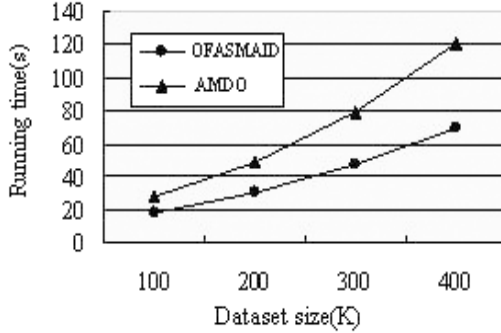


Fig. 2. Running times of MOFASIED and DOMA

Intuitively, Fig. 2 shows that running time with different data set size by two algorithms. As the data set size increases, it can be seen that the running times of the two algorithms increase linearly. Meanwhile, compared with DOMA algorithm, the advantage of MOFASIED is conspicuous.

For MOFASIED, the improved-Euclidian-distance-based local outlier mining algorithm IEDLOMA does not detect all the objects in data set D. It refers to the objects only in each local observation window. Hence, as contrasted with DOMA, the time cost is decreased effectively.

5.2 Performance Test Analysis

In this section, there are two main standards for detecting faults. One is DT (Detection Rate), the other is FPR (False Positive Rate). We must give consideration to the DT and FPR while detecting faults. The DT and FPR on six test data set by using MOFASIED and DOMA. The experimental result is shown as Table 2.

Table 2. DT and FPR on six test data sets

Data set	DOMA		IEDLOMA	
	DT%	FPR%	DT%	FPR%
Mix2	88.3	2.35	91.4	2.06
PROBE	95.1	5.66	97.3	2.33
R2L	43.5	20.75	50.5	13.28
U2R	28.4	24.69	29.1	11.45
DOS	92.2	1.69	93.0	1.48

It can be seen from Table 2, no matter how the mixed attack data sets or the single type attack data sets, the DT and FPR of MOFASIED are both better than the corresponding in DOMA. The Performance tests are more stable and reliable.

According to the definition of the improved Euclidian distance in MOFASIED, the distance between the objects is computed according to the definition of the improved Euclidian distance. The distribution of the objects in data set will be uniform. The distance between the objects in dense area is increased. On the contrary, the corresponding distance in sparse area is decreased. Ultimately, our algorithm is better in the detection rate and the false positive rate than DOMA.

6 Conclusions

In this paper, a novel algorithm MOFASIED for mining outliers' feature attribute set based on improved Euclidian distance is discussed. The high dimensional space is divided into some subspaces. The distance between the objects is computed by applying the definition of the improved Euclidian distance in each subspace. The outlier set can be obtained. Moreover, the concept of the feature attribute set is discussed to formalize the outliers. According to the idea of the anomaly-based intrusion detection research, the outliers' feature attributes can be mined to direct the software faults feature by using an effective method FAMOFAS, which is based on mining outliers' feature attribute set. Experimental results and analysis show that our algorithm is better than DOMA in performance test and time cost. Finally, generation of quality of software faults feature can be improved effectively.

Acknowledgment

This work is partially supported by the National High Technology Research and Development Program ("863" Program) of China (No. 2009AA01Z433) and the Natural Science Foundation of Hebei Province P. R. China(No. F2008000888). The authors also would like to express their gratitude to the reviewers, whose valuable comments are very helpful in revising the paper.

References

1. Soares, M.S., Julia, S., Vrancken, J.: Real-time Scheduling of Batch Systems Using Petri nets and Linear Logic. *Journal of Systems and Software* 81(11), 1983–1996 (2008)
2. Ortmeier, F., Schellhorn, G.: Formal Fault Tree Analysis-Practical Experiences. *Electronic Notes in Theoretical Computer Science*, vol. 185, pp. 139–151 (2007)
3. Li, J.J., Li, S.H., Fang, Z.C.: The Research of Fault Diagnosis in Aluminum Electrolysis Based on Rough Set. In: *Proc. of the 2008 ISECS International Colloquium on Computing, Communication, Control, and Management*, Guang Zhou, pp. 162–166 (2008)

4. Luan, S.M., Dai, G.Z.: An Approach to Diagnosing a System with Structure Information. *Chinese Journal of Computers* 28(5), 801–808 (2005)
5. Deng, W., Yang, X.F., Wu, Z.F.: An Efficient Genetic Algorithm for System-Level Diagnosis. *Chinese Journal of Computers* 30(7), 1116–1125 (2007)
6. Li, Y., Xu, S.Y., Han, C.S., Yu, T., Xing, Z.B.: Application of quantitative fault tree analysis to software development for space camera. *Optics and Precision Engineering* 16(11), 2181–2187 (2008)
7. Zhao, J., Zhang, R.B., Gu, G.C.: Study on Software Reliability Growth Model Considering Failure Dependency. *Chinese Journal of Computers* 30(10), 1714–1721 (2007)

Unifying Web-Scale Search and Reasoning from the Viewpoint of Granularity

Yi Zeng¹, Yan Wang¹, Zhisheng Huang^{2,3}, and Ning Zhong^{1,4}

¹ International WIC Institute, Beijing University of Technology
Beijing, 100124, P.R. China

{yzeng,wang.yan}@emails.bjut.edu.cn

² Department of Artificial Intelligence, Vrije University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

huang@cs.vu.nl

³ School of Computer Science and Engineering, Jiangsu University of Science and
Technology, Jiangsu, 212003, P.R. China

⁴ Department of Life Science and Informatics, Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi 371-0816, Japan

zhong@maebashi-it.ac.jp

Abstract. Considering the time constraints and Web scale data, it is impossible to achieve absolutely complete reasoning results. Plus, the same results may not meet the diversity of user needs since their expectations may differ a lot. One of the major solutions for this problem is to unify search and reasoning. From the perspective of granularity, this paper provides various strategies of unifying search and reasoning for effective problem solving on the Web. We bring the strategies of multilevel, multiperspective, starting point from human problem solving to Web scale reasoning to satisfy a wide variety of user needs and to remove the scalability barriers. Concrete methods such as network statistics based data selection and ontology supervised hierarchical reasoning are applied to these strategies. The experimental results based on an RDF dataset shows that the proposed strategies are potentially effective.

1 Introduction

The assumption of traditional reasoning methods do not fit very well when facing Web scale data. One of the major problems is that acquiring all the relevant data is very hard when the data goes to Web scale. Hence, unifying reasoning and search is proposed [1]. Under this approach, the search will help to gradually select a small set of data (namely, a subset of the original dataset), and provide the searched results for reasoning. If the users are not satisfied with the reasoning results based on the sub dataset, the search process will help to select other parts or larger sub dataset prepared for producing better reasoning results [2]. One detailed problem is that how to search for a good or more relevant subset of data and do reasoning on it. In addition, the same strategy may not meet the diversity of user needs since their backgrounds and expectations may differ a lot. In this paper, we aim at solving this problem.

Granular computing, a field of study that aims at extracting the commonality of human and machine intelligence from the viewpoint of granularity [23], emphasizes that human can always focus on appropriate levels of granularity and views, ignoring irrelevant information in order to achieve effective problem solving [34]. This process contains two major steps, namely, the search of relevant data and problem solving based on searched data. As a concrete approach for problem solving based on Web scale data, the unification of search and reasoning also contains these two steps, namely, the search of relevant facts, and reasoning based on rules and searched facts. A granule is a set of elements that are drawn together by their equality, similarities, indistinguishability from some aspects (e.g. parameter values) [5]. Granules can be grouped into multiple levels to form a hierarchical granular structure, and the hierarchy can also be built from multiple perspectives [3]. Following the above inspirations, the web of data can be grouped together as granules in different levels or under different views for searching of subsets and meeting various user needs. From the perspective of granularity, we provide various strategies for unifying user driven search and reasoning under time constraints. From the multilevel point of view, in order to meet user needs in different levels, unifying search and reasoning with multilevel completeness and multilevel specificity are proposed. Furthermore, from the multiperspective point of view, the unifying process can be investigated based on different perspectives of the knowledge source. We also propose unifying search and reasoning with a starting point, which is inspired by the basic level advantage from cognitive psychology [6], to achieve diversity and scalability.

The rest of this paper focuses on introducing various strategies for unifying search and reasoning from the viewpoint of granularity: Section 2 introduces the multilevel completeness strategy. Section 3 introduces unifying strategy with multilevel specificity. Section 4 discusses the starting point strategy. Section 5 investigates on the multiperspective strategy. For each strategy introduced in this paper, we provide some preliminary experimental results based on a semantic Web dataset SwetoDBLP, an RDF version of the DBLP dataset [7]. Finally, Section 6 discusses some related work and makes concluding remarks.

2 Multilevel Completeness Strategy

Web scale reasoning is very hard to achieve complete results, since the user may not have time to wait for a reasoning system going through the complete dataset. If the user does not have enough time, a conclusion is made through reasoning based on a searched partial dataset, and the completeness is not very high since there are still some sets of data which remain to be unexplored. If more time is allowed, and the reasoning system can get more sub datasets through search, the completeness can migrate to a new level since the datasets cover wider range.

There are two major issues in this kind of unifying process of search and reasoning: (1) Since under time constraint, a reasoning system may just can handle a sub dataset, methods on how to select an appropriate subset need to be developed. (2) Since this unification process require user judges whether the completeness of reasoning results is good enough for their specific needs, a prediction

method for completeness is required. We name this kind of strategy as unifying search and reasoning with multilevel completeness, which provides reasoning results in multiple levels of completeness based on the searched sub dataset under time constraints, meanwhile, provides prediction on the completeness value for user judges. In this paper, we develop one possible concrete solution.

For issue (1), searching for a more important sub dataset for reasoning may be a practical approach to select the subset effectively [11], and may be an approach to handle the scalability issue, since in most cases, the amount of important data is relatively small. Under the context of the Semantic Web, the semantic dataset can be considered as a graph that contains a set of nodes (subjects and objects in RDF dataset) and a set of relations (predicates in RDF dataset) on these nodes. Hence, in this paper, we borrow the idea of “pivotal node” from network science [8], we propose a network statistics based data selection strategy. Under this strategy, we use the node degree (denoted as $degree(n)$) to evaluate the importance of a node in a dataset. The nodes with relatively high value of node degree are selected as more important nodes and grouped together as a granule for reasoning tasks. There might be many types of predicates which are associated with the nodes in the RDF dataset, and different meanings of the various predicates may meet user needs from different perspectives. According to a specific need from a perspective, (which will be explained in detail in Section 5), we choose one type of predicate to investigate on the importance of a node. When we only consider this type of predicate and neglect other types, a subgraph of the original RDF dataset can be selected out. In this subgraph, the node degree considering a special type of predicate P can be denoted as $degree(n, P)$.

For issue (2), here we give a formula to produce the predicted completeness value ($PC(i)$) when the nodes which satisfy $degree(n, P) \geq i$ (i is a nonnegative integer) have been involved.

$$PC(i) = \frac{|N_{rel(i)}| \times (|N_{sub(i)}| - |N_{sub(i')}|)}{|N_{rel(i)}| \times (|N| - |N_{sub(i')}|) + |N_{rel(i')}| \times (|N_{sub(i)}| - |N|)}, \quad (1)$$

where $|N_{sub(i)}|$ represents the number of nodes which satisfy $degree(n, P) \geq i$, $|N_{rel(i)}|$ is the number of nodes which are relevant to the reasoning task among the involved nodes $N_{sub(i)}$, and $|N|$ is the total number of nodes in the dataset. The basic idea is that, first we can obtain a linear function which go through $(|N_{sub(i)}|, |N_{rel(i)}|)$ and $(|N_{sub(i')}|, |N_{rel(i')}|)$ (i' is the last assigned value of $degree(n, P)$ for stopping the reasoning process before i). Knowing $|N|$ in the dataset ($|N|$ only needs to be acquired once and can be calculated offline), by this linear function, we can predict the number of satisfied nodes in the whole dataset, then the predicted completeness value can be acquired.

As an illustrative example, we take the reasoning task “Who are authors in Artificial Intelligence (AI)?” based on the SwetoDBLP dataset. For the most simple case, following rule can be applied for reasoning to find relevant authors:

$$haspaper(X, Y), contains(Y, \text{“Artificial Intelligence”}) \rightarrow author(X, \text{“AI”})$$

where $haspaper(X, Y)$ denotes that the author X has a paper titled Y , while $contains(Y, \text{“Artificial Intelligence”})$ denotes that the title Y contains the term

“Artificial Intelligence”, and $author(X, \text{“AI”})$ denotes that the author X is an author in the field of AI . Since the SwetoDBLP contains too many publications (More than 1,200,000), doing reasoning based on a dataset like this may require an unacceptable period of time, it is better that more important authors could be provided to the user first. Here we choose the predicate that indicate an author has a coauthor (denoted as P_{cn}). Under this perspective, the authors with more coauthors, namely, has a higher value of $degree(n, P_{cn})$, are more important. In order to illustrate the levels of completeness, we randomly choose some $degree(n, P_{cn})$ to stop the reasoning process, as shown in Table 1. The reasoning process will start from the nodes with the biggest value of $degree(n, P_{cn})$, reduce the value gradually as time passed by, and will stop at the chosen $degree(n, P_{cn})$ for user judges. In order to meet users’ specific needs on the levels of completeness value, using the proposed completeness prediction method introduced above, the prediction value has also been provided in Figure 1. This prediction value serves as a reference for users to judge whether they are satisfied. If more time is allowed and the user has not been satisfied yet, more nodes are involved, one can get reasoning results with higher levels of completeness. In this way, we provide solutions for the various user needs.

Table 1. Unifying search and reasoning with multilevel completeness and anytime behavior

$degree(n, P_{cn})$ value to stop	Satisfied authors	AI authors
70	2885	151
30	17121	579
11	78868	1142
4	277417	1704
1	575447	2225
0	615124	2355

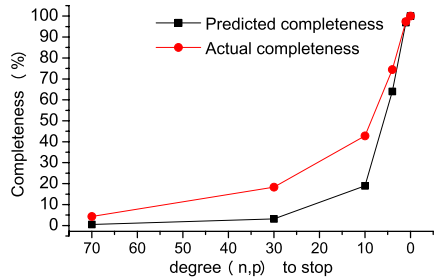


Fig. 1. Comparison of predicted and actual completeness value

3 Multilevel Specificity Strategy

Reasoning results can be either very general or very specific. If the user has not enough time, the search and reasoning process will just be on a very general level. And if more time is available, this process may go to a more specific level which contains results in a finer level of grain size (granularity). Namely, the unification of search and reasoning can be with multilevel specificity, which provides reasoning results in multiple levels of specificities under time constraints.

The study of the semantic networks emphasizes that knowledge is stored as a system of propositions organized hierarchically in memory [9]. The concepts in various levels are with different levels of specificities. Hence, the hierarchical knowledge structure can be used to supervise the unification of search and reasoning with multilevel specificity. In this process, the search of sub datasets is

based on the hierarchical relations (e.g. sub class of, sub property of, instance of, etc.) among the nodes (subjects and objects in RDF) and is forced to be related with the time allowed. Nodes which are not sub classes, instances or sub properties of other nodes will be searched out as the first level for reasoning. If more time is available, more deeper levels of specificity can be acquired according to the transitive property of these hierarchical relations. The specificity will just go deeper for one level each time before the next checking of available time (Nodes are searched out based on direct hierarchical relations with the nodes from the former direct neighborhood level).

As an illustrative example, we use the same reasoning task in the upper section. For the very general level, the reasoning system will just provide authors whose paper titles contain “Artificial Intelligence”, and the reasoning result is 2355 persons (It seems not too many, which is not reasonable.). Since in many cases, the authors in the field of AI do not write papers whose titles include the exact term “Artificial Intelligence”, they may mention more specific terms such

Table 2. Answers to “Who are the authors in Artificial Intelligence?” in multiple levels of specificity according to the hierarchical knowledge structure of Artificial Intelligence

Specificity	Relevant keywords	Number of authors
Level 1	Artificial Intelligence	2355
Level 2	Agents	9157
	Automated Reasoning	222
	Cognition	19775
	Constraints	8744
	Games	3817
	Knowledge Representation	1537
	Natural Language	2939
	Robot	16425

Level 3	Analogy	374
	Case-Based Reasoning	1133
	Cognitive Modeling	76
	Decision Trees	1112
	Proof Planning	45
	Search	32079
	Translation	4414
	Web Intelligence	122

Table 3. A comparative study on the answers in different levels of specificity

Specificity	Number of authors	Completeness
Level 1	2355	0.85%
Level 1,2	207468	75.11%
Level 1,2,3	276205	100%

as “Agent”, “Machine Learning”, etc. If more time is given, answers with a finer level of specificity according to a hierarchical domain ontology of “Artificial Intelligence” can be provided. Based on all the AI related conferences section and subsection names in the DBLP, we create a “three-level Artificial Intelligence ontology” automatically (This ontology has a hierarchical structure representing “Artificial Intelligence” related topics. Topic relations among levels are represented with “`rdfs:subClassOf`”), and we utilize this ontology to demonstrate the unification of search and reasoning with multilevel specificity¹. The rule for this reasoning task is:

$$\text{hasResttime}, \text{haspaper}(X, Y), \text{contains}(Y, H), \text{topics}(H, \text{“AI”}) \rightarrow \text{author}(X, \text{“AI”})$$

where *hasResttime* is a dynamic predicate which denotes whether there is some rest time for the reasoning task², *topics*(*H*, “AI”) denotes that *H* is a related sub topic from the hierarchical ontology of AI. If the user allows more time, based on the “`rdfs:subClassOf`” relation, the subtopics of AI in Level 2 of the ontology will be used as *H* for reasoning to find more authors in the field of AI. Further, if the user wants to get results finer than Level 2, then the subtopics in Level 3 are used as *H* to produce an even more complete result list. As shown in Tables 2 and 3, based on the hierarchy of Artificial Intelligence, Since Levels 2 and 3 contain more specific sub branches, it is not surprising that one can get more authors when deeper levels of terms are considered, hence, the completeness of the reasoning result also goes to higher levels, as shown in Table 3.

4 Starting Point Strategy

Psychological experiments support that during problem solving, in most cases, people try to investigate the problem starting from a “basic level” (where people find convenient to start according to their own background knowledge), in order to solve the the problem more efficiently [6]. In addition, concepts in a basic level are used more frequently than others [10]. Following this idea, we define that during the unification of search and reasoning process on the Web for a specified user, there is a starting point (denoted as *SP*), which consists of a user identity (e.g. a user name, a URI, etc.) and a set of nodes which serve as the background for the user (e.g. user interests, friends of the user, or other user familiar or related information). A starting point is used for refining the unification of search and reasoning process in the form that the user may prefer.

Following the idea of starting point, the search of important nodes for reasoning can be based on the following strategies:

¹ Here we ignore the soundness of this ontology, which is not the focus of this paper (Supporting materials on how we build the ontology can be found from : <http://www.iwici.org/user-g.>). One can choose other similar ontologies instead.

² For implementation, logic programming languages such as Prolog does not allow a dynamic predicate like *hasResttime*. But we can consider *resttime*(*T*) as a counter which would return a number. Then, we can check the number to know whether there is any rest time left. Namely: $\text{resttime}(T), T > 0 \rightarrow \text{hasResttime}$.

Table 4. A comparative study of the multilevel completeness strategy without and with a starting point (User name: John McCarthy)

Completeness	Authors (coauthor numbers) without a starting point	Authors (coauthor numbers) with a starting point
Level 1 $degree(n, P_{cn}) \geq 70$	Carl Kesselman (312) Thomas S. Huang (271) Edward A. Fox (269) Lei Wang (250) John Mylopoulos (245) Ewa Deelman (237) ...	Hans W. Guesgen (117) * Carl Kesselman (312) Thomas S. Huang (271) Edward A. Fox (269) Lei Wang (250) John Mylopoulos (245) ...
Level 2 $degree(n, P_{cn}) \in [30, 70)$	Claudio Moraga (69) Virginia Dignum (69) Ralph Grishman (69) Biplav Srivastava (69) Ralph M. Weischedel (69) Andrew Lim (69) ...	Virginia Dignum (69) * John McCarthy (65) * Aaron Sloman (36) * Claudio Moraga (69) Ralph Grishman (69) Biplav Srivastava (69) ...
...

- Strategy 1 (Familiarity-Driven): The search process firstly select out the nodes which are directly related to the SP for the later reasoning process, and SP related results are ranked to the front of others.
- Strategy 2 (Novelty-Driven): The search process firstly select out the nodes which are not directly related to the SP , then they are transferred to the reasoning process, and SP related nodes are pushed to the end of others.

Strategy 1 is designed to meet the user needs who want to get more familiar results first. Strategy 2 is designed to meet the needs who want to get unfamiliar results first. One example for strategy 2 is that in news search on the Web, in most cases the users always want to find the relevant news webpages which have not been visited. Here we give an example using strategy 1, and this example is a synergy of the multilevel completeness strategy and the starting point strategy. Following the same reasoning task in the above sections, “John McCarthy”, is taken as a concrete user name in a SP , and his coauthors³ whom he definitely knows (with * after the names) are ranked into the top ones in every level of the “Artificial Intelligence” author lists when the user tries to stop while an arbitrary $degree(n, P_{cn})$ of the relevant nodes has been involved (Since the coauthors are all persons whom the author should know. These information helps users get more convenient reasoning results.). Some partial output in some levels is shown in Table 4. The strategy of multilevel specificity and starting point can also be

³ In this study, we represent the coauthor information for each author in an RDF file using the FOAF vocabulary “foaf:knows”. The coauthor network RDF dataset created based on the SwetoDBLP dataset can be acquired from <http://www.iwici.org/dblp-sse>. One can utilize this dataset to create a starting point for refining the reasoning process.

integrated together, which provide reasoning results based on starting point in every level of specificity to produce a more user-preferred form of results.

5 Multiperspective Strategy

User needs may differ from each other when they expect answers from different perspectives. In order to avoid the failure of understanding in one way, knowledge needs to be represented in different points of view [11]. If the knowledge source is investigated in different perspectives, it is natural that the search and reasoning results might be organized differently. Each perspective satisfies user needs in a unique way. As another key strategy, unifying search and reasoning from multiperspective aims at satisfying user needs in multiple views.

We continue the reasoning task of “Who are authors in Artificial Intelligence?”. As proposed in the above sections, we use node degree under a perspective (*degree* (n, P)) to search for a subset of the original data for reasoning. Here we consider following perspectives: the number of coauthors, and the number of publications. Firstly, We choose the perspective of the number of coauthors. From this perspective, we find following characteristics of the SwetoDBLP dataset: Coauthor number distribution is shown as in Figure 2. In the left side of Figure 3, there is a peak

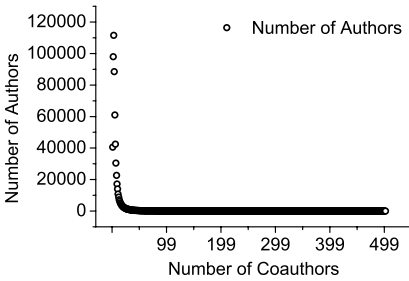


Fig. 2. Coauthor number distribution in the SwetoDBLP dataset

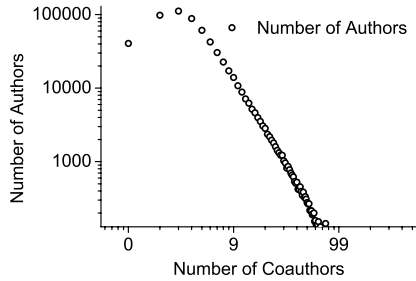


Fig. 3. log-log diagram of Figure 2

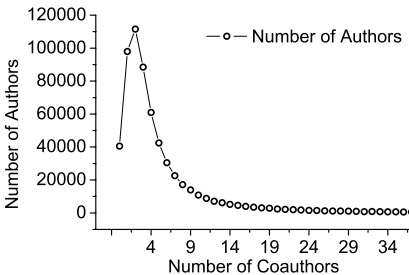


Fig. 4. A zoomed in version of Figure 2

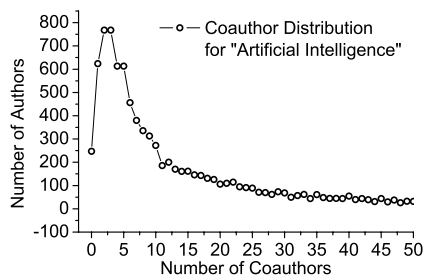


Fig. 5. A zoomed in version of coauthor distribution for “Artificial Intelligence”

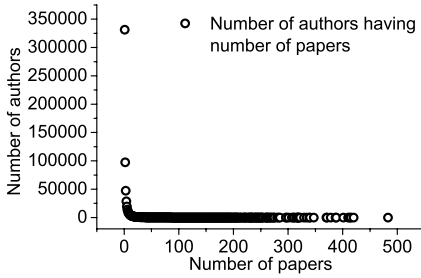


Fig. 6. Publication number distribution in the SwetoDBLP dataset

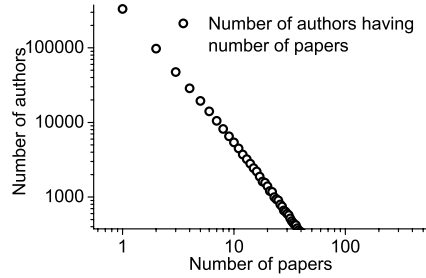


Fig. 7. log-log diagram of Figure 6

value in the distribution, and it does not appear at the point of 0 or 1 coauthor number (as shown in Figure 4). Hence, the shape of the distribution is very much like a log-normal distribution. These phenomena are not special cases that just happen to all the authors, we also observed the same phenomenon for authors in many sub-fields in computer science, such as Artificial Intelligence (as shown in Figure 5), Software Engineering, Data Mining, Machine Learning, the World Wide Web, Quantum Computing, etc. As a comparison of the coauthor number, we provide some partial results from the view point of publication number. We observe that, different from the perspective of coauthor number distribution, the publication number distribution follows very much like a power law distribution, without a peak value in the middle of the distribution curve, as shown in Figures 6 and 7.

It is clear that since the distribution of node degree under the above two perspectives are different, and for the same node, the node degree under these two perspectives are different, we can conclude that using different perspectives, both of the sequence of nodes provided for reasoning and the reasoning results are organized differently. In this way, various user needs can be satisfied.

6 Discussion and Conclusion

The study of unifying reasoning and search at Web scale [11] is the framework that this paper is based on. The strategies introduced in this paper aim at providing some possible solutions for how the unification can be done in a more user-oriented way from the viewpoint of granularity. They are developed based on many existing studies. Here we introduce two major related areas, namely, variable precision logic and previous studies on reasoning with granularity.

Variable precision logic is a major method for reasoning under time constraints, which provides two reasoning strategies, namely, variable certainty and variable specificity reasoning [12]. Concerning time constraint, given more time, a system with variable specificity can provide a more specific answer, while a system with variable certainty can provide a more certain answer [12]. Some strategies on unifying search and reasoning introduced in this paper, for example, the multilevel specificity strategy is inspired by variable specificity reasoning.

The major difference is that: variable specificity reasoning uses “if-then-unless” rule, while multilevel specificity strategy uses hierarchical knowledge structure to supervise the unification process of search and reasoning. In this paper, we did not investigate on the idea of variable certainty. Since it belongs to non-monotonic reasoning, and the certainty won’t necessarily go higher as more data is involved (since there might be contradictions [13] or inconsistency [14] on the facts, especially in the dynamic changing context of the Web). How it can be applied to a more user-centric environment still needs further investigations.

The study of reasoning with granularity starts from the logic approaches for granular computing [15][16][17], etc. Under the term of granular reasoning, it has also been studied from the perspectives of propositional reasoning [18], Aristotle’s categorial syllogism [19], and granular space [20]. These studies concentrate on the logic foundations for reasoning under multi-granularity (mainly on zoom-in and zoom-out). In this paper, our focus is on how to unify the search and reasoning process from the viewpoint of granularity, namely, how to search for a good subset of the original dataset, and do reasoning on the selected dataset based on the idea of granularity. Besides the inspiration from granular computing [3][4], especially granular structures [3]. The strategies proposed in this paper are also inspired from Cognitive Psychology studies on human problem solving (e.g. starting point) [6][21]. Further, we concentrate on how granularity related strategies can help to effectively solve Web scale reasoning problems according to different user context and time constraints.

We also need to point out that although the strategies introduced in this paper are inspired by some basic strategies in granular computing, the granular structures, more specifically granular knowledge structures that are mentioned in this paper are different from previous studies [3][22]. In granular computing, granules are organized hierarchically from larger grain sizes to smaller ones (or the other way around), and the granules in coarser levels contain the ones in finer levels. In this study, although granules are still in a hierarchy, the granules does not contain each other. In the multilevel completeness strategy, granules are organized into different levels by the node degree under a perspective, granules with higher value of $degree(n, P)$ do not contain those with lower values. In the multilevel specificity strategy, although the hierarchical knowledge structures of Artificial Intelligence has a typical granular structure (All the subtopics are covered under the terms one level coarser than them.), the granular structure of the reasoning results based on this hierarchy is different from the granular structures studied previously [3][22], since the results which were got from the coarser levels cannot cover finer levels (The reason is that if the user does not have enough time, nodes in finer levels, such as authors of “Decision Trees”, will not be selected for the reasoning task whether they are AI authors.).

As an approach for incomplete reasoning at Web scale, unifying search and reasoning from the viewpoint of granularity provides some strategies which aim at removing the diversity and scalability barriers for Web reasoning.

For the diversity issue: The strategy of starting point focuses on user specific background and the unification process is familiarity driven or novelty driven,

and is obviously user oriented. Multilevel completeness strategy is with anytime behavior [23], and provides predictions of completeness for user judges when the user interact with the system. Multilevel specificity strategy emphasizes on reasoning with multiple levels of specificity and users can choose whether to go into more specific or more general levels. Multiperspective strategy attempts to meet various user needs from multiple perspectives.

For the scalability issue: In the multilevel completeness strategy, although the partial results may have low completeness, more important results have been searched out and ranked to the top ones for reasoning based on their higher values of $degree(n, P)$. In other words, more important results are provided as a possible way to solve the scalability problems. The starting point strategy also provides two methods to select important nodes for reasoning. The multilevel specificity strategy concentrates on the appropriate levels of specificity controlled by the knowledge hierarchy and does not get into unnecessary levels of data. Hence, under limited time, the reasoning task and time is reduced.

Since user needs are very related to the satisfaction of reasoning results, in future studies, we would provide a comparison from the user perspective on the effects of multiple strategies mentioned in this paper. We would also like to investigate in great details on how these strategies can be combined together to produce better solutions⁴. Since the unification of Web scale search and reasoning from the viewpoint of granularity brings many human problem solving strategies to Web reasoning, it can be considered as an effort towards Web intelligence [24].

Acknowledgements

This study is supported by the European Commission under the 7th framework programme, Large Knowledge Collider (FP7-215535) and Vrije University Amsterdam, the Netherlands. This paper was prepared when Yi Zeng was visiting Vrije University Amsterdam. The authors would like to thank Stefan Schlobach, Christophe Guéret on their constructive comments and suggestions.

References

1. Fensel, D., van Harmelen, F.: Unifying reasoning and search to web scale. *IEEE Internet Computing* 11(2), 96, 94–95 (2007)
2. Yao, Y.: The art of granular computing. In: Kryszkiewicz, M., Peters, J.F., Rybiński, H., Skowron, A. (eds.) *RSEISP 2007. LNCS (LNAI)*, vol. 4585, pp. 101–112. Springer, Heidelberg (2007)
3. Yao, Y.: A unified framework of granular computing. In: *Handbook of Granular Computing*, pp. 401–410. Wiley, Chichester (2008)
4. Zhang, B., Zhang, L.: *Theory and Applications of Problem Solving*, 1st edn. Elsevier Science Inc., Amsterdam (1992)

⁴ Further investigations can be tracked through the USer-G (Unifying Search and Reasoning from the viewpoint of Granularity) website <http://www.iwici.org/user-g>

5. Yao, Y.: Perspectives of granular computing. In: Proceedings of 2005 IEEE International Conference on Granular Computing, vol. 1, pp. 85–90 (2005)
6. Rogers, T., Patterson, K.: Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General* 136(3), 451–469 (2007)
7. Aleman-Meza, B., Hakimpour, F., Arpinar, I., Sheth, A.: Swetodblp ontology of computer science publications. *Journal of Web Semantics* 5(3), 151–155 (2007)
8. Barabási, A.: *Linked: The New Science of Networks*. Perseus Publishing (2002)
9. Collins, A.M., Quillian, M.R.: Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior* 8, 240–247 (1969)
10. Wisniewski, E., Murphy, G.: Superordinate and basic category names in discourse: A textual analysis. *Discourse Processing* 12, 245–261 (1989)
11. Minsky, M.: *The Emotion Machine: commonsense thinking, artificial intelligence, and the future of the human mind*. Simon & Schuster, New York (2006)
12. Michalski, R., Winston, P.: Variable precision logic. *Artificial Intelligence* 29(2), 121–146 (1986)
13. Carnielli, W., del Cerro, L., Lima-Marques, M.: Contextual negations and reasoning with contradictions. In: Proceedings of the 12th International Joint Conference on Artificial Intelligence, pp. 532–537.
14. Huang, Z., van Harmelen, F., ten Teije, A.: Reasoning with inconsistent ontologies. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, pp. 454–459 (2005)
15. Hobbs, J.: Granularity. In: Proceedings of the 9th International Joint Conference on Artificial Intelligence, pp. 432–435 (1985)
16. Liu, Q., Wang, Q.: Granular logic with closeness relation λ and its reasoning. In: Ślęzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 709–717. Springer, Heidelberg (2005)
17. Zhou, B., Yao, Y.: A logic approach to granular computing. *The International Journal of Cognitive Informatics & Natural Intelligence* 2(2), 63–79 (2008)
18. Murai, T., Resconi, G., Nakata, M., Sato, Y.: Granular reasoning using zooming in & out: Propositional reasoning. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) RSFDGrC 2003. LNCS (LNAI), vol. 2639, pp. 421–424. Springer, Heidelberg (2003)
19. Murai, T., Sato, Y.: Granular reasoning using zooming in & out: Aristotle’s categorical syllogism. *Electronic Notes in Theoretical Computer Science* 82(4), 186–197 (2003)
20. Yan, L., Liu, Q.: Researches on granular reasoning based on granular space. In: Proceedings of the 2008 International Conference on Granular Computing, vol. 1, pp. 706–711 (2008)
21. Wickelgren, W.: Memory storage dynamics. In: *Handbook of learning and cognitive processes*, pp. 321–361. Lawrence Erlbaum Associates, Hillsdale (1976)
22. Zeng, Y., Zhong, N.: On granular knowledge structures. In: Proceedings of the first International Conference on Advanced Intelligence, pp. 28–33 (2008)
23. Vanderveen, K., Ramamoorthy, C.: Anytime reasoning in first-order logic. In: Proceedings of the 9th International Conference on Tools with Artificial Intelligence, pp. 142–148 (1997)
24. Zhong, N., Liu, J., Yao, Y.: *Web Intelligence*, 1st edn. Springer, Heidelberg (2003)

The Quest for Parallel Reasoning on the Semantic Web

Peiqiang Li¹, Yi Zeng¹, Spyros Kotoulas², Jacopo Urbani², and Ning Zhong^{1,3}

¹ International WIC Institute, Beijing University of Technology
Beijing 100124, P.R. China

{lpeiqliang,yzeng}@emails.bjut.edu.cn

² Department of Computer Science, VU University Amsterdam
De Boelelaan 1081, 1081HV, Amsterdam, The Netherlands

{kot,j.urbani}@few.vu.nl

³ Department of Life Science and Informatics, Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi 371-0816, Japan

zhong@maebashi-it.ac.jp

Abstract. Traditional reasoning tools for the Semantic Web cannot cope with Web scale data. One major direction to improve performance is parallelization. This article surveys existing studies, basic ideas and mechanisms for parallel reasoning, and introduces three major parallel applications on the Semantic Web: LarKC, MaRVIN, and Reasoning-Hadoop. Furthermore, this paper lays the ground for parallelizing unified search and reasoning at Web scale.

1 Introduction

The Semantic Web provides services for exchanging and utilizing data, information and knowledge in different forms on the Web. Its goal is to organize the web of data through its semantics in a machine-understandable format that can be utilized for problem solving on the Web. In addition, the Semantic Web is dedicated to providing an information sharing model and platform that is convenient for both human and machine to understand and cooperate [1]. Reasoning engines are one of the foundational systems that enable computer systems to automatically reason over Semantic Web data according to some inference rules.

Currently, Semantic Web reasoners are usually deployed on a single machine. This solution is applicable when the datasets are relatively small. In a Web context, it is unreasonable to expect fast enough reasoning mechanisms that work on a single processor, especially when dealing with Web-scale RDF and OWL datasets. Firstly, on the Web, large corpuses of semantically rich data is found, posing new challenges to processing techniques. Secondly, the Web of data is growing very fast and is dynamic. Thirdly, rules might be represented in different forms, which requires a reasoning task to do preprocessing and coordination with other reasoners [2,3]. In the face of these new requirements, existing reasoning methods have lost their effectiveness. Besides developing new forms of reasoning, another solution is to parallelize the reasoning process [3,4,5].

The rest of the paper is organized as follows. Section 2 summarizes the current status for parallel reasoning outside the Semantic Web. Section 3 describes in detail the problem of parallel reasoning on the Semantic Web. Section 4 describes LarKC, MaRVIN and Reasoning-Hadoop as three implementations. Finally, based on the analysis of present parallel reasoning techniques on the Semantic Web, we make some preliminary discussion on a related topic which is yet to be fully explored: parallelizing ReaSearch (the unification of reasoning and search) at Web scale.

2 Parallel Reasoning

It is commonly agreed that the goal of parallel reasoning is to accelerate the reasoning process. The major differences between a single processor-based method and a parallel environment-based method can be summarized from two perspectives: platform architecture and algorithm.

2.1 Platform Architecture Perspective

From the perspective of instruction and data streams, three types of architectures are considered, namely, SIMD (single instruction stream, multiple data streams), MISD (multiple instruction streams, single data streams), and MIMD (multiple instruction streams, multiple data streams) [6]. From the perspective of memory, three types of architectures are considered, namely, SMP (Symmetric Multiprocessing), DMP (Distributed Memory Parallel), and HMS (Hierarchical Memory systems) [7]. Systems with an SMP architecture are composed of several interconnected processors that have a shared memory. In DMP, each processor maintains its own memory. Local memory access is fast but distributed memory access has to be done through an interconnect network and is slow [7,8]. HMS is a hybrid architecture which is an integration of SMP and DMP. In this architecture, clusters of nodes have an SMP architecture internally and a DMP architecture across clusters. Commonly-found clusters of multi-processor or multi-core nodes are also considered to have an HMS architecture [7].

2.2 Algorithm Perspective

Besides the differences on system architecture, a set of parallel reasoning dispatching algorithms have also been developed. The goal for adopting parallel reasoning is to accelerate the reasoning speed. However, as the number of machines increases, the reasoning speed may increase only sublinearly because of communication overhead. Besides, each machine may not be assigned the same amount of processing and thus, not run in its maximum capacity [9]. Thus, one of the key issues for parallel reasoning algorithms is minimizing communication while maintaining an even distribution of load (load balancing) [10]. In order to solve this problem, several algorithms were proposed. Here we introduce two

of them. Although they are proposed before and outside the Semantic Web, we believe they may bring some inspiration.

The load balancing algorithm in [11] uses a matrix to denote the distance from two arbitrary reasoning engines. Another matrix is used to represent the mappings (task transfer from one to another) between two arbitrary reasoning engines [11]. The reasoning engine selection strategy can be summarized as follows: the algorithm adopts the shortest path algorithm. Firstly, it tries to map among the nearest reasoning engines, and after the mapping, remove those which have been occupied. Then, the mapping will try those reasoning engines which have the second shortest path. This progress repeats until all the reasoning engines are assigned. The advantage for this algorithm is that it can drastically reduce communication, its weakness is that it is centralized [11].

The load sharing by state (LDSHBS) algorithm [12] restricts the communication to the nearest reasoning engines. This algorithm uses a tree-structured allocator which has a structure similar to a binary tree. All reasoning engines are at the lowest level. Some nodes supervise reasoning engines in the system. If these nodes detect that some reasoning engines are overloaded, they will remove some tasks from them. If they find some engines are underutilized, the nodes will allocate some additional tasks to them [12]. The advantage of this algorithm is that it can allocate tasks locally and reduce communication. Besides, the algorithm also ensures that each reasoning engine gets a reasonable number of tasks. The disadvantage is that choosing an appropriate number of supervision nodes might be a hard problem [12], especially for Web scale reasoning.

3 Parallelizing Semantic Web Reasoning

Since the data on the Semantic Web is mainly presented using RDF, reasoning on the Semantic Web is mainly focused on RDFS and OWL. We can identify two major goals in Semantic Web reasoning. The first goal is to check the consistency of the web of data so that the data from different sources are well integrated [13]. The second goal is to find new facts (implicit semantic relations) based on existing facts and rules [13][14].

Compared to traditional parallel processing, the Semantic Web has some additional concerns. Firstly, there are too many nodes in a Web-scale RDF dataset, and each node may have many predicates associated with it. This makes data dependencies complex. Hence, partitioning RDF data is not easy. Secondly, configuring the parallel hardware environment may need to meet new challenges considering Web scale data. Since the dataset is dynamically changing and grows very fast, we cannot assume a static environment. Thirdly, so far, there are not many parallel reasoning algorithms which can be directly imported from other areas to solve Semantic Web parallel reasoning. Finally, load balancing on each machine is still a hard problem to solve, given the very skewed nature of Semantic Web terms. In this light, the current state-of-the art in practical parallel reasoning is rather poorly developed.

For Parallel Semantic Web Reasoning, two major trends are introduced to process reasoning in parallel, namely, data partitioning approaches, and the rule

partitioning approaches [15]. In data partitioning approaches, all rules are applied by all reasoners while data is split into smaller partitions and processed in parallel by several reasoners [15]. In rule partitioning approaches, the rules are partitioned into different reasoners to perform the reasoning tasks, and the data has to go through all the reasoners [15,16]. These studies have also shown that effective partitioning is not easy. Better partitioning methods need to be proposed. Some practical implementations of large-scale parallel reasoning are described in the following section.

4 Some Solutions

Very recently, some practical implementations for parallel Semantic Web reasoning have been published. In this section, we are going to introduce three parallel reasoning approaches for web-scale data: LarKC [3], MaRVIN [5], and Reasoning-Hadoop [4]. While all of these are in the direction of parallel reasoning on the Semantic Web, each of them has a different viewpoint.

4.1 LarKC

The Large Knowledge Collider (LarKC)¹ is an open architecture and a generic platform for massive distributed reasoning [3]. LarKC currently emphasizes on scalability through parallelization of the execution of an open set of software components. LarKC works as a scalable workflow engine for reasoning tasks. In each workflow, there are several components (plug-ins) which are responsible for diverse processing tasks, for example, identifying relevant data, transforming data, selecting data and reasoning over data. The execution of the workflow is overseen by a decider plug-in [3]. Since several plug-ins are invoked in a workflow, they can be distributed among several nodes and work in parallel [17]. LarKC parallelizes execution in the following ways:

- Invocation of plug-ins that have a parallel implementation;
- Invocation of distinct plug-ins in parallel;
- Execution of several workflows in parallel, or execution of the same workflow with different input in parallel.

Currently, a set of LarKC plug-ins already have a parallel implementation: A Sindice identifiers uses multiple threads (thus exploiting shared memory, multiple processor architectures), and a GATE transformer can run on supercomputers. MaRVIN, which will be introduced in the next section will be wrapped as a parallel and distributed reasoner for LarKC. As work in progress, USER-G (Unifying Search and Reasoning from the perspective of Granularity)² is also a series of methods that aims at working in a parallel environment for LarKC.

Currently, LarKC is considering some parallel programming models (e.g. OpenMP, HPF (High Performance Fortran), and MPI (Message Passing Interface)) and some frameworks (e.g. the Ibis framework [18]) to offer as an API

¹ <http://www.larkc.eu>

² <http://www.iwici.org/user-g>

for developing parallel components. The core mechanism of OpenMP is based on shared memory directives [19] and task decomposition, but does not provide how to do decomposition [7]. Although HPF provides specific data decomposition, the rules of HPF may cause a too high communication overhead [7]. MPI allows the developers to specify the distribution of the work and the data. How and when the communication is done can also be specified [7]. According to [7] LarKC may consider using the Ibis framework [18]. Ibis is a grid programming environment that combines portability, flexibility and high efficiency [18]. The parts which are relevant to LarKC are the Ibis portability layer (IPL) [18] and the MPJ/Ibis [20].

4.2 MaRVIN

As a part of the LarKC project, MaRVIN (Massive RDF Versatile Inference Network)³ is a parallel and distributed platform for processing large amounts of RDF data. MaRVIN is the first practical parallel reasoning implementation [5].

The work on MaRVIN is motivated by the observation that it is hard to solve Semantic Web problems through traditional divide-and-conquer strategies since Semantic Web data is hard to partition [5].

MaRVIN brings forward a method named divide-conquer-swap [5] to do inferring through forward chaining (i.e. calculate the closure of the input data). The main algorithm can be described in the following steps: First, the platform divides the input data into several independent partitions and assigns this partitions to compute nodes. Second, each compute node computes the closure of its partition using a conventional reasoner. Then, old and new data is mixed and new partitions are created in a distributed manner. This process is repeated until no new triples are derived. At this point, the full closure has been calculated.

In the context of MaRVIN, the SPEEDDATE routing strategy has been developed [5]. RDFS and OWL rules are triggered by triples that share at least one term. SPEEDDATE makes partitions in a distributed manner while increasing the triples with the same terms that belong to the same partition. This way, the number of partitioning-reasoning cycles that need to be performed to calculate the full closure is reduced.

The advantages of the MaRVIN platform are the following:

- Since the partitions are of equal size, the amount of data to be stored and the computation to be performed is evenly distributed among nodes;
- No upfront data analysis is required;
- It can support any monotonic logic by changing the conventional reasoner;
- It uses a peer-to-peer architecture. Thus, no central coordination is required;
- It shows anytime behavior, i.e. it produces results incrementally with time.

The latest experiments on real-world datasets show that MaRVIN can calculate the closure of 200M triples on 64 compute nodes in 7.2 minutes, yielding a throughput of 450K triples per second.

³ <http://www.larkc.eu/marvin/>

In terms of the definitions in sections 2 and 3, MaRVIN does data partitioning on a HMP architecture. Compared to traditional reasoners, MaRVIN shows higher loading speeds but is limited to calculating the closure of the input.

4.3 Reasoning-Hadoop

Reasoning-Hadoop [4] is a parallel rule-based RDFS/OWL reasoning system built on the top of the Hadoop framework [21]. Hadoop is an opensource framework mainly used for massive parallel data processing initially developed by Yahoo! and now hosted by the Apache foundation.

Hadoop implements the MapReduce programming model. The MapReduce programming model was developed by Google [22] and it requires that all the information is encoded as a set of pairs of the form $\langle key, value \rangle$. A typical MapReduce algorithm takes as input a set of pairs, processes them using two functions, *map* and *reduce*, and returns some new pairs as output. The program execution is handled by the framework which splits the input set in subsets and assigns computation to nodes in the network [22].

In the reasoning-hadoop project⁴ RDFS and OWL-Horst forward reasoning has been implemented with a sequence of MapReduce algorithms. In terms of the definitions in sections 2 and 3, Reasoning-Hadoop does both data partitioning and rule partitioning on a DMP architecture.

In [4], it is shown that a naive implementation for RDFS reasoning performs poorly. Thus, three non-trivial optimizations were introduced:

- the schema triples are loaded in the nodes' main memory and the rules are applied on the fly with the instance triples;
- the rules are applied during the *reduce* function, and the *map* function is used to group together the triples that could lead to a duplicated derivation;
- the rules are applied in a certain order so there is no need to apply the same rule multiple times.

The refined version of the algorithm proved to be have very high performance. The RDFS reasoner is able to compute the closure of the 1 billion triples of the 2008 Billion Triples challenge in less than 1 hour using 33 machines [4]. This approach currently outperforms any other published approach.

The performance of the OWL reasoner is not yet competitive. The optimizations introduced for the RDFS semantics do not apply for some of the rules of the OWL Horst fragment. Current research is focused on finding optimizations that apply to this fragment.

Concluding, Reasoning-Hadoop has shown that there are several advantages in using MapReduce for reasoning in Semantic Web. First, the reasoning can be done efficiently on large datasets because the Hadoop framework can be deployed in networks with thousand of nodes. Second, the execution is handled completely by the framework and the programmer can concentrate on the logic

⁴ <https://code.launchpad.net/jrhn/+junk/reasoning-hadoop>

of the program without worrying about the technical problems that are common in distributed systems. The main disadvantage of this approach lies in dealing with more complex logics. For example, the rules of the OWL Horst fragment are more complex and more difficult to encode efficiently. However, given the early stage of this research, a verdict is yet to be reached for the applicability of this approach to other logics.

5 Evaluation of Parallel Semantic Web Reasoning Approaches

Compared to traditional parallel reasoning, parallel Semantic Web reasoning approaches pose unique challenges in their evaluation. Thus, a new set of evaluation criteria needs to be defined. We can make a distinction between functional and non-functional characteristics of such systems.

Functional characteristics refer to the functions a system can perform. In the context of parallel Semantic Web reasoning, we can identify the following functional characteristics:

- *Logic applied*: depending on the logic implemented, various optimizations may be possible. For example, the antecedents of the RDFS match at most one instance triple [4]. The approach presented in the previous section exploits this fact to optimize RDFS reasoning by loading schema triples in memory and processing instance triples as a stream. Nevertheless, this optimization cannot be applied to OWL horst reasoning, because antecedents in the OWL horst ruleset may contain multiple instance triples.
- *Degree of completeness*: tolerating incomplete results can dramatically speed up the reasoning tasks [2]. This is a common compromise made in search engines. Furthermore, sometimes, the number of answers grows sublinearly with time. Depending on the task, it may be acceptable to return a fraction of the total answers spending a (smaller) fraction of the time that would be required to calculate all answers. MarVIN [5] is an example of a system with this characteristic.
- *Correctness*: similarly to completeness, accepting incorrect answers may also increase performance [2].

Non-functional characteristics refer to constraints in performing the prescribed functions. Some relevant non-functional characteristics for parallel reasoning are:

- *Triple throughput*: a central measure for the performance of a reasoning system is the number of triples it can process per second, indicating its efficiency [4,5].
- *Query throughput*: another performance measure for reasoning systems doing query answering is the number of queries they can process per second.
- *Query response time*: relevant to the previous characteristic, query latency refers to the amount of time between posting a query and getting the answer.

- *Scalability* refers to the capability of a system to handle larger input and to efficiently use additional computational resources. In parallel systems, computational resources are usually refer to computation nodes.
- *Maximum input size*: depending on the approach, there may be hard limits in the amount of data a system can handle, given some hardware. If these limits are reached, the approach becomes impractical. These limits may be imposed by restrictions in the available memory or hard disk space.

6 ReaSearch and Its Parallelization

Parallelizing reasoning is an attempt to solve the scalability problems for Web scale reasoning by introducing additional computational power. Nevertheless, because of the limitations and assumptions of traditional reasoning methods [2], complementary methods should be developed.

6.1 The Necessity of Parallelizing ReaSearch

ReaSearch⁵, which stands for unifying reasoning and search to Web scale, was proposed in [2] and is further developed in the LarKC project [3]. It emphasizes on an interweaving process of searching an important subset from the Web of data and do reasoning on it. The interweaving process will not stop until the user is satisfied with the reasoning result [2], as shown in Figure 1(a) (Note that this workflow design is a part of the search and reasoning part in LarKC [3]). On the Semantic Web, both the search and reasoning process need to handle massive data. Hence, the ReaSearch process need to be parallelized. ReaSearch is a framework for unifying reasoning and search, and concrete strategies on how to implement this framework can be developed through different methods.

USeR-G (Unifying Search and Reasoning from the viewpoint of Granularity)⁶ is a set of strategies that aim at combining the idea of granularity [23,24] and ReaSearch to provide concrete implementations of ReaSearch for the LarKC Project. The set of strategies include: unifying search and reasoning through multilevel completeness, multilevel specificity, multiperspective, etc [25]. In this paper, we will not go into details of these strategies, we just mention some processing steps in the implementation of these strategies where parallelization is needed. For the multilevel completeness strategy, the search process needs two parameters for the selection of important sub dataset. Namely, node degree calculation and node number statistics for the whole dataset. In one of our experiments for calculating the number of co-authors from the SwetoDBLP dataset [25], we spend more than one hour on the 1.08G semantic dataset, which is unacceptable, and should be parallelized. As indicated in [2], Web scale data may be over 10 billion triples. Hence, the task of calculating the node number for the multilevel completeness strategy can be parallelized to save time. For

⁵ <http://www.reasearch.net/>

⁶ <http://www.iwici.org/user-g>

the multilevel specificity strategy, nodes that are distributed in different levels of specificity can be assigned to multiple nodes in order to save processing time. For the multiperspective strategy, since the unification of search and reasoning can be done from multiple perspectives to meet the diverse user needs, different perspectives can be processed in parallel so that one can get reasoning results from all perspective almost at the same time. For all of the strategies, all of the reasoning parts can be and should be parallelized.

6.2 A Preliminary Design for the Parallel ReaSearch Architecture

As mentioned above, parallelizing ReaSearch is not only about parallelizing reasoning. Both search and reasoning need to be considered in the parallel environment. In this section, we propose a preliminary parallel ReaSearch architecture.

Following the design principles of parallel reasoning [26], for a parallel ReaSearch architecture, the nodes with different functionalities are distributed physically, but unified logically. As mentioned in Section 4.1, the unification of search and reasoning can be implemented in a workflow, hence the search process can be done in one node, and the reasoning process can be done in another. An alternative strategy, which scales better, is that the search and the reasoning processes themselves are also parallelized.

For the search part, search plugins are parallelized as shown in Figure 1(b),(c). In Figure 1(b), the search tasks are parallelized by dividing the dataset into several sub datasets and each subset is handled by one search plugin and the search

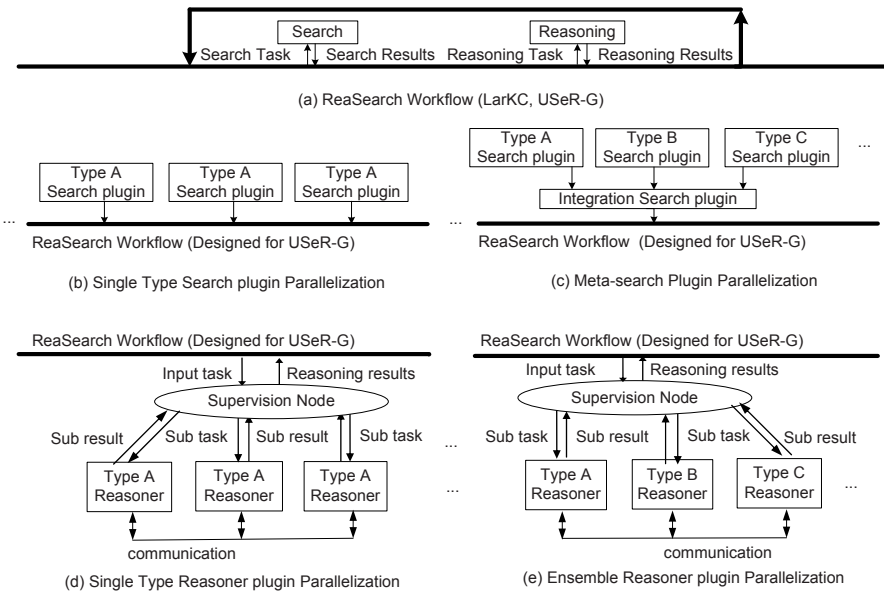


Fig. 1. A Parallel Architecture for ReaSearch

results from different search plugins are independently sent to the ReaSearch workflow for reasoning. Figure 1(c) provides an architecture for a meta-search plugin, which was inspired by the study of meta-search engine [27] for information retrieval. For this type of parallel search plugin, all the RDF/OWL datasets go through different types of the search plugins (Type A,B,C, etc.), and an integration search plugin is used to integrate all the search results from different type of search plugins and select out the most important subset of RDF/OWL data for reasoning. The selection criterion is that if a subset of the original dataset appears in all the search results from different search plugins, then it is considered as the most important subset and is delivered through the ReaSearch workflow for the reasoning task. If a subset appears in only some of the search results, then it is considered as a less important subset. Although this meta-search plugin has a parallel architecture, its aim is not to speed up the search process. Instead, it helps to select out the most important subset for reasoning.

For the reasoning part, two types of architectures are provided, as shown in Figure 1(d),(e). Each of them has a supervision node, hence both of them have centralized architectures. Reasoning is part of the ReaSearch workflow and is working in parallel with the parallel search architecture. In the architecture shown in Figure 1(d), reasoners are identical and apply the same rules. The dataset is divided in several parts to be processed on different reasoners(data partitioning). Its aim is to improve the speed of reasoning, which is similar in spirit with Reasoning-hadoop and MaRVIN. In Figure 1(e), the sub reasoning engines are different. This architecture refers to ensemble reasoning [28], which is inspired from ensemble learning in the field of machine learning. This architecture is not aimed at speeding up the reasoning process. The motivations behind this approach lie in: (1) Reasoning based on Web-scale data may produce too many results. Among these results, only some parts may be useful to a certain user. With the different reasoning plugins involved, various reasoning results will be obtained. If some results appear in the result sets from all reasoners, they can be considered as more important than others. The integration node can provide the most important reasoning results by selecting out the ones which appear in all or most sub result sets. (2) With different reasoning plugins involved, one may get more meaningful results compared to the first type shown in Figure 1(d). In this case, the integration node is responsible for merging all the reasoning results together for user investigation. In a more user-oriented scenario, the architecture also enables users to configure how many reasoners of each type they prefer to use for their own reasoning tasks. In this way, each type of reasoner will have a weight on the whole architecture. When these weights are changed, the reasoning results may also differ to meet different user needs.

7 Conclusion

The Semantic Web brings many new problems and insights to the field of parallel reasoning. New architectures and algorithms need to be developed to fit the context of Web scale reasoning. LarKC, MaRVIN, and Reasoning-Hadoop are

three practical systems which have touched this area and are potentially effective. Nevertheless, there is still much more that has not been well explored, such as how to develop concrete strategies for unifying reasoning and search in a parallel environment (i.e. parallelizing ReaSearch).

Moreover, through the preliminary design for the parallel ReaSearch architecture, we notice that for the reasoner part, a parallel architecture can improve the reasoning speed while also producing more fruitful reasoning results to meet diverse user needs. This topic needs further investigation. In this paper, we try to provide some preliminary discussion to inspire more research results in this area. We had some very preliminary discussions on where do ReaSearch (more specifically, the set of methods UseR-G) should be parallelized. In future work, we will go into deeper discussion and concrete implementations.

Acknowledgements

This study was supported by the European Union 7th framework project FP7-215535 LarKC (Large Knowledge Collider) and VU University Amsterdam. This paper was prepared when Yi Zeng was visiting VU University Amsterdam.

References

1. Berners-Lee, T.: The semantic web. *Scientific American* 6, 1–6 (2001)
2. Fensel, D., van Harmelen, F.: Unifying reasoning and search to web scale. *IEEE Internet Computing* 11(2), 96, 94–95 (2007)
3. Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Valle, E., Fischer, F., Huang, Z., Kiryakov, A., Lee, T., School, L., Tresp, V., Wesner, S., Witbrock, M., Zhong, N.: Towards lark: A platform for web-scale reasoning. In: *Proceedings of the International Conference on Semantic Computing*, pp. 524–529 (2008)
4. Urbani, J., Kotoulas, S., Oren, E., van Harmelen, F.: Scalable distributed reasoning using mapreduce. In: *Proceedings of the International Semantic Web Conference* (2009)
5. Oren, E., Kotoulas, S., Anadiotis, G., Siebes, R., Ten Teije, A., van Harmelen, F.: Marvin: distributed reasoning over large-scale semantic web data. *Journal of Web Semantics* (to appear)
6. Flynn, M.: Very high-speed computing systems. *Proceedings of the IEEE* 54(12), 1901–1909 (1966)
7. Gallizo, G., Roller, S., Tenschert, A., Witbrock, M., Bishop, B., Keller, U., van Harmelen, F., Tagni, G., Oren, E.: Summary of parallelisation and control approaches and their exemplary application for selected algorithms or applications. In: *LarKC Project Deliverable 5.1*, pp. 1–30 (2008)
8. Wilkinson, B., Allen, M.: *Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers*, 2nd edn. Prentice-Hall, Englewood Cliffs (2005)
9. Robert, B.: Japan's pipedream: The fifth generation project. *System and Software* September 3, 91–92 (1984)

10. Ehud, S.: Systolic programming: A paradigm of parallel processing. In: Proceedings of the international conference on Fifth Generation Computer Systems, pp. 458–470 (1984)
11. Liu, Z., You, J.: Dynamic load-balancing on a parallel inference system. In: Proceedings of the Second IEEE Symposium on Parallel and Distributed Processing, pp. 58–61 (1990)
12. Tan, X., Zhang, X., Gao, Q.: Load sharing algorithms for parallel inference machine epim-ldshbs, intlsh. Chinese journal of computers (5), 321–331 (1986)
13. Allemang, D., Hendler, J.: Semantic Web for the Working Ontologist. Elsevier, Inc., Amsterdam (2008)
14. Brachman, R., Levesque, H.: Knowledge Representation and Reasoning. Elsevier, Inc., Amsterdam (2004)
15. Soma, S., Prasanna, V.: Parallel inferencing for owl knowledge bases. In: Proceedings of the 37th International Conference on Parallel Processing, pp. 75–82 (2008)
16. Schlicht, A., Stuckenschmidt, H.: Distributed resolution for alc. In: Proceedings of the International Workshop on Description Logic (2008)
17. Oren, E.: Goal: Making pipeline scale. Technical report, LarKC 1st Early Adopters Workshop (June 2009)
18. van Nieuwpoort, R., Maassen, J., Wrzesinska, G., Hofman, R., Jacobs, C., Kielmann, T., Bal, H.: Ibis: a flexible and efficient java based grid programming environment. Concurrency and Computation: Practice and Experience 17(7-8), 1079–1107 (2005)
19. Chapman, B., Jost, G., van der Pas, R., Kuck, D.: Using OpenMP: Portable Shared Memory Parallel Programming. The MIT Press, Cambridge (2007)
20. Bornemann, M., van Nieuwpoort, R., Kielmann, T.: Mpi/ibis: a flexible and efficient message passing platform for java. In: Proceedings of 12th European PVM/MPI Users' Group Meeting, pp. 217–224 (2005)
21. Hayes, P.: Rdf semantics. In: W3C Recommendation (2004)
22. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation, pp. 137–150 (2004)
23. Hobbs, J.: Granularity. In: Proceedings of the 9th International Joint Conference on Artificial Intelligence, pp. 432–435 (1985)
24. Yao, Y.: A unified framework of granular computing. In: Handbook of Granular Computing, pp. 401–410. Wiley, Chichester (2008)
25. Zeng, Y., Wang, Y., Huang, Z., Zhong, N.: Unifying web-scale search and reasoning from the viewpoint of granularity. In: Liu, J., et al. (eds.) AMT 2009. LNCS, vol. 5820, pp. 418–429. Springer, Heidelberg (2009)
26. Serafini, L., Tamilin, A.: Drago: Distributed reasoning architecture for the semantic web. In: Proceedings of the European Semantic Web Conference, pp. 361–376 (2005)
27. Howe, A., Dreilinger, D.: Savvysearch: a meta-search engine that learns which search engines to query. AI Magazine 18(2), 19–25 (1997)
28. Chabuk, T., Seifter, M., Salasin, J., Reggia, J.: Integrating knowledge-based and case-based reasoning. Technical report, University of Maryland (2006)

A Model for Personalized Web-Scale Case Base Maintenance

Jingyu Sun^{1,2}, Xueli Yu¹, Ruizhi Wang³, and Ning Zhong^{2,4}

¹ College of Computer and Software, Taiyuan University of Technology
Taiyuan, Shanxi, 030024, China

² International WIC Institute, Beijing University of Technology
Beijing, 100022, China

³ Dept. of Computer Science and Technology, Tongji University
Shanghai, 200092, China

⁴ Dept. of Life Science and Informatics, Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi 371-0816, Japan
whitesunpersun@163.com, xueli13287@263.net,
ruizhi.ann.wang@gmail.com, zhong@maebashi-it.ac.jp

Abstract. The growing use of case-based reasoning (CBR) systems on the Web has brought with it increased awareness of the Web-scale case base maintenance (CBM). While most existing CBM policies and approaches, which were designed for smaller case bases with sizes ranges from thousands to millions of cases, are not suitable for Web-scale distributed case collection. In this paper, we propose a novel CBM model for personalized Web-scale CBM, which addresses the needs of the Web-based CBR systems. The proposed CBM model is constructed based on chunk activation of ACT-R theory and rule-based reasoning. In addition, a basic activation value computation method is given to rank the cases and an algorithm is proposed to select top-N active cases. Experiments on several real-world datasets such as the MovieLens dataset show the effectiveness of our model.

1 Introduction

With the development of Internet and Web technology, more and more new real world requirements emerge, e.g., to enhance semantic supports for the current Web, to provide personalized service, to reinforce abilities of processing Web-scale problem and so on. In the academic community, some trial solutions have been proposed, such as Web Intelligence (WI) [1,2], Semantic Web [3], Semantic Search [4]. In the industrial community, Web 2.0, cloud computing and others have been proposed and show a great development. However, it is too difficult to solve above problems completely by several approaches and technologies in a short period of time. Challenges still exist and new approaches need to be explored according to new real world applications.

Generally speaking, Web reasoning is one of the good solutions to enhance semantic supports for the current Web. Moreover, it is easy to provide personalized

service through reasoning on individual data (user profile). However, traditional logic based reasoning systems do not scale to the large amount of information that is required for the Web [5]. In other words, current reasoning methods cannot meet Web-scale requirement and are invalid in the large datasets (knowledge bases). Recently, Web-scale reasoning is considered as a key problem and many research interests have focused on on it. For example, *Semantic Web challenge* [6] is a related project and the specific goal of its Billion Triples Track is to demonstrate the scalability of applications as well as to encourage the development of applications that can deal with realistic Web data. Another related project is the *EU FP 7 Large-Scale Integrating Project LarKC* [7], which is to develop the *Large Knowledge Collider (LarKC)*, a platform for massive distributed incomplete reasoning that will remove the scalability barriers of currently existing reasoning systems for the Semantic Web [7].

Furthermore, Web-scale data are distributed on different Web sites and Web-scale reasoning methods should support to collect them from multi-data sources and reason on them. So some related technologies, such as Resource Description Framework (RDF), SparQL, and some reasoning frameworks, such as LarKC [7], have been proposed. Especially, there are massive experiential knowledge distributed on different Web sites, such as purchase experience, search histories, user's comments, tags of a Web page. However, for some cases, it is difficult to build a rule-based system, but relatively easy to extract cases from multiple Web data sources. Recently Zhong et al. proposed a framework of granular reasoning [8] to implement Web-scale reasoning by combining case-based reasoning (CBR) with other reasoning methods. Additionally, they pointed out that personalized services need to be considered on the same importance with Web-scale problem solving and reasoning.

We argue that CBR is one of more feasible Web reasoning methods and can be combined with rule-based reasoning (RBR), which makes up some shortcomings of RBR for Web-scale reasoning. There are at least two grounds supported it:

- The traditional rule-based systems have a few drawbacks, such as difficulties of knowledge acquisition, no memory of tackled problems or previous experience, poor efficiency of inference, ineffectiveness to deal with exceptions, and poor performance of the whole system. Whereas, CBR can just handle the above problems well [9].
- CBR has roots in cognitive psychology [10] and can be regarded as a human problem solving inspired method, which is easy to deal with massive data at a high speed.

Based on above observations, we firstly propose a realistic solution for Web-scale reasoning based on granular reasoning proposed by Zhong et al. [8]: Web-scale case-based reasoning for personalized services as shown in Fig. 1. In this solution, there are three key problems: personalized Web-Scale case representation, personalized Web-Scale case base maintenance, and personalized Web-Scale case utilization (retrieval). Additionally, there are two types of data which need to be processed and represented by RDF: cases and user profiles. A case may be

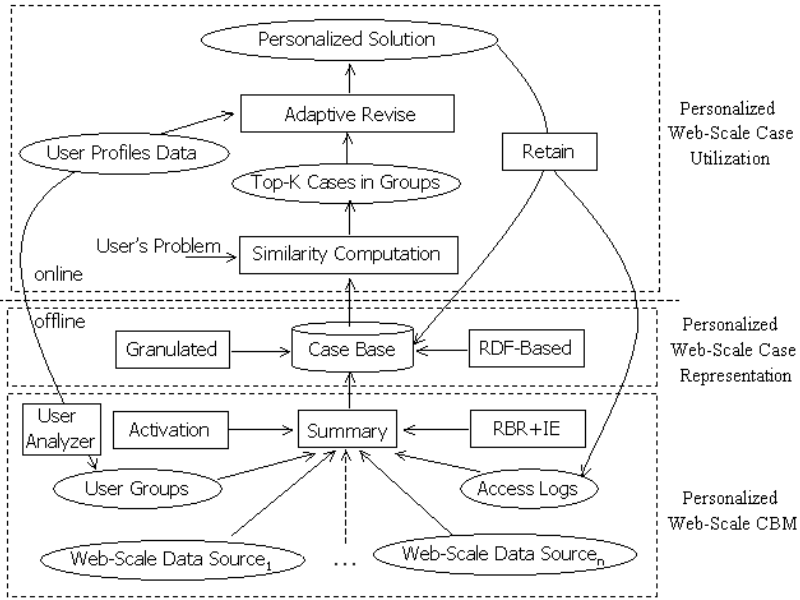


Fig. 1. Web-scale CBR model for personalized services

built or adapted by extracting items of multi-data sources. While user profiles include personalized needs and help with reasoning.

However, it is very crucial to maintain and evaluate a bigger case base in the Web-scale CBR system and most existing CBM policies and approaches are designed for smaller case bases with sizes ranges from thousands to millions of cases and not suitable for Web-scale distributed case base. In this paper, we focus on investigating personalized Web-Scale case base maintenance. Section 2 introduces related work about CBM. In Section 3, we firstly propose a novel CBM policy according to characteristics of Web case – a valuable case utilization policy and a model for personalized Web-scale CBM, which is based on chunk activation of ACT-R theory [11] and rule-based reasoning. Then we discuss a unified framework and propose a basic activation value computation model and a top-N active cases selection algorithm for analyzing valuable case based on the proposed CBM model. In Section 4, we show some experimental results on real applications to verify the effectiveness of the proposed approach. Finally, Section 5 concludes the paper by highlighting the major advantages of our method and pointing out our future work.

2 Related Work

Case-based reasoning (CBR) is a relatively older problem solving technique that is attracting increasing attention in the Web era. It has been used successfully

in diverse application areas [12], ranging from web search [13], recommender systems [14] to legal reasoning [15]. The work Schank and Abelson in 1977 is widely held to be the origins of CBR [16]. It is inspired by human problem solving and has roots in cognitive psychology [10]. Its key idea is to tackle new problems by referring to similar problems that have already been solved in the past [17]. A widely accepted framework for case-based reasoning is characterized by the so-called “CBR cycle” [18] and consists of the following four phases [19]:

- Retrieve the case(s) from the case base which is (are) most similar to the target problem.
- Reuse the information provided by this (these) case(s) in order to generate a candidate solution for the new problem.
- Revise the proposed solution according to the special requirements of the new problem.
- Retain the new experience obtained in the current problem-solving episode for future problem solving.

In recent years, researchers are paying more attention to Case Base Maintenance (CBM). In the revised model [20], it emphasizes the important role of maintenance in modern CBR and indeed proposes that the concept of maintenance encompass the retain, review and restore steps. There is an accepted definition of case base maintenance:

CBM implements policies for revising the organization or contents (representation, domain content, accounting information, or implementation) of the case base in order to facilitate future reasoning [21].

In addition, CBM involves revising of indexing information, links between cases, and/or other organizational structures and their implementations.

Simply, the current CBM approaches can be divided in two policies, one concerning optimization and the other a case base partitioning. The objective of these approaches is to reduce the case retrieval time. Usually, the optimization policy consists of deleting less by following two strategies: addition and the deletion of cases. Whereas, the partitioning policy consists of dividing the case base into several search spaces. Additionally, there are two the important criteria for evaluating case base:

- *Competence* is the range of target problems that can be successfully solved.
- *Performance* is the answer time that is necessary to compute a solution for case targets. This measure is bound directly to adaptation and result costs.

However, in our field of search, there is no literature about the Web-scale CBR and the Web-scale CBM. But they are very important problems in the Web era. In this paper, a personalized Web-scale CBM model is firstly discussed.

3 Personalized Web-Scale CBM Model

In recent years, researchers are paying more attention to Web-based CBR. Many Web-based CBR systems [13][22] are emerging. However, case base is regarded

as one of key components of CBR system. But existing CBM policies and approaches are not fit to maintain Web-scale case bases and there are at least three reasons as following:

- The case-deletion strategy is too limited to detect which Web-scale cases should be deleted owing to they are distributed, dynamic and personalized.
- The case-addition strategy is poor efficiency for a bigger case base and related algorithms are too complex to be fit for a Web-scale case base.
- Additionally, the partitioning policy is only a secondary strategy for a Web-scale case base because it is necessary to partition a bigger case base effectively.

We argue that the Web-scale case base maintenance needs a new policy, which is attune to characteristics of a Web-scale case base and can facilitate the Web-scale CBR. However, what is a right policy and method for Web-scale CBM? In our opinions, characteristics of Web cases are important factors and decide CBM policy and model. Firstly, we discuss them as follows.

3.1 Characteristics of Web Cases

A Web-scale case base consists of massive Web cases and Web cases usually shows many new characteristics in a Web-based CBR system as follows.

- Distributed: A Web case may be organized by some different kinds of items which may be extracted from different data sources and different Web cases may be distributed in many Web sites.
- Dynamic: Some item-values of a Web case may be changed and some sub-case bases distributed in different Web sites may be not visited. For example, plane ticket (an item of case) often has a different discount (item-value) in the different season; and a bigger case base may be dynamically divided into many sub-case bases.
- Personalized: Different Web cases are fit for different users and may include a little different individual information owing to they are accessed by many different users.
- Massive quantity: There may be massive Web cases in a Web-based CBR system.
- Hierarchical: Web cases may be hierarchically restructured according to items and their types in order to facilitate future reasoning.

Due to above characteristics, we think that it is very important to find most valuable cases from massive cases combining personalized needs in order to keep validity of Web case base and update some older or invalid items of cases in order to facilitate future reasoning. However, how to attain this goal? We will discuss a new CBM policy in the following section.

3.2 A Valuable Case Utilization Policy for CBM

In this paper, we propose a valuable case utilization policy for Web case bases. It shows two features as follows:

- Choose most valuable items which are extracted from different data sources in order to build cases combining with personalized needs.
- Choose most valuable cases from massive cases in order to compress size of Web case bases.

However, this policy is different from the traditional optimization and partitioning policy. It focuses on utilizing valuable cases which are built by most valuable items extracted from different data sources; however, optimization policies focus on compressing case base through deletion and addition strategy, while partitioning policy is only secondary policy in order to reduce search space.

3.3 A Personalized Web-Scale CBM Model

In order to implement above new policy, we propose a personalized Web-scale CBM model to pick most valuable cases in a Web-based CBR system through computing activation values of items and cases based on chunk activation of ACT-R theory [11]. And it also utilizes individual information (user profiles) and users' access logs based on rule-based reasoning. In detail, it includes two levels:

- The first level is based on chunk activation of ACT-R theory to choose most valuable data (item and case). It includes two sub-levels:
 - The first sub-level is an item activation value computation level. In this sub-level, activation values of items are computed in order to build cases through utilizing valuable items, which are extracted from multi data sources, according to related activation values.
 - The second sub-level is a case activation value computation level. In this sub-level, activation values of cases are computed in order to compress size of case base through retaining most valuable cases.
- The second level is based on rule-based reasoning and also includes two sub-levels:
 - The first sub-level is to choose items to build cases or compress size of case base according to rules which are designed based on activation values for each item or case.
 - The second sub-level is to revise a case or edit a case base according to reasonable rules which are produced based on metrics for cases and a case base.

Additionally, traditional partitioning policy is taken as a secondary approach to strengthen this model and is used to partition case base based on case similarity for the case storage and retrieval.

However, what is a feasible computation model to implement this model? We argue that it is a better way to simulate human problem solving methods and pick most valuable items and cases through utilizing their access time, frequency and so on. In addition, a complementary way is to utilize relationships between a case and items which are used to build this case.

3.4 An Activation Value Computation Model for Items and Cases

Specially, ACT-R is a cognitive architecture, a theory about how human cognition works. And the chunk activation computation is one core of ACT-R theory and it can reflect which chunk (a form of knowledge) should be picked according to its previous access time and frequency and relationships with other chunks. Based on it and combined with rule-based reasoning, we design a unified framework for the computation model, which is fit to pick valuable items and cases as shown in Fig. 2.

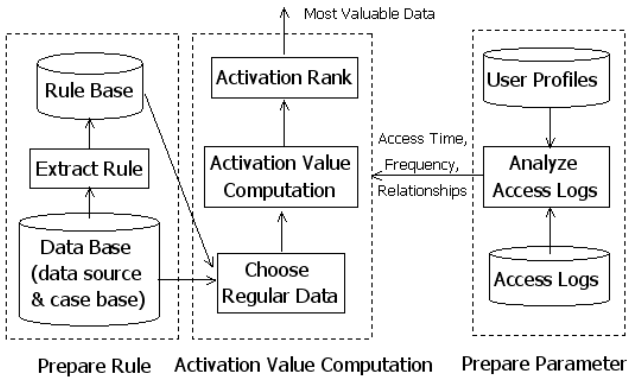


Fig. 2. A unified framework for the computation model

In this unified framework for the computation model, there are three key phases:

- The first phase is to prepare rules in order to wash data and output regular data according to a real-world application.
- The second phase is to prepare right parameters for activation value computation model through analyzing access logs, user profiles and so on.
- The third phase is to compute activation values for data (items or cases) with right parameters and to pick most valuable data according to the real-world application or characteristics of data.

The core of this framework is the activation value computation, which utilizes chunk activation computation of ACT-R theory, in order to pick valuable data. And there are two fundamental assumptions:

- The first one is that an item is taken as a simple chunk. Usually, it is extracted from one data source and is a simple chunk which only include one slot.
- The second one is that a case is taken as a regular chunk. Usually, it is organized by some items and is a regular chunk which includes several slots.

In ACT-R theory, the activation of a chunk is a sum of base-level activation, reflecting its general usefulness in the past, and an associative activation, reflecting its relevance in the current context. It is given by

$$A_i = B_i + \sum_j (W_j \cdot S_{ji}), B_i = \ln\left(\sum_l^h t_{il}^{-d}\right), \tag{1}$$

where A_i denotes the activation value of the chunk i , B_i denotes the base-level activation value and if higher if used recently, W_j denotes the attentional weighting of element j (or slot) of the chunk i , S_{ji} denotes the strength of association of element j to the chunk i , t_{il} denotes access time of the chunk i at the l th time, h denotes access times to the chunk i in total, and d is a weaken parameter (usually equals 0.5).

In order to discuss the activation value computation model for convenience, we firstly give some definitions as follows.

Definition 1 (Data base)

A data base is denoted by $D = \{d_1, d_2, \dots, d_n\}$. For anyone $d_i \in D$, d_i is an item which is extracted from one data source or a case which is organized by several different types of items. That is, D is a set for items or cases.

Definition 2 (Access log base)

An access log base is denoted by $L = \{l_1, l_2, \dots, l_m\}$. For anyone $l_j \in L$, l_j is produced by the system and includes *user's id*, *id of data (d_i)*, *remarks*, *access time* and others. Usually, there are k_i access logs for the same d_i and they are denoted by $L_i = \{l_{k_1}, l_{k_2}, \dots, l_{k_i}\}$.

Definition 3 (User group Set)

A user group set is made up of all user groups and is a partition according to d_i and L_i . Usually, it is denoted by $G = \{g_1, g_2, \dots, g_h\}$, and for any one user group $g_f \in G$, g_f includes some users who access same d_i or have related to l_j .

Then can discuss our computation model based on above definitions and Equation 1. Due to B_i can reflect the usefulness of a chunk in the past through its previous access time and frequency of utilization, we can take the base-level activation value as a value of an item or a case to denote how it is valuable. But A_i can reflect the relevance of a chunk in the current context through relationships with other chunks and usually we can use the associative activation value to denote the relationship between an item or a case and a user in current situation. So A_i can be taken as a personalized parameter of an item or a case and used to support personalized services. In the first step, we only consider the value of an item or a case and a basic computation model is given as follows:

$$A_i^{(f)} \approx B_i^{(f)} = \ln\left(\sum_l^h (t_{il}^{(f)})^{-d}\right), \tag{2}$$

where, $t_{il}^{(f)}$ denotes the l th access time by users in group g_f for data $d_i \in D$ until now, usually d_i is accessed by h times in total, d is a weaken parameter.

In order to meet requirements of a real-world application, a time factor α is introduced into Equation 2 for right time span as follows:

$$A_i^{(f)} \approx B_i^{(f)} = \ln\left(\sum_l^h (\alpha \cdot t_{il}^{(f)})^{-d}\right), \quad (3)$$

where, α is decided by time span extracted from access logs in the real-world application and can be estimated by $\max(t_{il}^{(f)})$ and $\min(t_{il}^{(f)})$.

As a result, Equation 3 is taken as a basic activation value computation model and fit to estimate the value of an item and a case in the proposed personalized Web-scale CBM model. However, we can utilize the associative activation value of an item or a case in order to enhance personalization and will discuss it in detail in our future work.

Furthermore, we propose an algorithm to pick valuable cases for the proposed CBM model, named *Selecting Top-N Active Cases (STAC)* algorithm as shown in Algorithm 1. This algorithm is only a framework and gives some key steps for picking top-K valuable cases. In practice, we should choose effective metrics for partitioning user group and right analysis methods for a access logs base according to real-world application. Also we should adjust right parameters of Equation 3 for different access logs bases and design some feasible rules for building case.

Algorithm 1. Selecting Top-N Active Cases.

Input: Data bases D_1, D_2, \dots, D_q , access logs bases L_1, L_2, \dots, L_q , user profiles, K, N .

Output: Some sets of top-K active cases for each group and a set of top-N active cases.

Steps:

1. S1: Build user group G based on user profiles and access logs bases according to
 2. metrics of the real-world application.
 3. S2: Pick top-K active items of every data base $D_v (v = 1, 2, \dots, q)$ to each group
 4. $g_f (f = 1, 2, \dots, h)$:
 5. for anyone g_f, D_v and L_v do
 6. (1). Analyze L_v in order to determine values for $t_{il}^{(f)} (i = 1, 2, \dots, n^{(v)};$
 7. $l = 1, 2, \dots, h^{(f,v,i)})$;
 8. (2). Compute activation value $A_i^{(f,v)}$ of $d_i^{(f,v)} \in D_v$ according to Equation 3;
 9. (3). Sort item by its $A_i^{(f,v)}$ and pick top-K item set $I^{(f,v)}$.
 10. end for.
 11. S3: Build cases for each group g_f based on all top-K item set $I^{(f,v)} (v = 1, 2, \dots, q)$
 12. and rules which are designed according to the real-world application:
 13. for anyone g_f do
 14. Build top-K case set $CK^{(f)}$ through picking a group items $I_r^{(f,1)}, I_r^{(f,2)}, \dots,$
 15. $I_r^{(f,q)}$ according to rules and Equation 3;
 16. end for.
 17. S4: Make an intersection for all sets of top-K active cases $CK^{(f)} (f = 1, 2, \dots, h)$
 18. in order to pick top-N active case set C of the case base.
 19. S5: Return all top-K case sets $CK^{(f)} (f = 1, 2, \dots, h)$ and a case set C .
-

In one word, we consider personalized needs for the proposed CBM model and STAC algorithm through partitioning user groups. And they are fit to process massive data set in parallel and support distributed analysis and processing in the offline due to users are divided into many different groups. All these merits are favorable for Web-scale problems and supplying personalized services. However, an effective evaluation is necessary to be designed according to the real-world application and will be discussed in our future work. In addition, there are some advantages for the proposed CBM model. The first one is that it can simulate human problem solving methods, for example, it focuses on recent events (items or cases accessed recently) to predict which cases are more valuable. The second one is that it utilizes RBR to update cases and so on. The third one is that it can supply personalized services through partitioning user groups.

4 Experiments

4.1 First Experiment

In this experiment, we take one of MovieLens dataset (it can be attained in the GroupLens Wet site: <http://www.grouplens.org/node/73>) as data base. It is the biggest one: “10 million ratings and 100,000 tags for 10681 movies by 71567 users” and used to compute the activation value for every item (movie) and to discover value of item (movie) based on its access time and frequency of utilization in the pastime as following steps.

- Firstly, four fifths of the dataset are randomly chosen as training set and remaining data are as testing set.
- Secondly, the activation value for every item (movie) in testing set is computed according to our model and Equation 3 and all items in testing set are sorted ascending by their activation value.
- Thirdly, the classic item-based collaborative filtering algorithm is run in training set to train similarity matrix in order to predict every item’s ratings in testing set. And we take mean absolute error (MAE) as a metrics for evaluation. Usually, an item (a movie) with smaller MAE can show that its related ratings, access time and frequency of utilization are more valuable and should lead to bigger activation value.
- Finally, we draw MAE change curves according to items’ rank by activation value and add linear trend and logarithmic trend for MAE to illustrate MAE change trend.

In this experiment, six time tests are conducted on different testing sets which are chosen at random and six MAE change curves are drawn. From every curve, a same fact can be discovered: an item (movie) with bigger activation value shows smaller MAE. We can draw a conclusion: an item (case) with bigger activation value should be utilized with high priority and an item (case) with smaller activation value can be deleted from data base if it is lower than a given threshold. This conclusion illustrates that our proposed utilizing valuable case policy is feasible and farther experiments will be made in our future work.

4.2 Second Experiment

We design a real world application based on our proposed CBM model – the home finder (HF) system. It is a Web-based CBR system and supply a service to search for real estates (apartment, house, hotel, etc.) for sale or to be rent. There are multi data sources (real estate or hotel information, map, public transportation, statistics) and massive users in this system. All data are represented by RDF and users' access histories are taken as cases. All cases are organized in hierarchic structure in order to find wanted cases rapidly through referring to hierarchic case-based reasoning [23].

In this system, the utilizing valuable case policy is used and the case base is maintained and updated at regular intervals according to our proposed CBM model. For each user group, top-K cases are picked according to the system's access logs. Personalized solutions are recommended according to users' queries and user profiles in the online. And the system also retains users' queries and choices as a new case.

In this experiment, some datasets collected from this demo system and preliminary experimental results show that our proposed CBM policy and model are feasible to maintain Web-scale case base. And More results will be produced in our future work.

5 Conclusions

In this paper, we proposed a new CBM policy – a valuable case utilization policy and a model for personalized Web-Scale CBM, which can simulate human problem solving methods through utilizing chunk activation of ACT-R theory and RBR. In addition, a basic activation value computation model and a top-N active cases selection algorithm for the proposed model were discussed in detail. Experiments on several real-world datasets show that the proposed CBM policy and model are suitable to maintain Web-scale case base and can be used in Web-based CBR system. However, it is very important to choose different rules and parameters for the proposed model according to different real-world applications. In our future work, we will analyze and extend our computation model and algorithm in theory in order to optimize them and develop better methods to find right rules and parameters for real-world applications.

Acknowledgments

The authors would like to acknowledge the following support for their research on hierarchic personalized Web-scale case base maintenance: Natural Science Foundation of China (No. 60873139); International collaboration Project of Shanxi (No. 2008081032); Youth Natural Science Foundation of Shanxi (No. 200821024).

References

1. Zhong, N., Liu, J.M., Yao, Y.Y.: In Search of the Wisdom Web. *IEEE Computer* 32(11), 27–31 (2002)
2. Zhong, N., Liu, J.M., Yao, Y.Y.: *Web Intelligence*. Springer, Heidelberg (2003)

3. Semantic Web (2009), <http://www.w3.org/2001/sw/>
4. Guha, R., Mccool, R., Miller, E.: Semantic Search (2009), <http://www2003.org/cdrom/papers/refereed/p779/ess.html>
5. Fensel, D., et al.: Towards LarKC: A Platform for Web-scale Reasoning. In: ICSC 2008, pp. 524–529 (2008)
6. Semantic Web Challenge (2009), <http://challenge.semanticweb.org/>
7. LarKC (2009), <http://www.larkc.eu/>
8. Zhong, N., Yao, Y.Y., et al.: Towards Granular Reasoning on the Web. In: NEFORs 2008, ASWC 2008 (2008)
9. Babaka, O., Whar, S.Y.: Case-based Reasoning and Decision Support Systems. In: Proc. of the IEEE Internal Conference on Intelligent Processing Systems, pp. 1532–1536 (1997)
10. Riesbeck, C.K., Schank, R.C.: Inside Case-based Reasoning. Hillsdale, New Jersey (1989)
11. Anderson, J.R., Bothell, D., et al.: An Integrated Theory of Mind. *Psychological Review* 111, 1036–1060 (2004)
12. Bergmann, R., Althoff, K.D., Breen, S., et al.: In: Bergmann, R., Althoff, K.-D., Breen, S., Göker, M.H., Manago, M., Traphöner, R., Wess, S. (eds.) *Developing Industrial Case-Based Reasoning Applications*, 2nd edn. LNCS (LNAI), vol. 1612, Springer, Heidelberg (2003)
13. Balfe, E., Smyth, B.: Case-based collaborative web search. In: Funk, P., González Calero, P.A. (eds.) *ECCBR 2004*. LNCS (LNAI), vol. 3155, pp. 489–503. Springer, Heidelberg (2004)
14. Lorenzi, F., Ricci, F.: Case-Based Recommender Systems: A Unifying View. *Intelligent Techniques for Web Personalization*, 89–113 (2005)
15. Bruninghaus, S., Ashley, K.D.: Combining Case-based and Model-based Reasoning for Predicting the Outcome of Legal Cases. In: Ashley, K.D., Bridge, D.G. (eds.) *ICCBR 2003*. LNCS, vol. 2689. Springer, Heidelberg (2003)
16. Watson, I., Marir, F.: Case-Based Reasoning: A Review. *The Knowledge Engineering Review* 9(4) (1994)
17. Kolodner, J.L.: *Case-based Reasoning*. Morgan Kaufmann, San Francisco (1993)
18. Hullermeier, E.: Credible Case-Based Inference Using Similarity Profiles. *IEEE Transactions on Knowledge and Data Engineering* 19(6), 847–858 (2007)
19. Aamodt, A., Plaza, E.: Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications* 7(1), 39–59 (1994)
20. Berghofer, T.R., Iglezakis, I.: Six Steps in Case-Based Reasoning: Towards a Maintenance Methodology for Case-based Reasoning Systems. In: *GWCBR 2001*, pp. 198–208. Shaker-Verlag, Aachen (2001)
21. Leake, D.B., Wilson, D.C.: Categorizing case-base maintenance: Dimensions and directions. In: Smyth, B., Cunningham, P. (eds.) *EWCBR 1998*. LNCS (LNAI), vol. 1488, pp. 196–207. Springer, Heidelberg (1998)
22. Toussaint, J., Cheng, K.: Web-based CBR (Case-based Reasoning) as a Tool with the Application to Tooling Selection. *The International Journal of Advanced Manufacturing Technology* 29, 24–34 (2005)
23. Smyth, B., Keane, M.T., Cunningham, P.: Hierarchical Case-Based Reasoning Integrating Case-Based and Decompositional Problem-Solving Techniques for Plant-Control Software Design. *IEEE Transactions on Knowledge and Data Engineering* 13(5), 793–812 (2001)

X3D-Based Web 3D Resources Integration and Reediting

Zhoufan Zhou¹, Hisao Utsumi², and Yuzuru Tanaka²

¹ School of Mechanical Engineering, University of Science and Technology Beijing,
No. 30, Xuanyuan Road, Haidian District, Beijing, 100083 China

zhoufan.zhou@gmail.com

² Meme Media Laboratory, Hokkaido University,
North 13, West 8, Kita-ku, Sapporo, 060-8628 Japan
{utsumi, tanaka}@meme.hokudai.ac.jp

Abstract. X3D is an open standard for 3D shape content delivery, which combines both 3D shapes and animation behaviors into a single file. IntelligentBox is a prototyping software development system which represents any 3D object as a Box, i.e., a functional 3D visual object. X3D can provide more interactivity if it is empowered by IntelligentBox to add more reedit-ability. IntelligentBox provides an interactive environment for X3D shapes. This paper shows how X3D file shape information can be reused in IntelligentBox, and how IntelligentBox enables users to interactively add interactive animation functions to the original X3D models.

1 Introduction

In this age of information explosion, the convenience and speediness of Internet make it possible for us to get various information and resources timely. It is no doubt that graphic information visualization plays an important role in the information world. Along with the developing computer graphic technologies, Web is not a 2 Dimension (2D) world anymore. More and more Web 3D formats are being created and studied. Among them, a format called X3D, which is a kind of lightweight format of 3D shape model, is mounting the stage of the Web 3D world. However, there is still no such a platform which can integrate and reedit these 3D applications effectively, and the animation functions of X3D are not interactive enough, which has already been the bottleneck of X3D development. Therefore, the sharing and reusing of dynamic interactive 3D contents has been a topic of our research. If it becomes true, X3D will provide more powerful interactive animation functions which can be used in the fields of Electronic Commerce, Virtual Reality and so on.

X3D, namely Extensible 3D, is an open standard for 3D shape content delivery [1], which combines both 3D shapes and animation behaviors into a single file. X3D is generated by Virtual Reality Modeling Language (VRML97). Nowadays, many groups and teams are doing the research on X3D [2, 3, 4].

Compared to other kinds of 3D formats, X3D has many advantages as 1) International Organization for Standardization (ISO) open free format, 2) portability, 3) efficient programming, 4) easy code modification, 5) possibility of being translated from other 3D applications (AutoCAD, 3D Studio MAX, Google SketchUp, etc.), 6) abundant free Web

resources. Based on those advantages, nowadays X3D is widely used in the fields of engineering and scientific visualization, CAD and architecture, medical visualization, training and simulation, multimedia, entertainment, education, and so on [5, 6, 7, 8].

People can find free X3D resources easily in related web sites, such as Web3D Consortium who provides more than 2600 free samples by the web page <http://www.web3d.org/x3d/content/examples/X3dResources.html#Examples>, or translate other 3D format files into X3D using appropriate converter programs.

However, X3D has some disadvantages as well, such as, 1) the modification of X3D is not convenient enough for common users; 2) it cannot create the real sense of interactive functions, users are namely, allowed to use only those functions provided by designers.

Meanwhile, there is no interactive visual environment to easily integrate or to reedit those free resources, even though many researchers and companies are dedicating themselves to related works.

Aiming at solving these problems, we have used IntelligentBox [9, 10, 11], which is a constructive visual software development system proposed in 1995 for interactive 3D graphic applications, and developed an environment which not only focuses on the resource integration, but also enables users to easily modify and reedit X3D files.

For the purpose of easy editing and composition of X3D models, many companies are providing X3D editors today. Web3D Consortium, who started X3D, is also encouraging researchers and enterprises to devote their time to the related application development. We have tried four kinds of X3D editors which are popular in X3D designers, they are X3D-Edit [12], Vivaty Studio [13], BS Editor [14] and SwirlX3D Editor [15].

X3D-Edit is the primary authoring tool used to create X3D files. It almost covers all the functions of X3D. It provides a pure coding environment and a preview browser, which does not enable users to directly manipulate visual objects to edit their shapes and functions, so it is suitable for professional users. Vivaty Studio, BS Editor and SwirlX3D Editor all provide the visual window by which users can edit the model easily, nevertheless, there still exists some shortcomings of them respectively. Vivaty Studio is a visually oriented application from Vivaty. It provides some basic functions by a menu. Three-view interface ensures the objects can be observed from three different viewpoints, at the same time, users can move them just by dragging mouse. BS Editor is another X3D editor which can display the model in a view window. Users can observe the objects as in a normal browser. However, even though Vivaty Studio and BS Editor have already removed the necessity of pure coding, the editing and modification are still based on node operations. For example, users can add or remove the node from the node tree. It is impossible to edit or to separate objects just by manipulating the 3D shape objects, i.e., move the object directly. The basic edit functions they provide are limited. SwirlX3D Editor is also based on node operation, but it is the most suitable tool for us to integrate and reedit X3D files. Users can import different X3D files into the same environment, furthermore, users can pick up the shape which they want to edit directly. Nevertheless, the theory of node-tree based editing limits the flexibility and maneuverability. That is the reason why we hope to find a new method to solve this problem.

IntelligentBox is a prototyping software development system. The studying and developing of it since 1995 has established a dynamic functional combination mechanism

of interactive 3D graphic objects, which led us to realize a new approach to implant a large variety of interactive 3D applications.

IntelligentBox represents any 3D object as a Box which can transmit data and messages by the variables called Slot, in this way, the real time interactive operation can be easily implanted through slot connections among different boxes.

The creation and operation of boxes are simple and visual. Users just need to use the menu under the IntelligentBox environment to create a new Box and manipulate the mouse to combine two Boxes for constructing a new composite Box, which is more intuitive and effective than coding.

X3D itself provides some animation functions and supports the dynamically changing of the scene via programming and scripting languages. However, IntelligentBox can provide the real time interactive functions, which X3D is short of.

The remainder of this paper is organized as follows: Section 2 introduces the realization mechanism of X3D resource integration. Section 3 describes the mechanisms of X3D recourse assembly through reediting and reusing. Section 4 shows two application examples in the fields of mechanical engineering and e-commerce. In Section 5, we have a discussion to evaluate this new proposal. At last, our conclusion is given in Section 6.

2 X3D Resource Integration: A Realization Mechanism

How to import different X3D files into the IntelligentBox environment is the first question we have to answer. In this section, we are going to introduce the realization mechanism of X3D resource integration.

2.1 Tree Structures of X3D and IntelligentBox

Every X3D file has a node-tree structure, which includes all the information that describes the attributes of the file. There are a variety of nodes, each of which describes an object or its property. For instance, the Background Node includes the information of the background color and the Transform Node defines the value of its child node transformation. Using such a tree structure, each part of the final model is built up little by little, level by level. Fig. 1 shows a simple example of an X3D file. Fig. 1a shows its basic tree structure, in which node0 is the topmost node, which describes the transformation of its child nodes, namely node1-node2. Fig. 1b shows two boxes automatically created from the descriptions of node1 and node5, at the same time, the position of box1 is described by node2 while the color and the texture of box2 is described by node3 and node4.

IntelligentBox adopts a similar tree structure as X3D, furthermore, the tree structure used by IntelligentBox provides more formidable functions which we will discuss later. For the purpose of importing X3D file successfully, we plan to transmit the information of each node into the corresponding box. As shown in Fig. 2, we translate each node-tree structure into a box-tree structure. Some information nodes, such as appearance nodes and material nodes, are combined with other nodes like shape nodes. Once the corresponding box tree is obtained, a new 3D shape model will be created in the IntelligentBox environment. The new model includes all of the shape information of the original X3D model.

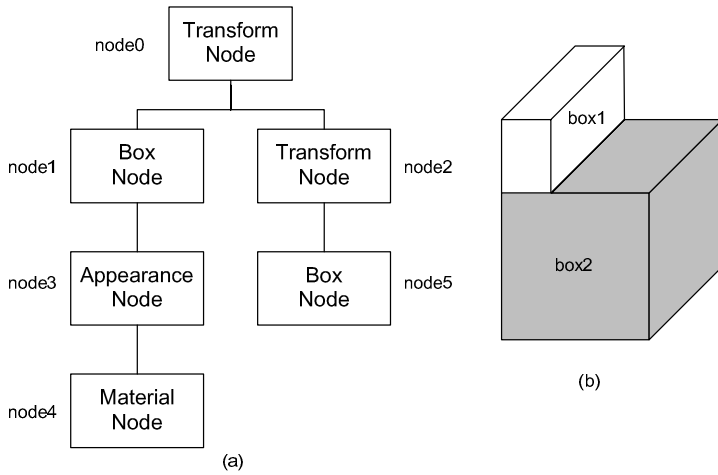


Fig. 1. A basic structure of an X3D model: (a) the tree structure; (b) the rendered result shown by the browser

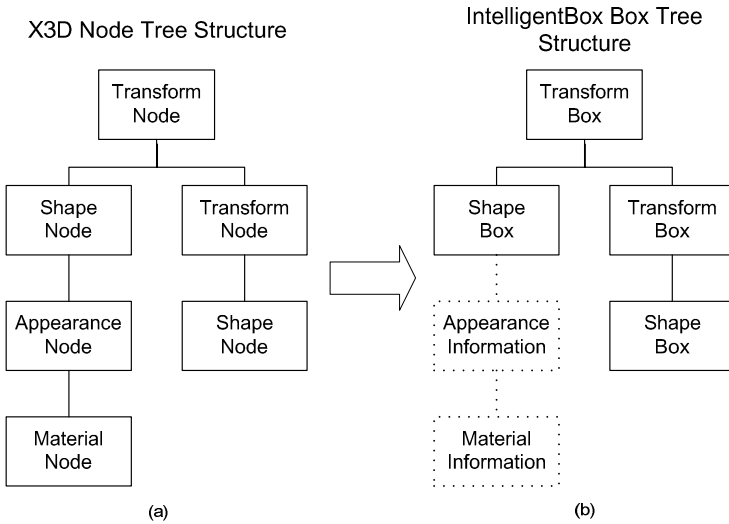


Fig. 2. The translation between: (a) an X3D node-tree structure and (b) IntelligentBox box-tree structure

2.2 Realization Method

By the translation between the two types of tree structures, we can successfully convert every shape node to an independent box, which means each shape may get a box function independently. In accordance with the diversity of X3D nodes, we have defined some corresponding boxes such as the BoxBox, SphereBox and so on.

The boxes we talked above provide similar shape descriptions as node descriptions. Therefore, the first step of the task is to read and analyze the node-tree structure of X3D. Because we just aimed at the X3D shapes, the nodes that provide animations and sensor functions need to be omitted. By analyzing the target X3D file, we get its shape node tree, which describes the basic shape structure of the model. The shape node tree is a key to open a new visual world, with which we can create the same model in IntelligentBox environment. After being read and analyzed, the shape node tree will be stored temporarily. It will be subsequently reread to recreate a 3D model.

After the new model is created, each sub-shape can be easily moved or separated since they are all independent boxes. Not only do they have the same shapes as the original files, but also have the attributes every box has. These attributes include geometry parameters, color specification parameters and texture parameters.

3 X3D Resources Assembly, Reediting and Reusing

We got the X3D shapes successfully by the method above. However, those models were still just 3D shapes without any function. For the reason that we aimed at the real interactive 3D environment and the adequate reusing of X3D files, the next step is to assemble, reedit and reuse those shape boxes.

3.1 Basic Mechanism of IntelligentBox

At first, let us see the basic mechanism of IntelligentBox. Each box of IntelligentBox consists of a model, a view, and a controller, called model-view-controller structure. A model contains the box state, which is stored in a set of slot variables defined in the model. A view is a significant factor of a box, by which, the appearance of a box on a computer screen is defined. Some slot variables are defined in the view part. Such slots include those holding geometry parameters and color parameters. A controller makes a definition of the box reaction to user operation. The message transmission among them is shown in Fig. 3.

A slot connection is shown in Fig. 4. Every box contains multiple slots which store the values of the corresponding attributes. By connecting different slots of different boxes, the message transmission channel is established. Slot connection is carried out by any of the three standard messages when there is a parent-child relationship between two boxes. As shown in Fig. 4, the slot3 of the box1 is connected with the slot2 of the box0, thus it becomes possible to exchange messages between these two slots.

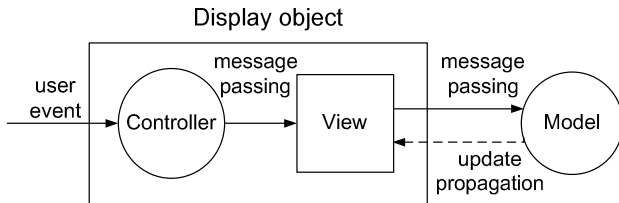


Fig. 3. The message transmission among the controller, the view and the model

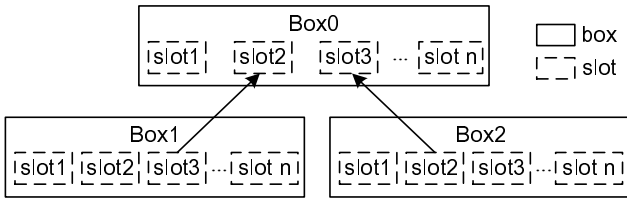


Fig. 4. An example of slot connections

The message transmission between two boxes is based on a slot connection. The prerequisite of a slot connection is the establishment of a parent-child relationship between the same pair of boxes. The box that embeds the other box in its local coordinate system is called the parent box while the embedded box is called a child box. There are three standard messages, namely, a 'set' message, a 'gimme' message and an 'update' message as shown in Fig. 5. These messages have the following formats:

- (1) ParentBox set <slotname> <value>.
- (2) ParentBox gimme <slotname>.
- (3) ChildBox update.

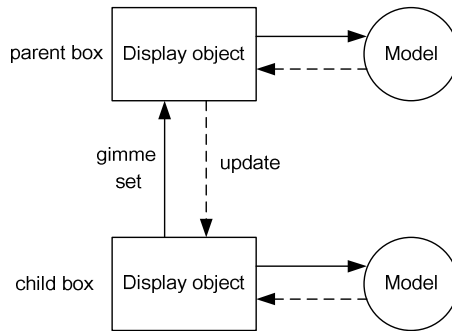


Fig. 5. Three standard messages between two boxes

The 'set' message and the 'gimme' message both go from a child box to its parent box. At the same time, the 'update' messages go in the opposite direction. A <value> in a format (1) indicates any value kept by the child box which issues this 'set' message, and a <slotname> in formats (1) and (2) indicates a user-defined slot of the parent box which receives these two messages. Each box has, for each slot, two procedures associated with a 'set' message and a 'gimme' message. These procedures are executed when the box receives either a 'set' message or a 'gimme' message to this slot.

IntelligentBox provides another method to establish a message connection between two boxes without parent-child relationship. As shown in Fig. 6, two boxes share the same model part because each box has a model and a display object. This method makes it possible to transmit data when the slot connection between them does not work because of the lack of a parent-child relationship. Two shared copies of a basic box storing a single value may work as two terminals of a cable. Users may use such

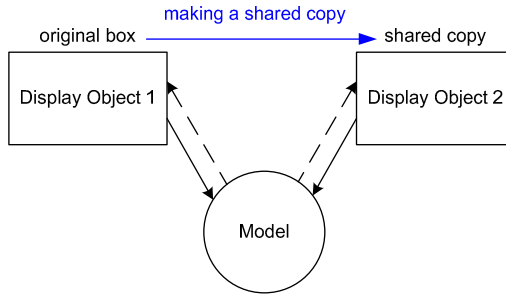


Fig. 6. A model shared by two boxes

a cable to connect two slots of two different boxes by connecting the two terminals of this cable to these slots.

3.2 New Model Assembly and Reediting

The most important point of object assembly is how to place the objects at specified locations in a certain environment. There are many kinds of assembly methods in Computer Aided Design (CAD) area. Take account of the real operation environment and realization mechanism, we decided to continue using the method adopted by X3D.

In X3D there is a node named Transform which controls the translation, rotation and scaling parameters by inputting corresponding values. Compared to this method, IntelligentBox provides an easy way to operate the objects. Users can move, rotate, scale any object just through mouse operation. Nevertheless, this method could not place an object accurately, what is why we wanted to import the same parameter controlling method.

Each box contains the parameters describing the real display mode of the object. Those parameters can be modified from outside through slot connections to change, for example, the position, size or orientation of the object. Since IntelligentBox provides a direct interface for data transmission named slot, we can store those transformation parameters into slots.

Such transformation parameters depend on a reference frame or a coordinate system. There are two types of reference frames: an absolute reference frame and a relative reference frame. Each absolute reference frame uses the system coordinate system. If a box has no parent box, the default reference frame is absolute. However, if a box has its parent box for some composition, we should use the relative reference frame defined by the parent box.

In IntelligentBox, each child box is bound by the local coordinate system spanned by its parent box. The child box will do the same transformation as its parent box does. Therefore, if we need the local coordinate system of any box, we just need to build the connection between those two boxes.

By these two methods, absolute reference frame and relative reference frame can be easily removed or built. For instance as shown in Fig. 7, there are two box models imported from X3D files. Now we are going to assemble box2 to box1, thus we need to use the local coordinate system of box1. The concrete operation step is to cut the

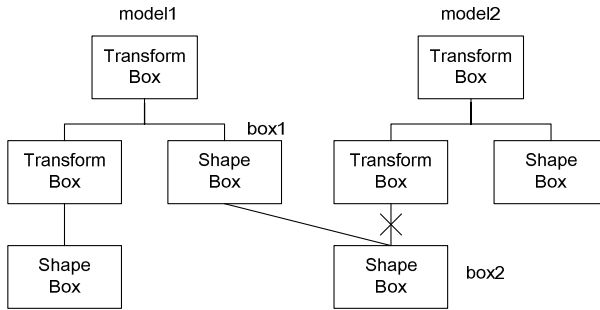


Fig. 7. An example of establishing a connection between two models

connection between box2 and its parent box at first, then connect box2 and box1. In this way, any modification of the transform parameters of box2 is based on the local coordinate system of box1.

What we just discussed above is the method of assembly of X3D models. The 3D shape boxes still hold the character parameters which describe the shape of the model, such as the radius of a sphere, the size of a box. We store those parameters in the slots, too. Therefore, we can reedit the model shape by changing slot values. A more detailed example about slot will be given in Section 3.3.

3.3 New Model Reusing

The other target of our research is to create a real interactive system. After reading, re-creating, assembling and reediting X3D files, the 3D shapes are displayed in the IntelligentBox environment. However, if they have no function, this research is meaningless. Actually, IntelligentBox has already provided many kinds of function boxes by which we can give full scope to our imagination to compose an interactive model.

Aiming at using those existed function boxes, three methods are proposed. The first method is to transfer the X3D shape view into a function box directly as shown in Fig. 8. In this way, we use the shape view of X3D and the controller and model part of a function box. It is a very simple method which can make sure those three parts in one single box. The message flow is direct and effective. However, this method is just suitable under some certain situation, such as transporting a shape view into a rotate box. If the view part does not match with the model part of function box, we need to use the second method.

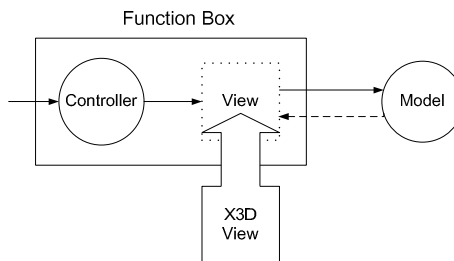


Fig. 8. The method for directly converting an X3D box to a function box

The second method is to build the connection between the X3D box and the function box as shown in Fig. 9, for example, a TimerBox. By sharing the slot value, X3D box will make the corresponding changes which are determined by the function box.

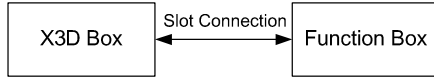


Fig. 9. Motion control of an X3D box by a function box through a slot connection

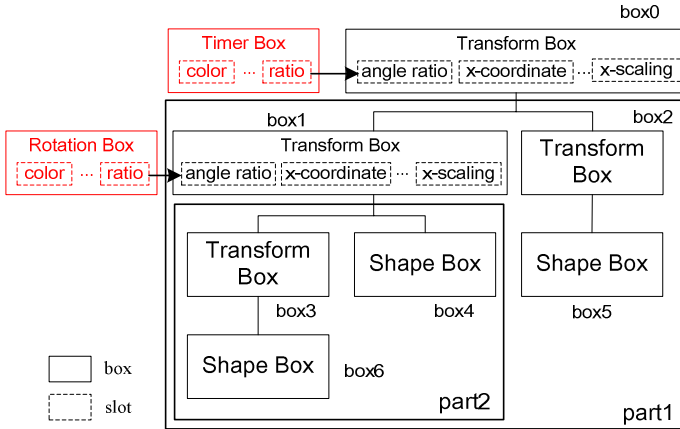


Fig. 10. The method of externally modifying the transformation information in the parent transform box by connecting function boxes through slot connections

As discussed in Section 2.1, the transformation information of X3D Shape Box is stored in the parent Transform Box. Meanwhile, the transformation of a child box is determined by the local coordinate system of its parent box. According to this, we can use the third method as shown in Fig. 10, to modify the transformation information in the parent box to control the child box. In Fig. 10, if we modify the transformation information of box0, the part1 will be transformed. In the same way, if we modify the transformation information of box1, part2 will be transformed. In this way, we can modify the transformation of each box easily. For instance, suppose that we connect a TimerBox with box0. After the calculation of the TimerBox, part1 will make the corresponding transformation along with the time change.

Each Transform Box has the following list of slots to control the transformation of its child box, i.e., the three coordinates, orientation, the angle ratio, and the axis of rotation. By controlling the slot values, we can transform corresponding child boxes. For example, as shown in Fig. 10, we connect a TimerBox to box0, and a RotationBox to box1 with the slots named 'ratio' and 'angle ratio'. After the calculation of the TimerBox, the value of rotation ratio will be stored in the 'ratio' slot, and then the value will be sent to the 'angle ratio' slot of box0, then part1 will be rotated. If we connect another RotationBox to box1, part2 will be transformed by namely rotating the RotationBox.

The bridge between an X3D box and a function box is built successfully by the above three methods. It opens an important door of the box gallery. IntelligentBox provides several kinds of functions, such as, RotationBox, ScalingBox, SliderMeterBox, StringBox, CameraBox and so on. We will give some application examples later.

4 Applications

This section introduces application examples to clarify our approach of integrating and reediting X3D under the environment of IntelligentBox.

We created two examples to explain how our approach works. The first one is a piston-crank mechanism [16] as shown in Fig. 11. In this system, the piston motion drives the crank motion, and then the crank drives the crankshaft rotation, so the fan rotates for the same ratio as the crankshaft.

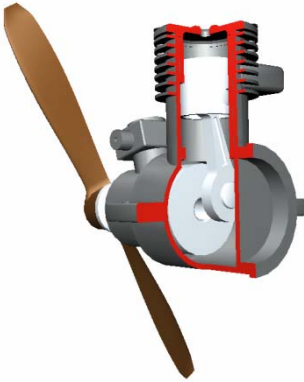


Fig. 11. The example of a piston-crank mechanism

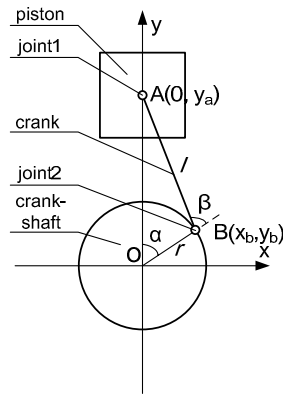


Fig. 12. The simplified model

In this example, we used the third method to control the transformation of each object. Fig. 12 is the simplified model of the piston-crank mechanism while Fig. 13 is the tree structure of this model. When the crankshaft rotates, by calculating the angle α , we can get the value of y_a which describes the position of the piston and joint1. Then the data is transmitted to the piston via the parent box which works as the transmission medium. At the same time, the value of angle β can be calculated using the value of α and is sent to the crank, thus the crank will rotate for the given angle value. In this way, no matter where the position of crankshaft is, all the other objects can get their own positions. The similar motion propagation will be made when we move the piston or the crank. Therefore, the real time interactivity for users to play with the composed object is successfully achieved.

The second example is a jeep model [17] as shown in Fig. 14. In this model, the body part was borrowed from one X3D file, the four wheels came from another X3D file. We transported the two axles into RotationBox, which can be controlled by a mouse for rotation. We connected four wheels to the two axles. By this way, wheels

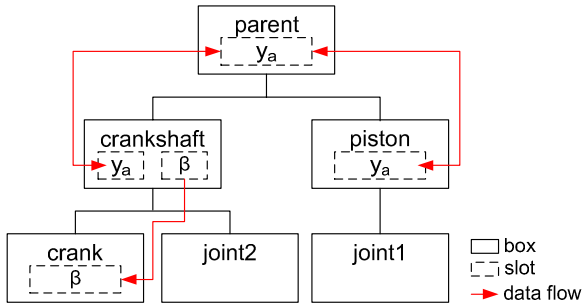


Fig. 13. The tree structure of the piston-crank mechanism

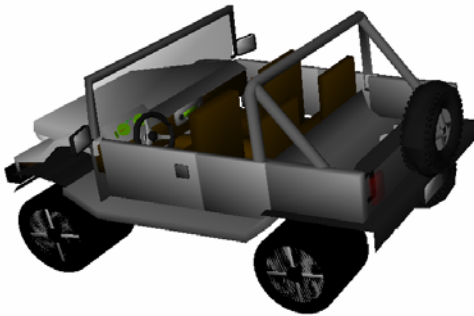


Fig. 14. The example of a jeep car

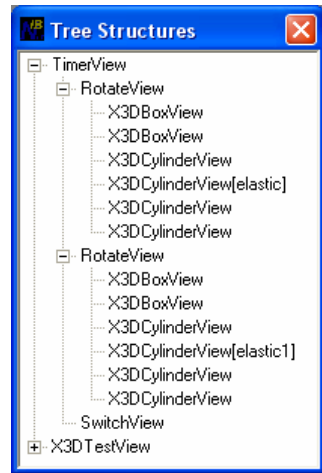


Fig. 15. The tree structure of the jeep car model

will rotate for the same ratio as the axles. Then we connected the two wheels with a TimerBox, which makes the ratio increase automatically. Then we connected a SwitchBox to the TimerBox, by which we can start or stop the animation manually. From the tree structure of this model as shown in Fig. 15, we can understand the whole structure clearly.

We added a RotationBox with which we can open the door of the car. We also built the rotation ratio relationship between the steering wheel and the front wheels, which made it possible to control the front wheels by rotating the steering wheel. Furthermore, by using a CameraBox and an EyeClickMoverBox, we can simulate the driver viewpoint, which is shown in Fig. 16, or change the viewpoint conveniently, and by using a LightBox, we can simulate the headlights and the taillights.

These examples proved that our research can be used in the fields of Virtual Reality, Web-based Customization, On-line Design, and so on.



Fig. 16. The simulation of the diver viewpoint

5 Discussion

We have already made some examples as in Section 4. IntelligentBox provides a very powerful platform for users to make their own ideas come true. By using it, the X3D shape and the box function are combined very well. X3D can provide more interactivity if it is empowered by IntelligentBox to add more reedit-ability.

However, we are still studying about how to convert the original function and sensor nodes of X3D file, i.e., event detection and dispatch mechanism in X3D, to composite boxes. At the same time, we are still trying to make it possible to utilize the JavaScript in IntelligentBox as the original X3D file does. If they are available, the function of the new system will be much stronger.

6 Conclusion

This paper introduces a method of how to integrate and reedit X3D shape resources under the environment of IntelligentBox. IntelligentBox is a platform for interactive 3D graphics. It defines an object as a reactive visual object named box.

Our research provides a real interactive environment for reading, re-creating, re-editing, assembling and reusing X3D format files, and more easy compositions of animation for X3D shape view. Compared to other X3D editors, the system we made developed a new way to control objects. By using the message transmission mechanism of IntelligentBox, X3D shapes own new attributes for receiving or sending information, which is the most important breakthrough. Therefore, users can easily use any X3D files to create their own model.

However, our research has some demerits as well. The new system could not utilize script program of the original X3D files yet.

This paper explains the methods we using for X3D file reading, re-creating, reediting, assembling and reusing, and gives some examples and applications using the new system.

References

1. Web 3D Consortium (2009), <http://www.web3d.org>
2. Brutzman, D., Daly, L.: X3D: extensible 3D graphics for Web authors. Academic Press, London (2007)

3. Geroimenko, V., Chen, C.: Visualizing information using SVG and X3D: XML-based technologies for the XML-based Web. Springer, Heidelberg (2005)
4. Daly, L., Brutzman, D.: X3D: extensible 3D graphics standard. *IEEE Signal Processing Magazine*, 130–135 (2007)
5. Anslow, C., Noble, J., Marshall, S., Biddle, R.: Web software visualization using extensible 3D (X3D) graphics. In: *Proceedings of the 4th ACM symposium on Software visualization*, Ammersee, Germany, pp. 213–214 (2008)
6. Jung, Y., Recker, R., Olbrich, M., Bockholt, U.: Using X3D for Medical Training Simulations. In: *Proceedings of the 13th international symposium on 3D web technology*, Los Angeles, California, pp. 43–51 (2008)
7. Brutzman, D.: Computer graphics teaching support using X3D: extensible 3D graphics for web authors. In: *ACM SIGGRAPH ASIA 2008 courses*, Singapore, vol. 23, pp. 161–162 (2008)
8. Jung, Y., Keil, J., Behr, J., Webel, S., Zöllner, M., Engelke, T., Wuest, H., Becker, M.: Adapting X3D for multi-touch environments. In: *Proceedings of the 13th international symposium on 3D web technology*, Los Angeles, California, pp. 27–30 (2008)
9. Okada, Y., Tanaka, Y.: IntelligentBox: a Constructive Visual Software Development System for Interactive 3D Graphic. In: *Proceedings of the Applications Computer Animation 1995*, Geneva, Switzerland, pp. 114–125, 213 (1995)
10. Okada, Y., Tanaka, Y.: Collaborative environments of IntelligentBox for distributed 3D graphics applications. *The Visual Computer* 14, 140–151 (1998)
11. Okada, Y.: IntelligentBox as component based development system for body action 3D games. In: *ACM International Conference Proceeding Series*, Valencia, Spain, vol. 265, pp. 454–457 (2005)
12. Bailey, M., Brutzman, D.: <https://savage.nps.edu/X3D-Edit> (2009)
13. Vivaty (2009), <http://www.vivaty.com>
14. BitManagement (2009), <http://www.bitmanagement.de>
15. Pinecoast Software (2009), <http://www.pinecoast.com>
16. Tecno Lution (2009), <http://www.tecnolution.com>
17. X3D Resources (2009), <http://www.web3d.org/x3d/content/examples>

Providing Relevant Answers for Queries over E-Commerce Web Databases

Xin Li¹, Jun Zhang², and Liping Li¹

¹ Computer center, Liaoning University of Technology,

² College of Electronic and Information Engineering, Liaoning University of Technology,
Jinzhou, China, 121001

lg_lx@163.com, marxj@163.com, liping_li@163.com

Abstract. Users often have vague or imprecise ideas when searching the e-commerce Web databases such as used cars databases, houses databases etc. and may not be able to formulate queries that accurately express their query intentions. They also would like to obtain the relevant information that meets their needs and preferences closely. In this paper, we present a new approach – QRR (query relaxation and ranking), for relaxing the initial query over e-commerce Web databases in order to provide relevant answer to the user. QRR relaxes the query criteria by adding the most similar values into each query criterion range specified by the initial query, and then the relevant answers which satisfy the relaxed queries could be retrieved. For relevant query results, QRR speculates the importance of each attribute based on the user initial query and assigns the score of each attribute value according to its “desirableness” to the user, and then the relevant answers are ranked according to their satisfaction degree to the user’s needs and preferences. Experimental results demonstrate that QRR can effectively recommend the relevant information to the user and have a high ranking quality as well.

1 Introduction

With the rapid growth of electronic commerce (e-commerce) all over the world, business transactions are carried out over the Web easily and speedily. The web has introduced an entirely new way of doing business, and also made it imperative for companies to optimize their electronic business. The ability of delivering personalized goods and services to the customers is becoming a key to the success of online businesses. Knowledge about the customer is fundamental for the establishment of viable e-commerce solutions. One way to achieve customization and personalization in e-commerce is the use of recommendation systems. Recommendation systems are usually used in e-commerce sites to suggest products to their customers and to provide consumers with more relevant information to enlarge their scale of purchase options. Nowadays, more and more e-commerce Web databases (in this paper “Web database” refers to the online database that is accessible only via Web form based interface) like scientific databases, used car databases, and houses databases etc. are available for lay users. Database query processing models have always assumed that the user knows what she wants and is able to formulate queries that accurately express her query intentions. However, users often have

insufficient knowledge about database contents and structure, and their query intentions are usually vague or imprecise as well. They would like to see more relevant information that relevant to their queries. Therefore, the query user submitted should be soft constraints for the query results in order to present more relevant information that can meet user's needs and preferences closely.

Example: Consider a used Car-selling e-commerce Web database D from Yahoo!Autos web site consisting of a single table $CarDB$ with attributes: Make, Model, Color, Year, Price, Location, Transmission, Mileage. Each tuple in $CarDB$ represents a used car for sale. Based on the database, the user may issues the following query:

Q: $CarDB$ (Make = Ford \wedge Model = Focus \wedge Price \leq 35000)

On receiving the query, the query processing model used by $CarDB$ will provide a list of *Ford Focus* that is priced below \$35000. However, given that *Buick LaCrosse* is a similar car, the user may also be interested in viewing all *Buick LaCrosse* priced around \$35000. The user may also be interested in a *Ford Focus* or *Buick LaCrosse* priced \$35500. Unfortunately, in the example above, the query processing model used by e-commerce web database $CarDB$ would not suggest the *Buick LaCrosse* or the slightly higher priced *Ford Focus* as possible answers of interest as the user did not specifically ask for them in her query. This will force the user to reformulate the queries several times for obtaining all relevant and similar results satisfying her needs. Therefore, providing some flexibility to the initial query can help the user to improve her interaction with the e-commerce system for capturing more relevant information.

In this paper, we propose a novel approach QRR (query relaxation and ranking), which can provide more relevant information by relaxing both categorical and numerical query criteria ranges of the initial query and does not require user feedback. This approach uses data and query workload mining in order to assist the relaxation process. Furthermore, based on the Web databases, too many relevant answers will be returned after a relaxed query. QRR ranks the relevant answers according to their satisfaction degree to the user's needs and preferences,

The rest of this paper is organized as follows. Section 2 reviews some related work. Section 3 proposes the query relaxation approach. Section 4 presents the relevant answers ranking approach. Section 5 discusses the key implementation details of our approach. The experiment results are presented in Section 6. The paper is concluded in Section 7.

2 Related Work

Data mining technologies have been around for decades, without moving significantly beyond the domain of computer scientists, statisticians, and hard-core business analysts. During the last years, researchers have proposed a new unifying area for all methods that apply data mining to Web data, named Web Mining. Web mining for E-commerce is the application of mining techniques to acquire this knowledge for improving E-commerce. Web mining tools aim to extract knowledge from the Web, rather than retrieving information. Commonly, Web mining work is classified into the following three categories [3]: Web Content Mining, Web Usage Mining (WUM) and Web Structure Mining. Web mining is concerned with the extraction of useful knowledge from the content of Web pages and databases, by using data mining. Web usage

mining, aims at discovering interesting patterns of use, by analyzing Web usage data. Finally, Web structure mining is a new area, concerned with the application of data mining to the structure of the Web graph. In this paper, we only take advantages of web data mining techniques to capture the most customers' preferences, which are used to relax the query criterion of the initial query.

Several researches have been proposed to handle the flexibility of queries in the database systems for presenting more information that relevant to the initial precise query. These researches can be classified into two main categories. The first one is based on Fuzzy Set Theory [12]. Tahani [11] firstly advocated the use of fuzzy sets for querying conventional databases. In recent years, the approaches proposed in [4] and [8] relax the query criteria by using the membership functions, domain knowledge and α -cut operation of fuzzy number. The second category focused on relaxing the query criteria range (query rewriting) such as [6] [7], and [9], which handle flexibility based on distance notion, linguistic preferences, and etc. However, it should be noted that the flexibility approaches based on fuzzy sets are highly dependent on the domain knowledge while in most cases the query rewriting approaches are not fully automatic and require the users to specify the distance metrics.

3 Query Relaxation

In order to recommend the similar answers to the user, we need to measure the similarity between the different pairs of values. The idea of query relaxation is to expand the original query criteria with similar values.

3.1 Problem Definition

Consider an autonomous e-commerce Web database table D with categorical and numerical attributes $A = \{A_1, A_2, \dots, A_m\}$ and a selection query Q over D with a conjunctive selection condition of the form $Q = \sigma_{\wedge_{i \in \{1, \dots, s\}} (A_i \theta a_i)}$, where $s \leq m$ and $\theta \in \{>, <, =, \geq, \leq, \text{between}, \text{in}\}$. Note that, if θ is the operator *between* (*not between*) and a_i is an interval which is represented by $[a_{i1}, a_{i2}]$, $A_i \theta a_i$ has the form of " A_i (*not between*) a_{i1} and a_{i2} ". Each A_i in the query condition is an attribute from A and a_i is a value (interval) in its domain. The set of attributes $X = \{A_1, \dots, A_s\} \subseteq A$ is known as the set of attributes specified by the query. When the query leads to empty (or little) or unsatisfactory answers, the original query should be relaxed to provide most similar answers for users, which can help users to broaden their purchasing options.

3.2 Relaxation of Categorical Query Conditions

Based on CarDB mentioned above, in order to recommend the similar cars to the user, it is necessary to evaluate the similarity between the different pairs of cars. For example, as mentioned above, *Buick LaCrosse* and *Ford Focus* are quite similar, they are family sedans, of comparable quality, and targeted to the same market segment. Thus, the similarity coefficient between *Buick LaCrosse* and *Ford Focus* may be 0.8, while the similarity coefficient between *Buick LaCrosse* and *Benz E-Class* may be 0.01.

We discuss an approach which is adapted from [7] for deriving such similarity coefficient by using database workload-*log of past user queries* on the database. The database workload information can reflect that the frequency with which database attributes and values were often requested by users and thus may be interesting to new users. The intuition is that if certain pairs of values $\langle u, v \rangle$ often “occur together” in the query workload, they are similar. For example, there may be queries with conditions such as “*Model IN (Camry, LaCrosse, Focus)*”. Such database workload information suggests that these manufacturers are more similar to each other than to, say *Benz E-Class*.

Let $f(u, v)$ be the frequency of the values u and v of categorical attribute A occurring together in a *IN* clause in the workload. Also let $f(u)$ be the frequency of occurrence of the value u of categorical attribute A in a *IN* clause in the workload, and $f(v)$ be the frequency of occurrence of the value v of categorical attribute A in a *IN* clause in the workload. Then, the similarity coefficient between u and v can be measured by using the following Equation (1).

$$VSim(u, v) = \frac{f(u, v)}{\max(f(u), f(v))} \quad (1)$$

The Equation (1) indicates that, the more frequently occurring together of the same pair of attribute values is, the larger their similarity coefficient is. By executing the equation (1) on the workload of CarDB, the similar values are obtained shown in Table 1.

Table 1. Similarity estimation

Value	Similar values	Similar degree
Make = Toyota	Honda	0.6
	Nissan	0.55
	Mazda	0.48
Model = Focus	LaCrosse	0.82
	Camry	0.74
	Avenger	0.52

According to the similarity coefficient between different pairs of categorical attribute values, QRR can provide the most similar information for the user when she submits a query with a relaxation threshold. According to the Table 1 shows above, given a query “*Model = Focus*”, if the relaxation threshold user provides is 0.8, QRR would translate the initial query into the relaxed query as: “*Model IN (Focus, LaCrosse)*”.

3.3 Relaxation of Numerical Query Conditions

In order to make the numerical query range contain more nearby numerical values, we need to evaluate the similarity between the different pairs of numerical values. Let $\{v_1, v_2, \dots, v_n\}$ be the values of numerical attribute A occurring in the database. Then the similarity coefficient $NSim(v, q)$ between v and q can be defined by Equation (2) as follows,

$$NSim(v, q) = 1 - \frac{|q - v|}{q} \tag{2}$$

Let α be the relaxation threshold user provided, q is the numeric value specified by the query, and then we can get the expanded range as,

$$[q - (1-\alpha)q, q + (1-\alpha)q] \tag{3}$$

Given the relaxation threshold user provided is 0.8, the query condition “*Price* ≤ 35000” of Q can be relaxed as “*Price* ≤ 42000”.

Generally speaking, with different threshold that the user chooses for the original query, our solution can translate the user’s original queries into the relaxed queries. For example, when the user issues a query “*Model = Focus and Price* ≤ 35000” with a relaxation threshold 0.8, according to our above discussion, the QRR will relax the query criteria and reformulate it as: “*Model IN (Focus, LaCrosse) and Price* ≤ 42000”.

4 Ranking Relevant Answers

In this section, we first discuss how to assign the attribute weights, and then present the approach for ranking the relevant answers.

4.1 Attribute Weight Assignment

We measure the importance of attributes (includes both the specified and unspecified attributes) by estimating the distribution difference of values of attribute A_i in database table D and query result T , the Kullback-Leibler distance is used for resolving this problem proposed by Su *et al.* [10]. Suppose that A_i is a categorical attribute with value set $\{a_{i1}, a_{i2}, \dots, a_{ik}\}$. Then KL-divergence of A_i from D to T is:

$$D_{KL}(D \parallel T) = \sum_{j=1}^k prob(A_i = a_{ij} | D) \log \frac{prob(A_i = a_{ij} | D)}{prob(A_i = a_{ij} | T)} \tag{4}$$

in which $prob(A_i = a_{ij} | D)$ refers to the probability that in D and $prob(A_i = a_{ij} | T)$ refers to the probability that in T .

To calculate the KL-distance in Equation (4) we need to obtain the distribution of attribute values over D (resp. T). The algorithm used in our paper is to build a histogram for each attribute A_j in the preprocessing stage. If A_j is a categorical attribute, each categorical value of A_j is used as a query to get its occurrence count. If A_j is a numerical attribute, an equal-depth histogram is built for A_j . After getting the histogram of A_j over D and T , the histograms are converted to a probability distribution by dividing the frequency in each bucket of the histogram by the bucket frequency sum of the histogram. Finally, the weight of each attribute can be assigned.

4.2 Ranking

Satisfaction Ranking

We rank the relevant tuples in query results according to the similarity between Q and an answer tuple t . The similarity measuring method is shown in Equation (5) as follows,

$$similarity(Q, t) = \sum_{i=1}^k W(A_i) \times \begin{cases} VSim(Q.A_i, t.A_i), & \text{if } Domain(A_i) = \text{Categorical} \\ NSim(Q.A_i, t.A_i), & \text{if } Domain(A_i) = \text{Numerical} \end{cases} \quad (5)$$

here, k is the number of attributes specified by the query, $W(A_i)$ is the importance weight of each specified attribute, $VSim(,)$ measures the similarity between the categorical values as explained above and $NSim(,)$ measures the similarity between the numerical values.

Obviously, our approach is to restrict similarity calculations only to the specified attributes, i.e., we only consider the projection of the database on the columns that are referenced in the query. Based on the Web databases, however, this similarity ranking function will partition the relevant answer into several equivalence classes, where tuples within each class share the same similarity score. To break ties among the tuples in each class, it is thus necessary to look beyond the attributes specified in the query (i.e. missing attributes).

Relevance Ranking

Consider the CarDB, for the query “Make = Ford and Model = Focus”, a car that is also has low “price” gets high rank because low priced cars are globally popular. In another word, the attribute value “Price = low” is globally important even though it is not specified by the car buyer in the query.

More formally, according to Section 3.2, the workload W is represented as a set of “tuples”, where each tuple represents a query and is a vector containing the corresponding values of the specified attributes [2]. Consider a query Q which specifies a set of attribute values. Recall the notion from Section 3.1, where X is the set of attributes specified in the query, and Y is the remaining set of unspecified attributes. Based on this assumption, for every value v in the domain of unspecified attribute Y_j , the global score of value v can be defined by

$$RF_i(v) = (F_i(v) + 1) / (FMax + 1) \quad (6)$$

where $F_i(v)$ be the frequency of occurrence of the value v of attribute A in the database workload; $FMax$ be the frequency of the most frequently occurring value in the database workload. For the numerical attribute, we discretize its domain into buckets, effectively treating a numerical attribute as categorical.

Based on the discussion above, we can get the relevance score of unspecified attributes values to the user preferences as follows,

$$Relevance(t) = \sum_{j=1}^n W(A_j) \times RF(v) \quad (7)$$

in which n is the number of unspecified attributes, $W(A_j)$ is the importance weight of each unspecified attribute. Consequently, according to the similarity and relevance score, the relevant answer tuples can be distinguished.

5 Implementation

In order to identify the QRR algorithm, we developed a monotype system illustrated in Figure 1 which contains pre-processing component and query processing component.

The main task of the pre-processing component is to compute and store a representation of the similarity function in auxiliary database tables. The categorical similarity analysis module is used for computing $Sim(u,v)$ for all distinct categorical values. Computing the similarity coefficient (i.e. $Sim(u, v)$) for similarity between all pairs of values u and v of any categorical attribute A involves scanning the database workload to compute co-occurrence frequencies of these values in the database workload, the results are stored in the table with columns $\{AttName, u_AttVal, v_AttVal, Similarity\}$ (we avoid space/time requirements quadratic in the size of A 's domain by only storing similarity coefficients that are above a certain threshold) and composite with a B+ tree index on $(AttName, u_AttVal, v_AttVal)$. The numerical similarity analysis module is used for computing the “distance” between two numerical values. For numerical attribute values, since we do not know what value q will be specified by a query, we cannot pre-compute the relaxed range of q ; thus we have to store an approximate representation of the smooth function (such as Equation (3)) so that the function value at any q can be retrieved at runtime. The approximated functions are stored in auxiliary tables. This intermediate layer now contains enough information for extending the original query and computing the ranking function.

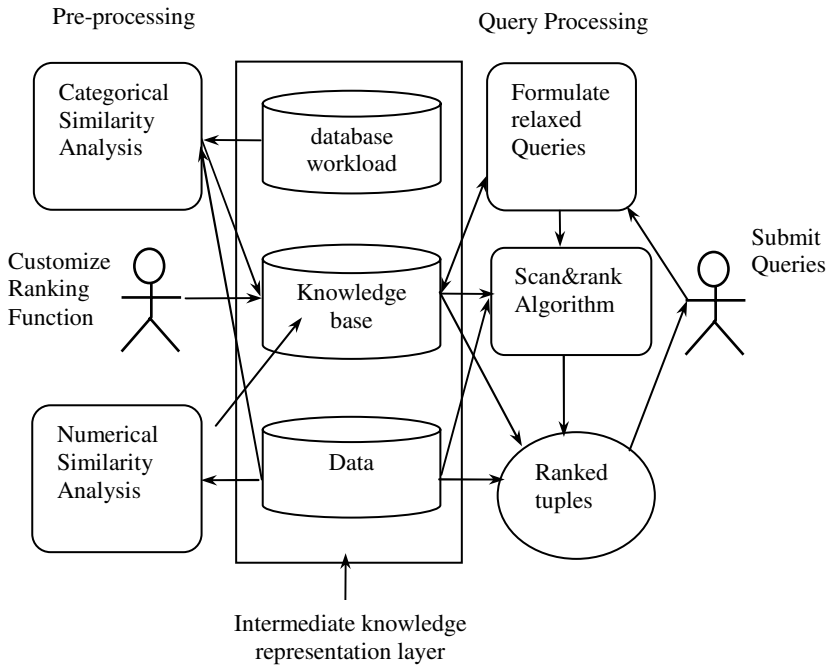


Fig. 1. QRR system architecture

The main task of the query processing component is, given a query Q and a relaxation degree α , to efficiently formulate the relaxed query and retrieve the relevant answers from the database using our similarity functions, and then ranks them using

ranking functions. Consider an incoming query Q , the formulate relaxed queries module translate the original query into relaxed query by interacting with knowledge base, and then use the scan&rank algorithm to select and rank the results.

6 Experiments

In this section, we describe our experiments, report the QRR experimental results and compare them with some related work.

6.1 Experimental Setup

The experiments aim at evaluating the ranking quality and performance quality of QRR algorithm, respectively. For our evaluation, we set up a used car database CarDB (*Make, Model, Year, Price, Location, Mileage*) containing 100,000 tuples extracted from Yahoo! Autos. The attributes *Make, Model, Year* and *Location* are categorical attributes and the attributes *Price* and *Mileage* are numerical attributes.

6.2 Quality of Ranking Evaluation

Besides QRR described above, we implemented two other ranking methods, which are described briefly below, to compare with QRR.

RANDOM ranking model: In the RANDOM ranking model, the tuples in the query results are presented to the user in a random order. The RANDOM model provides a base line to show how well QRR can capture the user behavior over a random method.

QFIDF ranking model: IDF technique which has been successfully used in the Information Retrieval field, which is used for ranking query results. The main focus of QFIDF proposed in [1] was on the Empty-Answers problem when the query is too selective, the idea of which is that takes advantage of the frequency of attribute values appearing in the workload as well as the database to determine the importance of attribute values and then rank the tuples in the answer according to the similarity between tuples and queries. The similarity between t and Q is simply the sum of corresponding similarity coefficients over all attributes.

To evaluate and compare the ranking precision of the various ranking algorithms, we requested ten people, some of them are actual used cars buyers, to provide us with queries that they would execute if they wanted to buy a used car. For the used car dataset, we generate 11 test queries. For each test query Q_i we generated a set H_i of 30 tuples likely to contain a good mix of relevant and irrelevant tuples to the query. We did this by mixing the top-10 results of each ranking algorithm, removing ties, and adding a few randomly selected tuples. Finally, we presented the queries along with their corresponding H_i 's (with tuples randomly permuted) to each user in our study. Each user's responsibility was to rank the top 10 tuples as the relevant tuples that they preferred most from the 30 unique tuples collected for each query. During ranking, they were asked to behave like real buyers to rank the records according to their preferences.

For formally comparing the ranking quality of the various ranking functions with the human responses, we used a standard *collaborative filtering metric R* which provided by Agrawal *et al* [1] to measure ranking quality. The Figure 2 shows the ranking precision of the different ranking algorithms for each test query. It can be seen that both QRR and QFIDF greatly outperform RANDOM. The averaged ranking precision of QRR and QFIDF were 0.74 and 0.55, respectively. While these preliminary experiments indicate that QRR is promising and better than the existing work, a much larger scale user study is necessary to conclusively establish this finding.

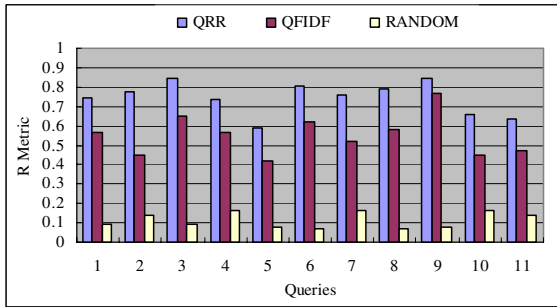


Fig. 2. The ranking precision of QRR, QFIDF and RANDOM for each test query

6.3 Quality of Ranking Evaluation

QRR system consists of two components: the pre-processing component and the query processing component. The similarity coefficients of categorical attribute values and standard deviations of numerical attribute values are computed in the pre-processing component. The similarity coefficients computing time depends on the number of records in the database workload. These quantities are stored as database auxiliary tables in a knowledge base.

The query processing component includes five modules: the query relaxation module, the attribute weight assignment module, the attribute-value similarity score assignment module, the ranking score calculation module and the ranking score sorting module. The first module has a time complexity of $O(n)$, where n is the number of the records in the auxiliary table, which was computed at the pre-processing component. Each of the second, third and fourth module has a time complexity of $O(n)$, where n is the number of query results, and the ranking score sorting module has a time complexity of $O(n \log(n))$. Hence, the overall time complexity for the query processing stage is $O(n \log(n))$.

Figure 3 shows the query execution time of the queries over CarDB as a function of the number of tuples in the query results. It can be seen that the execution time of QRR grows almost linearly with the number of tuples in the query results. This is because most of the running time is spent in the second, third and fourth modules of query processing part.

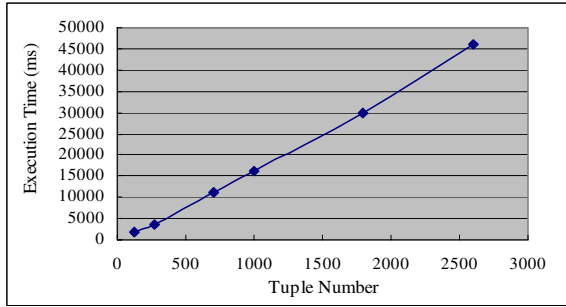


Fig. 3. Execution times for different numbers of query results for CarDB

7 Conclusion

In this paper, we present QRR (query relaxation and ranking), for relaxing the initial query over e-commerce Web databases in order to provide relevant answers to the user. QRR relaxes the query criteria by adding the most similar values into each query criterion range specified by the initial query, and then the relevant answers which satisfy the relaxed queries could be retrieved. For relevant query results, QRR speculates the importance of each attribute based on the user initial query and assigns the score of each attribute value according to its “desirableness” to the user, and then the relevant answers are ranked according to their satisfaction degree to the user’s needs and preferences. The QRR we proposed is domain independent. Experimental results demonstrate that QRR can effectively improve the answer recall and have a high ranking quality.

References

1. Agrawal, S., Chaudhuri, S., Das, G., Gionis, A.: Automated ranking of database query results. In: Proceedings of CIDR, pp. 171–183 (2003)
2. Chaudhuri, S., Das, G., Hristidis, V., Weikum, G.: Probabilistic ranking of database query results. In: Proceedings of VLDB, pp. 102–111 (2004)
3. Cooley, R., Srivastava, J., Mobasher, B.: Web mining: Information and pattern discovery on the World Wide Web. In: Proceedings of the ITCAL, pp. 121–130 (1997)
4. Hachani, N., Ounelli, H.: A knowledge-based approach for database flexible querying. In: Proceedings of DEXA, pp. 420–424 (2006)
5. Ichikawa, T., Hirakawa, M.: ARES: A relational database with the capability of performing flexible interpretation of queries. *IEEE Transactions on Software Engineering* 12(5), 624–634 (1986)
6. Kieling, W.: Foundations of preferences in database systems. In: Proceedings of VLDB, pp. 311–322 (2002)
7. Meng, X.F.: Providing flexible queries over web databases. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 601–606. Springer, Heidelberg (2008)

8. Ma, Z.M., Meng, X.F.: A knowledge-based approach for answering database fuzzy queries. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 623–630. Springer, Heidelberg (2008)
9. Rabitti, F.: Retrieval of multimedia documents by imprecise query specification. In: Bancilhon, F., Tsichritzis, D.C., Thanos, C. (eds.) EDBT 1990. LNCS, vol. 416, pp. 202–218. Springer, Heidelberg (1990)
10. Su, W.F., Wang, J.Y., Huang, Q.: Query result ranking over e-commerce web databases. In: Proceedings of CIKM, pp. 575–584 (2006)
11. Tahani, V.: A conceptual framework for fuzzy querying processing: a step toward very intelligent databases systems. *Information Processing Management* 13, 289–303 (1997)
12. Zadeh, L.A.: Fuzzy Sets. *Information and Control* 8(3), 338–356 (1965)

Detecting News Event from a Citizen Journalism Website Using Tags

Alton Y.K. Chua, Dion Hoe-Lian Goh, and Khasfariyati Razikin

Wee Kim Wee School of Communication & Information,
Nanyang Technological University, 31 Nanyang Link, Singapore 637718, Singapore
{altonchua, ashlgoh, khasfariyati}@ntu.edu.sg

Abstract. The accelerated news cycle and constantly emerging news-worthy events have led to ‘citizen journalism’ where people who are non-journalists collect, analyze and disseminate news pieces. This paper seeks to leverage tags drawn from iReport, an active citizen journalism Website to detect major news events. The goal is to examine the coverage and efficacy of news detected in iReport vis-à-vis those reported in the mainstream media. The data collection procedure involved manually culling major news events reported in Fox News between April 8 2008 and June 6 2008. Additionally, 81,815 tags from 15,216 documents were drawn from iReport during the same study period. Relative frequencies of all unique tags were used to check for spikes and bursts in the dataset. The results show that out of the 10 major news events reported in Fox News, five could be detected in iReport. This paper concludes by presenting the main findings, limitations and suggestions for future research.

Keywords: Social tagging, Citizen Journalism, Event Detection, Relative Frequency, User-generated Content.

1 Introduction

The advent of user-generated content has altered the way in which breaking news reaches audience globally. Users are capturing real-life events in text, photos and videos, and uploading them to the Internet on a daily basis. With an accelerated news cycle and numerous news-worthy events to cover, a new type of open-source reporting has emerged. Known as ‘citizen journalism’ (Gillmore, 2004), people who are non-journalists are beginning to collect, analyze and disseminate news pieces, not unlike what professional journalists do. The rising prevalence of citizen journalism is driven by factors including the lowering cost of Internet connectivity (Ahlers, 2006), the development of user-friendly online content management tools (Thurman, 2008) and the vast improvements made on consumer electronics such as cameras, video recorders and mobile phones which enable users to capture and distribute pictures and videos in electronic formats easily.

Citizen journalism was first thrust into the limelight in the aftermath of the 9/11 as eyewitnesses posted stories and images of the attack on the Internet. The 2006 Asian Tsunami disaster further accentuated the role of citizen journalism when video footages from survivors appeared on the Internet shortly after the disaster. Citizen journalism

entered the mainstream news media with the 2005 London terrorist bombings. The earliest photos of the blasts captured by ordinary citizens on their mobile phones were published on blogs and photo-sharing sites such as Flickr before appearing in national newspapers and television newscasts around the world the next day (Good, 2006).

Recognizing the value of grassroots' reporting, many mainstream media have augmented their print and television programming with citizen journalism. In particular, CNN launched iReport (www.ireport.com) in 2006 as part of such effort. In iReport, the disclosure of contributors' real identity is optional. Also, since submitted entries are not subjected to any editorial censorship, they become available immediately once they have been posted. Readers can access the submitted entries in multiple ways including listing them by viewership popularity, or filtering them on the basis of tags used by contributors. To date, iReport has attracted more than 100,000 postings from citizen journalists, along with some half a million of tags associated to the postings.

Even as the credibility and legitimacy of iReport have become more established, little research has been done to exploit the huge reservoir of data available. For this reason, this paper seeks to leverage tags drawn from iReport to detect major news events. The goal is to examine the coverage and efficacy of news detected in iReport. Coverage refers to the extent to which news reported in the mainstream media can also be detected in iReport, while efficacy refers to the promptness of news reported in the mainstream media vis-à-vis those detected in iReport.

To achieve this goal, major news events that took place between April 8 2008 and June 6 2008 were manually culled from Fox News. Thereafter, tags associated with news items submitted by citizen journalists within the same time period from iReport were analyzed. Instead of absolute frequencies, the relative frequencies of tags were used (Thelwall, 2006). The relative frequencies of tags are vestige of emerging popular news pieces. A spike in a time-series analysis of tags, for example, helps identify the day on which a certain event is highly discussed. Events detected in iReport through the tags were then compared against those reported in Fox News.

To the best of our knowledge, the use of relative frequencies of tags to detect news events has not been attempted. Our work can therefore be used as a springboard to develop more ideas and approaches for detecting events from social tagging systems. The next section presents the literature related to citizen journalism and social tagging. Following that, the Methodology section explains the data collection and analysis procedures. The detailed results are discussed in the Results and Analysis section. Finally, the Discussion and Conclusion section highlights two main findings and offers a few future research directions for scholars interested to study news event detection using tags.

2 Literature Review

For most of the twentieth century, news was primarily delivered by the press and television/radio broadcasting (Curran & Seaton, 2003). However, as the world becomes more connected in the twenty-first century through the Internet, the public's consumption pattern of news and attitude towards news reporting change. Not only do people expect to have access to the breaking news anytime and anywhere, the once mere consumers of news start to participate in the process of citizen journalism, helping to create a massive

conversation among themselves and anyone interested. Amid the mushrooming of blogs and wikis that publish independent news-related content, several mainstream media organizations are making efforts to involve non-journalists who are keen on reporting news. For example, The Washington Post (www.washingtonpost.com) embeds live Technorati updates for each of its stories, paving the way for its readers to become citizen journalists and commentators in the web community (Good, 2006). BBC (www.bbc.com) probes its news users of their views about the news and then publishes them in a particular section of the news product while MSNBC (www.msnbc.com) makes provision for editors to suggest assignments for anyone who wish to report on specified aspects of news stories that are unfolding (Nip, 2006). The response from citizen journalists has hitherto been overwhelming. OhmyNews.com, a South Korean online newspaper, has more than 37,000 registered contributors; Britain's second most popular news website, Guardian.co.uk, hosts a 'News' message board to which readers contributed more than 600,000 messages between 1999 and 2005 (Thurman, 2008).

One of the major advantages of citizen journalism is that citizens with cameras can capture images of news events more promptly than professional journalists, at least in the early minutes of the events. Also, if equipped with mobile access to the Internet, citizens can broadcast their photos, along with text content, immediately to the world (Tilley & Cokley, 2008). Furthermore, given its open and participatory nature, diverse views of a given news event can be expected (Angelo, 2008). Thus, the corpora of content created by citizen journalists could potentially be used to mine for the occurrences of major events.

Current event detection techniques generally seek to determine whether a news story contains an event by comparing the similarity of features between the new story and past news stories (Wei & Lee, 2004). Grouping of events by their relative similarities and differences helps to track events across time. This has been introduced in text-based topic detection and tracking (TDT), which uses lexical similarity of document texts to generate coherent clusters, in which elements in the same cluster share an identical topic (Allan, et al., 1998). Another approach is to consider the time gap between events. Time gap between bursts of topically similar stories is often an indication of different events and the incorporation of a time window for event scoping has commonly been adopted (Yang, et. al., 1998, Wei and Lee, 2004).

Event detection techniques can be classified into two forms: retrospective detection and online detection. Retrospective event detection (RED) consists of the discovery of previously undefined events from a chronologically ordered accumulation of news stories (Li, et al., 2005) while online detection strives to identify the onset of events emerging from live news feeds in real time. (Wei and Lee, 2004) Both forms of detection rely on historical news stories which contain two kinds of information, namely, contents and timestamps. Many previous studies tend to focus on the exploitation of contents but the usefulness of time information has often been ignored (Li et al., 2005). Taking the RED approach, this study used time information in the analysis.

In a parallel line of development, social tagging has been gaining traction on the Internet (Razikin, 2008). Since 2004, an increasing number of Websites including del.icio.us, Flickr, YouTube as well as those dedicated to citizen journalism allow users to annotate Web resources such as Web pages, images and videos using user-defined tags (Sen, et al., 2006). The confluence of the rising popularity of social tagging and citizen journalism presents the opportunity to investigate the use of tags to

detect news events. When an event of wide-spread significance occurs, a sharp rise in Web activity related to that event is observable (Neo, 2007). In the case of citizen journalism Websites, the occurrence of an event is characterized by an increased in the number of uploads. Also, different users are likely to be re-using the same tags in their postings.

Hitherto, there are very few works that rely on tags to detect events from news corpora. Tags are preferred over titles and descriptions for three reasons. One, tags are usually short and do not contain stop-words. Two, tags were found to be different from words found in titles and descriptions, suggesting their role in explicating the content (Greisler & Burns, 2007). Three, the process of generating tags appear to be similar to the process of generating search terms for subsequent retrieval (Furnas, et al., 2006). This means they could be suitable proxy of the content to which they have been associated (Goh, et al., 2009). This study is therefore a timely endeavour to push the frontier of event detection research. The findings can also broaden our understanding of the use of tags in making sense of a larger body of documents.

3 Methodology

A two-step data collection procedure was carried in this study. First, news on actual events that took place between April 8 2008 and June 6 2008 were manually drawn from Fox News. As with previous research (for example, Allan, et al., 1998), this was done in the absence of a baseline for the dataset. Representing a mainstream news source, Fox News was selected because it is a well-regarded organization known for its conservative news coverage (Zelizer, 2005). In the second step, all documents posted within the same time period from iReport was extracted. iReport has been chosen because it represents a thriving citizen journalism Website, attracting more than 100,000 news postings since its launch in 2006. Furthermore, it possesses the features of a social tagging system that allows contributors to freely assign tags to their documents.

A total of 19,960 documents were extracted from iReport. Documents which do not contain any tags or date information were removed from the dataset. Tags from the resulting 15,216 documents were used for analysis. These documents, which attracted 5.04 tags each, were contributed by 6,640 users. The total number of tags collected was 81,815, of which 9,204 were unique. The tags were analyzed using relative frequency, $r(d)$, which is computed by dividing the frequency of a tag on a given day by the total number of tags on that day (Thelwall, et al., 2006).

To detect news events, two methods were used, namely, a spike and a burst (Gruhl, et al., 2004). Both were used because preliminary experiments revealed neither performed consistently better than the other (Thelwall, et al., 2006). A spike is defined as the occasion on which the relative frequency of a tag, $r(d)$, on a given day, d , is at least five times higher than the average relative frequency of the tag of all previous days $\frac{1}{d-1} \sum_{i=1}^{d-1} r(i)$. Indicating a sudden surge in the relative frequency of a tag, a spike could point to the occurrence of an event to which the tag is associated.

A burst is defined as the occasion on which the minimum $r(d, 3) = \min\{r(d), r(d+1), r(d+2)\}$ of the relative tag frequencies on three consecutive days $d, d+1, d+2$

was at least five times higher than the average relative word frequency of all previous days. A burst thus indicates a period of at least three consecutive days in which an event was highly discussed.

All tags collected during the study period were compiled. Their relative frequencies on a daily basis were also computed, giving rise to three sets of tags for in-depth analysis. The first set comprised the top 20 tags ranked according to relative frequencies. Correlation analysis which was used to check for co-occurrences among these tags helped establish if two tags were associated to the same event. The second set comprised the top 20 tags ranked according to relative frequencies which have at least a spike. The final set comprised the top 20 tags ranked according to relative frequencies which have at least a burst.

4 Results and Analysis

Table 1 shows 10 major news events manually drawn from Fox News. The themes of these events include natural disasters, terrorism, politics and issues related to the economy.

Table 1. Events manually detected from Fox News between April 8 and June 6 2008

	Event	Date Reported
1	Olympic torch relay heads to San Francisco after violent protests in London and Paris.	Apr 8
2	Food prices rise	Apr 15
3	A magnitude 5.2 earthquake occurs outside of West Salem, Illinois, one of the strongest earthquakes in the midwestern states in 40 years.	Apr 18
4	Chaiten Volcano erupts in Chile, forcing the evacuation of more than 4,500 people.	May 2
5	Over 130,000 people in Myanmar are killed by Cyclone Nargis	May 3
6	Earthquake of magnitude 7.2 occurs in Sichuan, China, killing over 60,000 people	May 12
7	A series of bomb blasts kills at least 63 and injures 216 in Jaipur, India.	May 13
8	Fuel price rises. Protests due to rise in fuel prices.	May 23
9	Earthquake aftershocks in China destroy 71000 homes	May 25
10	Barack Obama becomes the presumptive nominee of the Democratic Party, becoming the first African-American to do so in a major U.S. political party.	Jun 4

Table 2 shows the first set of top 20 tags sorted in reverse order of their relative frequencies. Given that relative frequencies are computed on a daily basis, a given tag could appear more than once in the list. Pair-wise correlation analysis was performed among unique tags. Due to space constraints, only correlated tags with values more than 0.6 are shown.

The tag ‘Gas_prices’ was not correlated to any tags in the Table 2. Neither were the tags ‘Food_cost’ and ‘Comment’. The tags ‘China’, ‘Torch’, ‘Olympics’, ‘Olympic_torch’, ‘Sports’, and ‘Beijing_Olympics’ were highly correlated among themselves. Associated to the same event, they spotted high relative frequencies unanimously on

Table 2. Top 20 tags in reverse order of relative frequency and their pair-wise correlations

r(d)	Date	Tags	Correlated Tags
0.158	Jun 4	Gas_prices	-
0.098	Apr 10	China	Earthquake (.63); Olympics (.62); Olympic_torch (.61)
0.096	May 8	Election08	Obama (.85); Situation_room (.62)
0.087	May 23	Economy	Energy (.7); Big_oil (0.69)
0.083	May 23	Gas_prices	-
0.078	Apr 10	Torch	Sports (.97); Olympic_torch (.9); Olympics (.89); Beijing_olympics (.89)
0.076	Apr 10	Olympic_torch	Beijing_olympics (.99); Torch (.9); Sports (.89); China (.61)
0.076	Apr 10	Olympics	Beijing_olympics (.98); Torch (.89); Sports (.87); China (.62)
0.076	Apr 10	Sports	Torch (.97); Beijing_olympics (.9); Olympic_torch (.89); Olympics (.87)
0.075	Apr 10	Beijing_olympics	Olympic_torch (.99); Sports (.9); Olympics (.89); Torch (.89);
0.074	May 8	Obama	Situation_room (.89); Election08(.85)
0.073	May 23	Big_oil	Energy (.98); Economy (.69)
0.073	May 23	Energy	Big_oil(.98); Economy (.7)
0.068	May 12	Food_costs	-
0.066	May 19	China	Earthquake (.63); Olympics (.62); Olympic_torch (.61)
0.065	May 19	Earthquake	China (.63)
0.062	Apr 9	China	Earthquake (.63); Olympics (.62); Olympic_torch (.61)
0.061	May 8	Situation_room	Obama (.89); Election08(.62)
0.060	May 15	China	Earthquake (.63); Olympics (.62); Olympic_torch (.61)
0.057	Apr 29	Comment	-

April 10, representing a two-day delay from the time the news was reported in Fox News. The tags ‘Big_oil’, ‘Energy’ and ‘Economy’ were also highly correlated. So were two pairs of tags, namely, ‘China’ and ‘Earthquake’; and ‘Obama’ and ‘Election08’.

Table 3 shows the second set of top 20 tags sorted in reverse order of their relative frequencies which have at least one spike during the study period. The tag ‘Gas_prices’ was found to have a high relative frequency. It also saw two spikes, one on May 23 and the other on June 4. From Fox News, it was found that oil prices rose above \$135 a barrel for the first time on May 23. Furthermore, from May 28 till the beginning of June, there were numerous protests by lorry drivers against the hike in fuel price from various countries such as UK, France, Spain and Thailand. This coincides with the second spike observed on June 4.

It is interesting to note that while tags such as ‘China’, ‘Torch’, ‘Olympics’, ‘Olympic_torch’, ‘Sports’, and ‘Beijing_Olympics’ are listed in Table 2 as having high relative frequencies, they are not featured in Table 3 because they did not see any spikes. It could be attributed to the fact that the Olympic torch relay event started on March 24, prior to the period of study. During the month of April, the event had been drawing sustained attention from iReport contributors, but did not attract any sudden increase in the number of postings. By May and June, the newsworthiness of the event had diminished. Similarly, the tag ‘Food_cost’ had a high relative frequency May 12 but did not see any spike. The news about rising food prices was first reported in Fox News on April 15, almost a month earlier, and became a persistent issue in both the mainstream media and iReport.

Another group of co-occurring tags, namely, ‘Big_oil’, ‘Economy’ and ‘Energy’ also saw a spike on May 23. This could also be traced to the increase in gas prices described earlier.

Table 3. Top 20 tags which saw at least a spike during the study period

r(d)	Ave r(d)	r(d)/ave r(d)	Date	Tags
0.158	0.009	17	Jun 4	Gas_prices
0.096	0.014	7	May 8	Election08
0.087	0.010	8	May 23	Economy
0.083	0.004	21	May 23	Gas_prices
0.074	0.007	11	May 8	Obama
0.073	0.000	247	May 23	Energy
0.073	0.002	46	May 23	Big_oil
0.068	0.004	18	May 12	Food_costs
0.065	0.006	11	May 19	Earthquake
0.057	0.005	12	Apr 29	Sound_off
0.057	0.005	12	Apr 29	Opinion
0.057	0.005	12	Apr 29	Comment
0.055	0.007	8	May 12	Weather
0.055	0.006	10	May 12	Storm
0.053	0.003	16	May 15	Earthquake
0.052	0.002	33	Jun 1	2008_election
0.049	0.009	5	May 2	Opinion
0.047	0.001	55	May 22	Cancer
0.047	0.002	25	May 2	Immigration
0.046	0.009	5	May 25	Earthquake

The tag ‘Earthquake’ saw three spikes on May 15, May 19 and May 25. Furthermore, it was highly correlated to the tag ‘China’, as seen in Table 2. Reported in Fox News was the earthquake in Sichuan, China on May 12. There seems to be a three-day delay before the tag ‘Earthquake’ saw the initial spike. A closer examination of the iReport postings revealed that between May 15 and May 19, hundreds of videos and photos were uploaded by contributors who managed to capture the scene of the disaster and the on-going rescue operations. Reported in Fox News on May 25 was a catastrophic aftershock in Sichuan, China which destroyed thousands of homes. This news event corresponded to the third spike on May 25.

The highly correlated pair of tags ‘Obama’ and ‘Election08’ saw a spike on May 8. From news archives, it was found that Barack Obama won the North Carolina Democratic primary at the Super Tuesday III on May 6. The event was detected in iReport, albeit after a two-day lag. The tag ‘2008_election’ which saw a spike on June 1 was probably indicative of the built-up leading to the presidential nomination on June 3 but news about Obama’s victory was not detected in iReport.

The tags ‘Storm’ and ‘Weather’ both saw a spike on May 12. When their individual relative frequencies were plotted over time, it was found that both tags showed high co-occurrences. A check on Fox News revealed that major storms struck three states of the United States on May 11.

The tags ‘Opinion’, ‘Sound_off’ and ‘Comment’ which saw a spike on April 29 could not be used to detect any major news events. On closer examination, these tags were used mainly by contributors of iReport who wanted to share their perspectives on the U.S. election. Likewise, the tags ‘Immigration’ and ‘Cancer’ spiked on May 2 and May 22 respectively but were not traceable to any breaking news. Postings with the tag ‘Immigration’ concerned activists demanding citizenship opportunities for U.S. illegal immigrants while those tagged with ‘Cancer’ were found to be related to real-life stories of contributors who survived from cancer or had relatives who suffered from cancer.

Table 4 shows the final set of top 20 tags which saw a burst during the study period. The list is sorted in reverse order of their minimum relative frequencies which are at least five times their average relative frequencies of all previous days.

Table 4. Top 20 tags which saw a burst during the study period

Min r(d, 3)	Ave r(d)	Min r(d, 3)/ ave r(d)	Date	Tags
0.0422	0.0058	7	May 19 – 21	Earthquake
0.0306	0.0007	45	May 13 – 15	Earthquake
0.0242	0.0048	5	Apr 28 – 30	Opinion
0.0224	0.0040	6	May 23 – 25	Gas_prices
0.0220	0.0003	64	May 21 – 23	Cancer
0.0212	0.0016	14	Jun 1 – 3	2008_election
0.0176	0.0025	7	May 24 – 26	Memorial_day
0.0078	0.0009	9	May 22 – 24	Cancer
0.0064	0.0011	6	Apr 16 – 18	Lifestyle
0.0058	0.0000	197	May 29 – 31	Commute
0.0057	0.0006	10	Apr 18 – 20	Support
0.0055	0.0010	6	Jun 3 – 5	Morning_express
0.0048	0.0003	16	May 23 – 25	Energy
0.0046	0.0003	13	Jun 3 – 5	France
0.0044	0.0003	13	May 1 – 3	Immigration
0.0034	0.0007	5	May 29 – 31	Crisis
0.0033	0.0003	12	May 25 – 27	Space
0.0026	0.0001	39	Jun 3 – 5	Nascar
0.0025	0.0005	5	May 12 – 15	Freedom
0.0022	0.0001	19	May 25 – 27	Memorial

Table 5. Summary of news events detection in iReport

	Event	Date Reported in Fox News	Detected in iReport	Delay in news detection
1	Olympic torch relay heads to San Francisco after violent protests in London and Paris.	Apr 8	Yes	2 Days
2	Food prices rises	Apr 15	Yes	Almost a month
3	A magnitude 5.2 earthquake occurs outside of West Salem, Illinois, one of the strongest earthquakes in the midwestern states in 40 years.	Apr 18	No	NA
4	Chaiten Volcano erupts in Chile, forcing the evacuation of more than 4,500 people.	May 2	No	NA
5	Over 130,000 people in Myanmar are killed by Cyclone Nargis	May 3	No	NA
6	Earthquake of magnitude 7.2 occurs in Sichuan, China, killing over 60,000 people	May 12	Yes	3 days
7	A series of bomb blasts kills at least 63 and injures 216 in Jaipur, India.	May 13	No	NA
8	Fuel price rises. Protests due to rise in fuel price	May 23	Yes	No delay
9	Earthquake aftershocks in China destroy 71000 homes	May 25	Yes	No delay
10	Barack Obama becomes the presumptive nominee of the Democratic Party, becoming the first African-American to do so in a major U.S. political party.	Jun 4	No	NA

Seven tags which saw at least a spike also saw a burst. These were 'Earthquake', 'Gas_prices', '2008_election', 'Cancer', 'Opinion', 'Immigration' and 'Energy'. The tag 'Earthquake' could be traced to the earthquake in China while the tag 'Gas_prices' was related to the rising fuel cost and the tag '2008_election' referred to the U.S. elections. Having been featured in the mainstream media during the period between April and June, these news events could also be detected from iReport.

The tag 'Memorial_day' saw a burst from May 24 through May 26. This was attributed to the U.S. public holiday, Memorial Day, which falls on May 26.

The rest of the tags in Table 4 could not be used to detect any major events. For example, the burst seen in the tag 'France' from June 3 through June 5 was attributed to the surge in postings uploaded by contributors who wanted to share their holiday experiences in France. The burst seen in the tag 'Space' from May 25 through May 27 stemmed from a sustained interest among contributors who uploaded numerous photos related to outer space. Likewise, the tag 'Freedom' which saw a burst from May 12 through May 15 was related to postings that dealt with the issue of freedom of speech but could not be associated to any specific major news event.

Table 5 summarizes the detection of news events in iReport during the study period.

5 Discussion

News typically follows a life-cycle after it has been published. A lead-time may transpire before the public's interest is aroused, and intense discussion ensues. Thereafter, interest will wane until some new related events trigger the next wave of discussion (Zhao, et al., 2006). The fluctuations in the number of postings, along with the variations of user-assigned tags and their relative frequencies open the possibility for major news events to be detected.

Riding on the increasing popularity of citizen journalism, this paper takes the RED approach and details the use of spikes and bursts seen in the relative frequencies of tags to detect major news events from iReport. The performance of analysing the top 20 tags with at least a spike turned out better than that of the top 20 tags with at least a burst. Of the 10 major news events reported in Fox News, two were detected without delay. They were related to the fuel prices hikes reported on May 23 and the earthquake aftershocks in China reported on May 25. Two other news events related to the Olympic torch relay reported on Apr 8 and the Sichuan earthquake reported on May 12 were detected after a two-day and three-day delay respectively. The news event that was detected after almost a month was related to food prices hikes reported on April 15. The seemingly long delay could be attributed to the fact that the full impact of the rising costs in food was not immediately felt at the point of reporting. Furthermore, the increases could be staggered among different food items.

Five major news events which could not be detected were the earthquake in Illinois, the volcano eruption in Chile, the cyclone in Myanmar, the bomb blast in India and the presidential nomination of Obama. One possible reason why they were not detected could be that people residing at the site of the event, for example, in Chile or in India, either did not engage in citizen journalism or that they had used local Websites instead of iReport to upload their postings. Also, the fact that Myanmar is known to be a closed economy that restricts its citizen from providing information to the

world outside could also explain why the cyclone disaster was not detected. The country's poor Internet connectivity and low mobile phone penetration also curtail active citizen journalism. Obama's victory on the Super Tuesday III could be detected, but the euphoria probably had subsided by the time he became the presidential nominee.

While not all news events reported in the mainstream media could be detected, the results revealed a number of interesting news pieces from iReport. Among these were real-life stories about surviving Cancer patients and those whose relatives had suffered from Cancer, the demand for citizenship opportunities for U.S. illegal immigrants, and Memorial Day.

6 Conclusion

From this study, two main findings can be drawn. First, even though citizen journalism transcends national boundaries via the Internet, the results suggest that news uploaded in iReport may not necessarily be international in outlook. Major events which occurred in Chile, Myanmar and India, for example, were not detected. The postings related to these events did not surge sufficiently to see a spike or a burst. iReport, which is part of the U.S.-based CNN, tend to attract contributors who appeared also to be from the U.S.

Second, despite the fact that tags were freely assigned by contributors, majority of the tags with high relative frequencies appears to be descriptive and are of an acceptable quality, insofar as detecting major news events is concerned. When a major news event occurs, users tend to use identical tags or tags with similar themes. This causes a sudden surge to the relative frequencies of tags, creating spikes or bursts. In a way, this phenomenon supports the Wisdom of Crowd theory (Vapnik, 1995) which postulates that the knowledge that comes from a large group of users will be more reliable than a knowledge that comes from an individual.

Two limitations inherent in this paper must be acknowledged. One, the computation of relative frequencies did not take into account whether a given tag had been reused multiple times by the same contributor on a given day. Should a contributor submit a large number of postings with the same tag repeatedly, a spike or a burst might have been created. Two, in this current study, the original list of 10 major news events were manually culled only from Fox News. For a more robust list, other mainstream media such as BBC and newswire such as AFP could have also been referred.

With massive amounts of new materials emerging from the Internet daily, news consumers are invariably inundated with information. It would be difficult, for example, to find and track news events which are of interest to them. News event detection through tags certainly offers sufficient depth and breath for further investigation. Thus, one direction for future research would be to replicate this study and compare results obtained among other citizen journalism Websites such as www.citizenside.com, (affiliated to The France-Press Agency) www.jasminenews.com (from Sri Lanka) and www.merineews.com (from India). A second suggestion for future research is to augment the event detection technique used in this paper. Possible data points admitted for analysis could include the tagging pattern of individual contributors, the growth of tag vocabulary (Farooq et al., 2007), the rate of tag reuse (Sen et al., 2006) and tag entropy (Chi and Mytkowicz, 2008).

Acknowledgements

This work is partly funded by A*STAR grant 062 130 0057. The authors also wish to Anuksha Jeeha for her assistance in data collection.

References

1. Ahlers, D.: News consumption and the new electronic media. *Harvard International Journal of Press-Politics* 11(1), 29–52 (2006)
2. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 37–45 (1998)
3. Angelo, F.: Citizen-powered journalism fills a void. *Communication World* 25(1), 8–9 (2008)
4. Chi, E.H., Mytkowicz, T.: Understanding the efficiency of social tagging systems using information theory. In: *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pp. 81–88 (2008)
5. Curran, J., Seaton, J.: *Power without Responsibility: The Press, Broadcasting and New Media in Britain*, 6th edn. Routledge, London (2003)
6. Farooq, U., Kannampallil, T.G., Song, Y., Ganoë, C.H., Carroll, J.M., Giles, L.: Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In: *Proceedings of the 2007 international ACM conference on Supporting group work*, pp. 351–360 (2007)
7. Furnas, G.W., Fake, C., von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, Marlow, c., Naaman, M.: Why do tagging systems work. In: *CHI 2006 Extended Abstracts on Human Factors in Computing*, pp. 36–39 (2006)
8. Geisler, G., Burns, S.: Tagging video: conventions and strategies of the YouTube community. In: *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 480–480 (2007)
9. Gillmor, D.: We the media: The rise of citizen journalists. *National Civic Review* 93(3), 58–63 (2004)
10. Goh, D.H., Chua, A., Lee, C.S., Razikin, K.: Resource discovery through social tagging: A classification and content analytic approach. *Online Information Review* (in press)
11. Good, K.: The Rise of the Citizen Journalist. *Feliciter* 52(2), 69–71 (2006)
12. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through Blogspace. Paper presented at the WWW 2004, New York (2004), <http://www.www2004.org/proceedings/docs/1p491.pdf>
13. Johnson, T.J., Kaye, B.K.: Wag the blog: how reliance on traditional media and the Internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly* 81(3), 622–642 (2004)
14. Li, Z., Wang, B., Li, M., Maq, W.Y.: A probabilistic model for retrospective news event detection. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 106–113 (2005)
15. Neo, S.Y., Ran, Y., Goh, H.K., Zheng, Y., Chua, T.S., Li, J.: The use of topic evolution to help users browse and find answers in news video corpus. In: *Proceedings of the 15th international conference on Multimedia*, pp. 198–207 (2007)
16. Nip, J.Y.M.: Exploring the second phase of public journalism. *Journalism Studies* 7(2), 212–236 (2006)

17. Razikin, K., Goh, D.H.-L., Chua, A.Y.K., Lee, C.S.: Can social tags help you find what you want? In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 50–61. Springer, Heidelberg (2008)
18. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging communities, vocabulary, evolution. In: Proceedings of the Conference on Computer Supported Cooperative Work, pp. 181–190 (2006)
19. Thelwall, M., Prabowo, R., Fairclough, R.: Are raw RSS feeds suitable for broad issue scanning? A science concern case study. *Journal of the American Society for Information Science and Technology* 57(12), 1644–1654 (2006)
20. Thurman, N.: Forums for citizen journalists? Adoption of user generated content initiatives by online news media. *New Media & Society* 10(1), 139–157 (2008)
21. Tilley, E., Cokley, J.: Deconstructing the discourse of citizen journalism: Who says what and why it matters. *Preview Pacific Journalism Review* 14(1), 94–114 (2008)
22. Vapnik, V.N.: *The nature of statistical learning theory*. Springer, New York (1995)
23. Wei, C.P., Lee, Y.H.: Event detection from online news documents for supporting environmental scanning. *Decision Support Systems* 36(4), 385–401 (2004)
24. Yang, Y., Carbonell, J.G., Brown, R.D., Pierce, T., Archibald, B.T., Liu, X.: Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems* 14(4), 32–43 (1999)
25. Zelizer, B.: The culture of journalism. In: Curran, J., Gurevitch, M. (eds.) *Mass media and society*, 4th edn., pp. 198–214. Hodder Education, London (2005)
26. Zhao, Q., Liu, T.Y., Bhowmick, S.S., Ma, W.Y.: Event detection from evolution of click-through data. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 484–493 (2006)

A New Mechanism for Job Scheduling in Computational Grid Network Environments

Nandagopal Malarvizhi¹ and V. Rhymend Uthariaraj²

¹ Research Student, Ramanujan Computing Centre, Anna University Chennai, India

² Professor & Director, Ramanujan Computing Centre, Anna University Chennai, India
nmv_94@yahoo.com, rhymend@annauniv.edu

Abstract. Computational grids have the potential for solving large-scale scientific problems using heterogeneous and geographically distributed resources. However, a number of major technical hurdles must overcome before this potential can be realized. One problem that is critical to effective utilization of computational grids is the efficient scheduling of jobs. This work addresses this problem by describing and evaluating a grid scheduling architecture and a job-scheduling algorithm. The architecture is scalable and does not assume control of local site resources. In our algorithm Grid Resource Manager or Grid Scheduler performs resource brokering and job scheduling. The scheduler selects computational resources based on job requirements, job characteristics and information provided by the resources. The main aim of these schedulers is to minimize the total time to release for the individual application. The Time To Release (TTR) includes the processing time of the program, waiting time in the queue, transfer of input and output data to and from the resource. Since grid resources are heterogeneous and distributed over many areas the transmission time is very important criteria. In this paper, an algorithm for minimum time to release is proposed. The proposed scheduling algorithm has been compared with other scheduling schemes such as First Come First Served (FCFS) and Min-Min. These existing algorithms does not consider the transmission time (in time and out time) when scheduling jobs to resources. The proposed algorithm has been verified through the GridSim simulation toolkit and the simulation results confirm that the proposed algorithm produce schedules where the execution time of the application is minimized. The average weighted response times of all submitted jobs decrease up to about 19.79%. The results have been verified using different workloads and Grid configurations.

1 Introduction and Related Work

Grid Computing is concerned with “coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations” [1]. The coordination between multiple administrative domain results heterogeneity in Grid Environment. The resources in Grid Computing include supercomputers, workstations, databases, storages, networks and so on. Resources owned by various administrative organizations are shared under locally defined policies that specify what is shared, who is allowed to access what, and under what conditions [2]. To achieve the promising potentials of computational Grids, an effective and efficient scheduling system is fundamentally important. A grid becomes useful and meaningful when it both encompasses a large set of resources and serves a sizable community [3].

Grid computing has been put into use and achieved great success in the research arena. The most well known example is the SETI@home project [4] initiated by the University of California at Berkeley. Some of the successful examples of grids include Legion [5], NetSolve[6],Nimrod/G[7] and DISCWorld[8]. AppLes[9] is an agent based scheduling system which targets to promote the performance of every individual application. Scheduling systems for traditional distributed environments do not work in Grid environments because the two classes of environments are radically distinct. Scheduling in Grid environments is significantly complicated by the heterogeneous and dynamics nature of Grids. Compared to traditional scheduling systems such as clustering computing, Grid scheduling systems have to take into account diverse characteristics of both various Grid applications and various Grid resources. The different performance goals also place great impacts on the design of scheduling systems.

The different phases of Grid Scheduling Process have been discussed in [10]. In a grid, the abilities of the computing nodes vary, and tasks often arrive dynamically. Because of these, scheduling methods of parallel computing [11] may not be applicable in a grid. It is important to properly assign and schedule tasks to computing nodes [12]. Through good scheduling methods, the system can get better performance, as well as applications can avoid unnecessary delays. Grid scheduling algorithms in different occasions are discussed in paper [13],[14],[15]. But only little work has been done on scheduling jobs according to the real time characteristics of grid resources, which may greatly improve performance of the whole system. [16] provides scheduling algorithm for grid computing that meets the QoS constraints of the jobs. Moreover, the work in [17] proposes a mechanism that schedules jobs to resources based on the processing requirements of the applications being executed. The paper [18] presents a scheduling algorithm that mixes the QoS and the priority of jobs to find the right resource for a job. The main objective of these algorithms is to satisfy the given requirements, rather than optimizing any specific objective such as cost or time. In contrast we optimize the objective while satisfying the constraints.

In order to get more details about the scheduling literature, readers can refer to [19] and [20]. Due to the real-time behavior of grid computing framework, the scheduling problem in grid computing requires separate attention that needs to be solved effectively and efficiently. A survey evaluation of scheduling algorithms is presented in [21],[22]. To the best of our knowledge, the proposed methodology does not match with any of the existing techniques.

In our work we focus only on the Grid Scheduling aspects and try to examine the impact of global grid computing on the scheduling quality for computational jobs. In detail, the effect of the geographical distribution of the resources on the machine utilization and the average response time for the user is analyzed. In this paper, we propose a scheduling algorithm that assigns tasks to machines in a Grid computing system. The scheduling algorithm determines the execution order of the tasks that will be assigned to machines. Since the problem of allocating independent tasks in heterogeneous computational resources is known as NP- complete, an approximation or heuristic algorithm is highly desirable. We assume that the scheduling algorithms are non preemptive, and all are independent jobs.

The remaining part of this paper is organized as follows. A Grid scheduling Architecture is presented in Section 2. In Section 3, the analysis of grid scheduling algorithm for minimizing the execution time of a job is reviewed. The mathematical model

description is presented in Section 4. An experimental setup along with the comparative results is explained in Section 5. Conclusion and future research direction is offered in Section 6.

2 Grid Scheduling Architecture

A grid scheduling architecture is described in Fig 1. The main components used in this architecture are Dispatcher, Grid Scheduler and Load Balancer. Each resource may differ from the rest of the resources with respect to number of processors, cost of processing, speed of processing, internal process scheduling policy, local load factor etc. A Grid Scheduler (GS) receives applications from grid users, selects feasible resources for those applications according to acquired information from the Grid Information Service and finally generates application- to- resource mappings. The function of dispatcher is to interface the various modules in the architecture. Interfacing of modules is possible by means of passing the parameters across various components. Scheduler performs matchmaking of the resources and the clients. Load Balancer re-schedules the results of the scheduler for optimized resource usages.

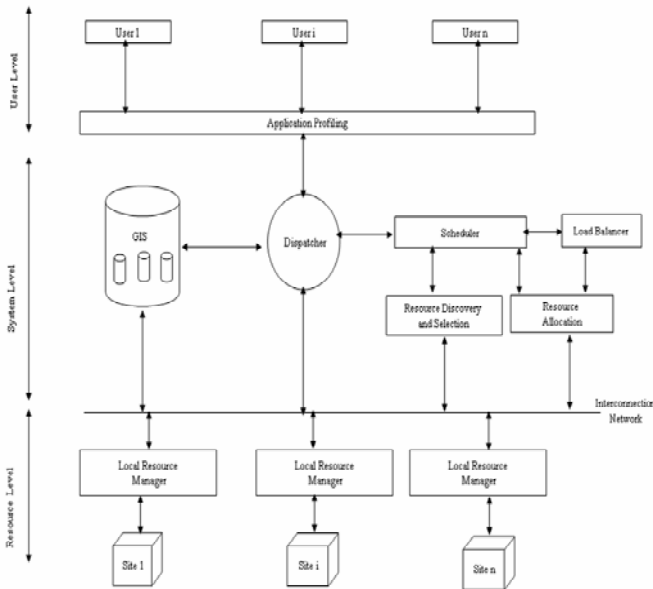


Fig. 1. Grid Scheduling Architecture

3 Grid Scheduling Algorithm

In step 1 of Algorithm 1, the user’s request is processed and split into individual job requests. In step 2, the scheduler discovers the available resources by contacting one or more index servers. By querying each individual resource the specific characteristics of

the resources (both static and dynamic) are identified in step 3. Each resource may provide static information such as the type of operating system installed, network bandwidth, processor speed and storage capacity (including physical and secondary memory) and dynamic information such as processor utilization, physical memory utilization, free secondary memory size, current load, availability and network bandwidth utilization. The actual scheduling process starts from step 4 that is repeated for each job request. In step 5, resources are evaluated according to the requirements in the job request and only the appropriate resources are kept for further processing. Step 6 predicts the performance of each resource by estimating TTR and then maps the job with each resource. Based on TTR, each combination of job and resource are sorted in non-increasing order of TTR. Step 7 removes all those combinations whose time of assignment is more than the deadline constraint and updated the sorted list. Step 8 releases the first combination of job and resource from the updated list. The released job is added to the job list of selected resource in step 9. If it is feasible to assign the released job to the selected resource based on the parameter values then that job is assigned to that resource. All the remaining combinations of the assigned job are removed from the list in step 10. This process is repeated for the next combination of job and resource in the updated list and so on. The algorithm keeps traversing the list until the list is empty as described in step 11.

Algorithm 1

- Step 1:** Create a list of all individual requests by validating the user specification(s).
- Step 2:** Contact one or more index servers to obtain a list of available resources.
- Step 3:** Query each resource to obtain static and dynamic information such as hardware and software characteristics, current status and load etc.
- Step 4:** For each job do the following steps.
- Step 5:** Filter out the resources that do not fulfill the job requirements and also the user is not authorized to use.
- Step 6:** Calculate TTR for each combination of job i and resource j (T_{ij}) and sort them in non-increasing order of T_{ij} .
- Step 7:** Remove all those combinations from the sorted list where T_{ij} is negative or $T_{ij} > D_i$.
- Step 8:** Let α be the job and β' be the resource in the first combination of the updated sorted list and η be the set of jobs that have already been assigned to β' .
- Step 9:** Combine all the jobs in set η with the job α and sort all of them in non-decreasing order of their deadline and ready time. Now try to assign all of these jobs to resource β' in the sorted order. If all these jobs can be assigned without violating the deadline constraint of any of the jobs or the availability constraint of resource β' , then the combination of α and β' is considered to satisfy the time constraint.
- Step 10:** If the time constraint is satisfied for the combination of α and β' , then assign α to β' and remove all the combinations of α from the sorted list.
- Step 11:** If the updated list is empty, then the algorithm terminates, otherwise go to Step 8.

3.1 Scheduling Interactions

Fig.2 presents a UML sequence diagram depicting the Grid Scheduling process. The interactions are between the Scheduler entity, Grid resource Information Service (GIS), Grid statistics, Grid shutdown and Grid report writer entities. Initially the Grid resources register themselves with GIS. A user sends a submit job request to the scheduler for computation. The scheduler then validates the request and interacts with index server for requesting resource information. Then it estimates TTR value and choose the best resource through the mapping procedure as described in Algorithm 1. Finally Submit Job request is sent to the selected resource only. When the job is finished, the resource entity sends back a completion event to scheduler. The report writer entity creates a report for entire simulation by getting the statistics from the Grid Statistics entity.

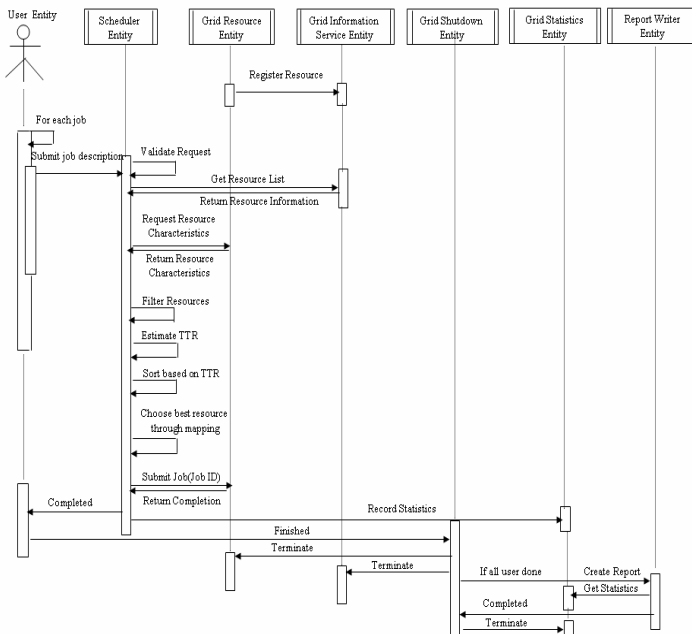


Fig. 2. Sequence Diagram for Grid Scheduling Process

3.2 Estimating Total Time to Release

The estimation of the total time to release requires considering the time to perform the following operations:

- **Transfer in:** Transfer of input files and executable to the resource
- **Waiting Time:** Time spend in resource queue
- **Computation Time:** Actual time to execute the job
- **Transfer out:** Transfer of output files to the user.

3.2.1 Transfer Time Estimation

The estimation of the time for the transfer in and transfer out procedures are based on the actual size of the input file, executable file and the user provided estimates for the size of the output files and also the baudrate (network bandwidth) predictions.

3.2.2 Waiting Time Estimation

To calculate the waiting time of the job in each resource, the scheduler finds a place to fix the job in the resource queue that is based on the processing time of the already existing jobs in the resource queue. The scheduler traverses the queue and finally places the job in suitable position in the queue.

3.2.3 Total Time Estimation

In order to reduce the total time to release of each job the process time and waiting time is calculated by submitting the job to each suitable resource. The process time and the total time for each job i on resource j was estimated as follows:

$$\text{Process Time}_{ij} = \text{in time}_{ij} + \text{CPU time}_{ij} + \text{out time}_{ij}$$

where CPU time_{ij} is calculated by $\text{Length}_i / \text{Speed}_j$

$$\text{Total Time}_{ij} = \text{Process Time}_{ij} + \text{Waiting Time}_{ij}$$

4 Model Description

In this section, we present the mathematical model of the problem. We begin with defining the parameters of the problem and variables used in the model.

4.1 Parameters and Variables

Assume that we have m jobs and n resources. The known parameters about jobs and resources are as follows:

- tt_{ij} Total time required for the i^{th} job if assigned to the j^{th} resource
- dl_i Deadline of the i^{th} job
- ra_j Time from which j^{th} resource is available
- ru_j Time from which j^{th} resource is unavailable
- rt_i Ready time of the i^{th} job

The decision variables used in this model are described as follows:

- $x_{ij} = 1$ if i^{th} job is assigned to resource j ; 0 otherwise $\forall i, j$
- s_i start time of job $i \forall i$
- $y_{ijk} = 1$ if the i^{th} job is the k^{th} assignment on resource j ;
0 otherwise $\forall i$ and $\forall j, k$
- f_{jk} start time of the k^{th} assignment on resource j if resource j has at least k jobs assigned $\forall j, k$

4.2 Time Minimization Model

As discussed earlier, many organizations would like to minimize the time of allocating jobs to resources in their grid. In these models assign jobs to the resources that take less amount of time without considering the price of those resources. In practice, some of the grids do not allow the allocation of jobs to different resources. In other words, all jobs are assigned to the same resource in such grids as a single atomic unit. We present a linear integer-programming (IP) model for the time minimization problem in these grids.

$$\text{Minimize} \quad \sum_i \sum_j tt_{ij} x_{ij} . \tag{1}$$

$$\text{Subject to:} \quad s_i + \sum_j tt_{ij} x_{ij} \leq dl_i \quad \forall i . \tag{2}$$

$$s_i \geq ra_j x_{ij} \quad \forall i . \tag{3}$$

$$s_i + \sum_j tt_{ij} x_{ij} \leq \sum_j ru_j x_{ij} \quad \forall i . \tag{4}$$

$$s_i \geq rt_i \sum_j x_{ij} \quad \forall i . \tag{5}$$

$$\sum_j x_{ij} \leq 1 \quad \forall i . \tag{6}$$

$$\sum_i y_{ijk} \leq 1 \quad \forall j, k . \tag{7}$$

$$\sum_i y_{ijk} \geq \sum_i y_{ij(k+1)} \quad \forall j, k . \tag{8}$$

$$\sum_k y_{ijk} = x_{ij} \quad \forall i, j . \tag{9}$$

$$f_{j(k+1)} \geq f_{jk} + \sum_i tt_{ij} y_{ijk} \quad \forall j, k . \tag{10}$$

$$x_{ij} \in \{0,1\} \quad y_{ijk} \in \{0,1\} \quad f_{jk} \geq 0 \quad s_i \geq 0. \quad (11)$$

The equation (1) represents the time of all assigned jobs. The deadline constraint for each job is modeled in the constraint set (2). The constraint set (3) ensures that the processing of a job on a resource starts only after that resource is available. The constraint set (4) ensures that the processing of a job on a resource ends before that resource becomes unavailable. The constraint set (5) ensures that the processing of a job starts only after it is ready to be processed. The constraint set (6) makes sure that a job is assigned to only one resource. Similarly, the constraint set (7) ensures that there is at most one job at the k th assignment of resource j . The constraint set (8) guarantees that there must be a job at the k th assignment of a resource before the $(k+1)$ th assignment of that resource has a job. The constraint set (9) ensures that if a job is assigned to a resource it must be assigned at some k th assignment; otherwise the job should not be assigned at any assignment of that resource. The constraint set (10) represents the constraints for variables f_{jk} . The constraints set (11) represent the values of the variables x_{ij} and y_{ijk} .

5 Experimental Setup

In this section we analyze the performance of the proposed algorithm. We carried out a simulation-based study. The simulation was based on the grid simulation toolkit software GridSim ToolKit 4.0[23]. In this simulation environment, users can easily add various scheduling policies into the task scheduler. We simulated the grid-computing environment with a scheduler, five users with different time requirements and rates of task creating, and 30 nodes with different compute power.

In our simulation, various entities connected by a virtual network and every two entities' connectivity had an exclusive bandwidth. In order to make comparisons with other results we chose FCFS and Min-Min as the benchmarks because most related works evaluated these algorithms. Since these two algorithms have better performance in general grid job scheduling, always selected to comparative benchmark. But these benchmark algorithms does not consider the transmission time (in time and out time) when scheduling jobs to resources.

The following experimental setup has been used: A job may require one or more machines, which may have different number of CPUs (PE's) with different speeds. We performed the experiments on a PC (Core 2 Processor, 3.20GHz, 1GB RAM) and all of the time in this paper was the simulation time.

5.1 Experimental Results

The benchmark algorithms had a much longer completion time than proposed algorithm. After 1250 jobs, our algorithm took the time 19.79% less than the Min-Min algorithm did.

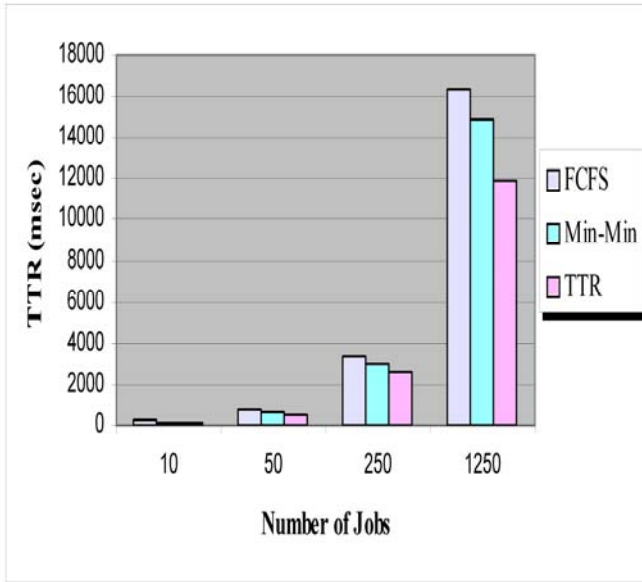


Fig. 3. The results of scheduling algorithm on the grid-computing platform

Table 1. Result Analysis

No.of Tasks	FCFS	Min Min	TTR	Improvement
10	230.24	160.67	157.89	1.73%
50	690.17	585.28	539.78	7.77%
250	3300.35	2986.78	2589.94	13.28%
1250	16435.11	14920.86	11967.93	19.79%

From Fig.3 and Table 1 we can see that the TTR algorithm can increase performance as compared to FCFS and Min-Min case. Note that when the system runs with a small amount of tasks, the improvement of the TTR is slight.

6 Conclusion and Future Work

The success of grid computing will depend on the effective utilization of the system for various computationally intensive jobs. Given a vast number of resources that are available on a Grid, an important problem is the scheduling of jobs on the grid with various objectives. This paper provides managerial insights for all three entities involve in the grid computing- job owners, resource owners and the scheduler. We adopt GridSim Tool Kit to carry out the simulation of our scheduling algorithm to

reduce the total time to release of user jobs and compares with FCFS and Min-Min algorithm. We draw the conclusion that the scheduling algorithm presented in this paper is better than algorithms FCFS and Min-Min.

The main issue in the application-centric scheduling algorithm is the prediction of performance potential of an application on a remote Grid. Now we discuss some of the limitations of this work and present some possible directions for future research. In this work, we assume that there is no precedence constraint among different jobs or different tasks of a job. Usually, the jobs are independent of each other in the grid, but different tasks of a job may have some precedence constraints. Hence, it is an interesting direction for future research. Such dependencies will not only make the problem extremely difficult to solve, but would also require estimating a very large number of parameters. In the future we should also consider some fault tolerant measures to increase the reliability of our algorithm.

References

1. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable virtual Organizations. *International Journal of Super Computer Applications* 15(3) (2001)
2. Foster, I., Iamnitchi, A.: On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. In: Kaashoek, M.F., Stoica, I. (eds.) IPTPS 2003. LNCS, vol. 2735. Springer, Heidelberg (2003)
3. Munter 2005, Grid computing whitepaper (2005)
4. SETI@homeHomePage, <http://setiathome.ssl.berkeley.edu>
5. Legion: A world wide virtual computer. University of Virginia (2007), <http://legion.virginia.edu/>
6. Seymour, K., Yarkhan, A., Agrawal, S., Dongarra, J.: Netsole: Grid enabling scientific computing environments. In: Grandinetti, L. (ed.) *Grid Computing and New Frontiers of High Performance Processing*. Elsevier, Amsterdam (2005)
7. Buyya, R., Abramson, D., Giddy, J.: Nimrod/G: An architecture for a resource management and Scheduling System in a Global Computational Grid (2002)
8. Coddington, P.: DISCWorld, virtual data grids and grid applications (2002), <http://www.gridbus.org/ozgrid/DiscWorld.ppt/>
9. Berman, F., Wolski, R., Casanova, H., Cirne, W., Dail, H., Faerman, M., Figueira, S., Hayes, J., Obertelli, G., Schopf, J., Shao, G., Smallen, S., Spring, N., Su, A., Zagorodnov, D.: Adaptive Computing on the Grid Using AppLeS. *IEEE Transactions on Parallel and Distributed Systems* 14(4) (April 2003)
10. Malarvizhi, N., Rhymend Uthariaraj, V.: A Broker-Based Approach to Resource Discovery and Selection in Grid Environments. In: *IEEE International Conference on Computer and Electrical Engineering*, pp. 322–326 (2008)
11. Feitelson, D.G., Rudolph, R.: Parallel Job Scheduling: Issues and Approaches. In: Feitelson, D.G., Rudolph, L. (eds.) *IPPS-WS 1995 and JSSPP 1995*. LNCS, vol. 949, pp. 1–18. Springer, Heidelberg (1995)
12. Harchol-Balter, M., Leighton, T., Lewin, D.: Resource Discovery in Distributed Networks. In: *18th ACM Symposium on Principles of Distributed Computing* (1999)
13. Benoit, A., Cole, M., Gilmore, S., Hillston, J.: Enhancing the effective utilization of Grid clusters by exploiting on-line performability analysis. In: *2005 IEEE International symposium on Cluster Computing and the Grid*, May 9-12, pp. 317–324 (2005)

14. Berman, F., Wolski, R., Casanova, H.: Adaptive computing on the Grid using APPLES. *IEEE Transactions on Parallel and Distributed Systems* 14(4), 369–382 (2003)
15. Ding, S.-L., Yuan, J.-B., Ju, J.-U.-B.: An algorithm for agent based task scheduling in grid environments. In: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, pp. 2809–2814 (2004)
16. He, X., Sun, X., Laszewski, G.V.: QoS guided min-min heuristic for grid task scheduling. *Journal of Computer Science and Technology* 18, 442–451 (2003)
17. Angulo, D., Foster, I., Liu, C., Yang, L.: Design and evaluation of a resource selection framework for grid applications. In: *Proceedings of IEEE International Symposium on High Performance Distributed Computing* (2002)
18. Min, R., Maheswaran, M.: Scheduling co-reservations with priorities in grid computing system. In: *Proceedings of 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGRID 2002* (2002)
19. Herroelen, W., De Reyck, B., Demeulemeester, E.: Resource constrained project scheduling: A survey of recent developments. *Computers and Operations Research* 25, 279–302 (1998)
20. Mokotoff, E.: Parallel machine scheduling problems: A survey. *Asia- pacific Journal of Operational Research* 18, 193–242 (2001)
21. Jain, R.: *A Survey of Scheduling Methods*. Nokia Research Center (September 1997)
22. Hawa, M.: *Stochastic Evaluation of Fair Scheduling with Applications to Quality-of-Service in Broadband Wireless Access Networks*. PhD dissertation, Univ. of Kansas (August 2003)
23. Buyya, R.: *A Grid simulation toolkit for resource modeling and application scheduling for parallel and distributed computing*, <http://www.buyya.com/gridsim/>

Efficient and Provably Secure Self-certified Signature Scheme

Jianhong Zhang, Hua Chen, and Qin Geng

College of Science, North China University of Technology,
Beijing 100041, P.R. China
jh Zhang@ncut.edu.cn

Abstract. Self-certified signature scheme is a better choice than that in traditional public key system, since it omits key escrow problem of ID-based crypto-system and avoids complex certificate management of traditional PKI. In the work, we first formalized the security models of self-certified signature by classifying its adversary types. Then based on Okamoto's signature idea, we proposed a concrete instance. And we also show that the proposed scheme is provably secure in the random oracle models and the security of the scheme is closely related to a extended security assumption: q -SDH+CDH assumption which is provided by us. Finally, by comparing the proposed scheme with Shao's scheme, we show that the proposed scheme is very efficient in terms of computational costs of generation and verification of a signature. No pairing operator is required to sign a message in our scheme.

1 Introduction

In an ordinary digital signature system, the signer uses his private key to produce signature for the given message, and the verifier uses the signer's public key to verify the validity of the signature. An important problem in all public key systems is how to guarantee a public key is indeed linked to the user who claims to be the legitimate owner of the public key.

In traditional public key infrastructures, a user's public key is authenticated by means of a Certification Authority (CA)'s explicit signature on the public key. And CA issues a digital certificate containing the user's public key and its related information Obviously, the realization of this authentication mechanism is not practical. The requisite explicit certificate and trusted authority are the main concern and the management of certificate is also very difficult. In 1985, Shamir proposed the idea of a cryptosystem based on identification information [1]. In the identity-based public key cryptosystem, a user's identity is used as his public key which is a meaning string. The identity information can be defined as the part of the identification information (Such as IP address, email etc.). If a user's identity is publicly known, the public key can be easily derived and verified. Hence a large public key file is not required. However, private key of each user is chosen by the system authority, not by the user himself, which makes the system place much reliance on the system authority. The main advantage

is that there is no certificate to store and to check. However, the disadvantage is key escrow. Namely, the secret key of the user is known to the authority, a malicious authority can do anything in name of a user.

In 1991, Girault [6] pointed out that most of the public key cryptosystem are vulnerable to the so-called active attacks, such as the adversary attempts to substitute or modify a genuine public key by a fake one during key distribution. In order to avoid such attacks, authenticity of the user's public key must be verified. Girault proposed a self-certified public key system to resolve the problem of public key verification, by combining the characteristics of RSA and discrete logarithms.

In a self-certified public key system, a user chooses his secret key and corresponding public key. He delivers his public key to a trusted party (or CA). The trusted party (CA) combines his public key with his identity to produce a witness. This witness may just be the trusted party's signature on the combination of the user's public key and identity [6], part of a signature [9] or the result of inverting a trapdoor one-way function based on the users public key and identity [12]. The user's public key and identity can be verified implicitly.

To better define the notion of CA's trust, Girault introduced the following three levels of trust in models:

Trust Level 1: The CA knows the users private keys and is capable of impersonating any user without being detected.

Trust Level 2: The CA does not know the users private keys, but it can still impersonate any user by generating a false certificate that may be used without being detected.

Trust Level 3: The CA does not know the users private keys, but it can impersonate the user. However, such impersonation can be detected.

With this definition, in schemes of levels 1 and 2, the CA must be fully trusted by the users, while in schemes of level 3, it is considered as a potentially powerful adversary that may use any means to impersonate the users.

Since 1991, self-certificated public key was introduced, it was further developed [9,13]. Moreover, Saeednia [12] points out that the scheme in [6] allows a cheating trusted party to extract users' private keys which suffers the same problem as IBC. Recently, Tseng [16] proposed a new digital signature scheme based on the self-certified public system. Subsequently, Y.Chang *et.al* [4] proposed another self-certified signature scheme based Tseng's scheme [16]. Unfortunately, Zhang *et.al* [17] showed that Chang *et.al*'s scheme was insecure and proposed an improved scheme. All the above schemes is based on the difficultiy of solving discrete logarithm problem. In 2007, Shao [15] proposed a novel self-certified signature scheme from pairings on elliptic curves. Up to now, previous references do not crispy formalize the model of the scheme as well as the security model. Moreover, previously proposed construction are less efficient in terms of communication and computation.

Our Contribution: In the paper, we first give formal definition on the model of self-certified signature, then we give security model of self-certificated signature

by classifying its adversary types. Finally, we provide a concrete signature scheme based on the idea of Okamoto’s signature, and show that the proposed scheme is provable secure in the random oracle model. The security of the scheme is closely related to a extended security assumption: q -SDH+CDH assumption and the CDH assumption. Finally, Comparison our scheme with Shao’s scheme in terms of computational costs of generation and verification of a signature, the result shows that our scheme is more efficient than Shao’s scheme. And no pairing operator is required to sign a message in our scheme.

2 Preliminaries

We review some fundamental backgrounds required in this paper, namely bilinear pairing, complexity assumption and security model on which our scheme is based.

2.1 Bilinear Maps

In the following, we recall the notations related to bilinear groups [12]. Let $\mathbb{G}_1, \mathbb{G}_2$ be bilinear groups as follows:

1. \mathbb{G}_1 and \mathbb{G}_2 are two cyclic multiplicative groups of prime order p , where possible $\mathbb{G}_1 = \mathbb{G}_2$
2. g_1 is a generator of \mathbb{G}_1 and g_2 is a generator of \mathbb{G}_2 .
3. e is a non-degenerate bilinear map $e : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$, where $|\mathbb{G}_1| = |\mathbb{G}_2| = |\mathbb{G}_T| = p$
 - (a) Bilinear: for all $u, v \in \mathbb{G}_1$ and $a, b \in Z_p$, $e(u^a, v^b) = e(u, v)^{ab}$;
 - (b) Non-degenerate: $e(g, g) \neq 1$;
4. e and the group action in $\mathbb{G}_1, \mathbb{G}_2$ can be computed efficiently.

2.2 Security Assumption

Here we first review the definition of the strong Diffie-Hellman (SDH) assumption introduced in [1], on which the security of our signature is based, and then extend it into a new security assumption, the extended strong Diffie-Hellman assumption, on which the security of a variant of our signature scheme is based.

Strong Diffie-Hellman Assumption: Let $\mathbb{G}_1, \mathbb{G}_2$ be bilinear groups as shown the above section. The q -SDH problem in $(\mathbb{G}_1, \mathbb{G}_2)$ is defined as follows: given $g_1 \in \mathbb{G}_1$, and the $(q + 1)$ -tuple $(g_2, g_2^x, \dots, g_2^{x^q}) \in \mathbb{G}_2^{q+1}$ as input, output a pair $(g_1^{\frac{1}{x+c}}, c)$ where $c \in Z_p$. Algorithm \mathcal{A} has advantage, $Adv_{SDH}(q)$, in solving q -SDH in $(\mathbb{G}_1, \mathbb{G}_2)$ if

$$Adv_{SDH}(q) \rightarrow Pr[\mathcal{A}(g_1, g_2, g_2^x, \dots, g_2^{x^q}) = (g_1^{\frac{1}{x+c}}, c)]$$

Where the probability is taken over the random choice of $g_2 \in \mathbb{G}_2$, $x \in Z_p^*$, and the coin tosses of \mathcal{A} .

Definition 1. Adversary $\mathcal{A}(t, \epsilon)$ —breaks the q -SDH problem if \mathcal{A} runs in time at most t and $Adv_{SDH}(q)$ is at least ϵ . The (q, t, ϵ) -SDH assumption holds if no adversary $\mathcal{A}(t, \epsilon)$ —breaks the q -SDH problem.

Definition 2 (Computational Diffie-Hellman (CDH) Assumption). Let \mathcal{G} be a CDH parameter generator. We say an algorithm \mathcal{A} has advantage $\epsilon(k)$ in solving the CDH problem for \mathbb{G}_1 if for a sufficiently large k ,

$$Adv_{\mathcal{G}, \mathcal{A}}(t) = Pr[\mathcal{A}(p, \mathbb{G}_1, g^x, g^y) = g^{xy} \mid (p, \mathbb{G}_1) \leftarrow \mathcal{G}^k, P \leftarrow \mathbb{G}_1, x, y \leftarrow \mathcal{Z}_p]$$

We say that \mathbb{G}_1 satisfies the CDH assumption if for any randomized polynomial time in t algorithm \mathcal{A} we have the $Adv_{\mathbb{G}_1, \mathcal{A}}(t)$ is negligible function.

Definition 3 (Inverse Computational Diffie-Hellman Problem (Inv-CDHP)). Let \mathcal{G} be a CDH parameter generator. We say an algorithm \mathcal{A} has advantage $\epsilon(k)$ in solving the Inv-CDH problem for \mathbb{G}_1 if for a sufficiently large k ,

$$Adv_{\mathcal{G}, \mathcal{A}}(t) = Pr[\mathcal{A}(p, \mathbb{G}_1, g, g^x) = g^{x^{-1}} \mid (p, \mathbb{G}_1) \leftarrow \mathcal{G}^k, g \leftarrow \mathbb{G}_1, x \leftarrow \mathcal{Z}_p]$$

We say that \mathbb{G}_1 satisfies the Inv-CDH assumption if for any randomized polynomial time in t algorithm \mathcal{A} we have the $Adv_{\mathbb{G}_1, \mathcal{A}}(t)$ is negligible function.

The following theorem relates these problems [5][14].

Theorem 1. The CDH problem and Inv-CDH problem are polynomial time equivalent.

In the following, we present a novel security assumption: q -SDH+CDH Assumption, by combining the strong Diffie-Hellman with the computational Diffie-Hellman.

q -SDH+CDH Assumption: The q -SDH+CDH problem in \mathbb{G}_1 is defined as follows: give $q + 1$ -tuple $(g_1, g_1^\alpha, \dots, g_1^{\alpha^q})$ and a random pair (g_1, g_1^r) of group \mathbb{G}_1 as inputs, output $(\rho \leftarrow (g_1)^{\frac{r}{(\alpha+c)}}, c)$, where g_1 is a generator of group \mathbb{G}_1 , $c, r \in_R \mathcal{Z}_p^*$. Note that α and r are unknown numbers. Algorithm \mathcal{A} has advantage, $Adv_{SDH}(q)$, in solving q -SDH+CDH in \mathbb{G}_1 if

$$Adv_{q-SDH+CDH}(q) \leftarrow Pr[\mathcal{A}(g_1, g_1^\alpha, \dots, g_1^{\alpha^q}, g_1^r) = (g_1^{\frac{r}{(\alpha+c)}}, c)]$$

Where the probability is taken over the random choices of $g_1 \in \mathbb{G}_1$, $\alpha, r \in \mathcal{Z}_p^*$, and the coin tosses of \mathcal{A} .

Definition 4. Adversary $\mathcal{A}(t, \epsilon)$ —breaks the q -SDH+CDH problem if \mathcal{A} runs in time at most t and $Adv_{q-SDH+CDH}(q)$ is at least ϵ . The (q, t, ϵ) -SDH+CDH assumption holds if no adversary $\mathcal{A}(t, \epsilon)$ -breaks the q -SDH+CDH problem.

According to the above definition, we know that the novel security assumption: **q -SDH+CDH Assumption**, is not easier than either q -SDH assumption. Because if the q -SDH+CDH Assumption can be solved in polynomial time, then we set $r = 1$, the above q -SDH+CDH assumption is converted into q -SDH assumption. It denotes that the q -SDH assumption can also be solved in polynomial time. Thus, we can obtain the lemma.

Lemma 1. *If the q -SDH+CDH Assumption can be solved in the polynomial time with non-negligible probability, then the q -SDH assumption is solvable.*

Proof. Suppose that q -SDH+CDH assumption is solved. Given an instance of q -SDH problem $(g_1, g^\alpha, \dots, g_1^{\alpha^q})$, we randomly $r \in Z_p$ to compute $B = g^r$. Take $(g_1, g^\alpha, \dots, g_1^{\alpha^q}, (g_1, g^r))$ as inputs of q -SDH+CDH problem, output $(g_1^{\frac{r}{\alpha+c}}, c)$. Then we can compute $(g_1^{\frac{1}{\alpha+c}}, c)$ by the known r . Thus the q -SDH problem can be solved.

Because we know that original proposition is equivalent to converse-negative proposition.

Lemma 2. *If the q -SDH assumption is hard to solve in polynomial time, then the q -SDH+CDH assumption is also hard to solve in the polynomial time.*

Lemma 3. *If the CDH problem is solvable in the polynomial time, then q -SDH+CDH problem is also solvable.*

Proof. Since the CDH problem is solvable, given (g, g^a, g^b) , then g^{ab} is able to be obtained. Given a q -SDH+CDH problem instance $(g, g^\alpha, g^{\alpha^2}, \dots, g^{\alpha^q}, g, g^r)$, randomly choose $c \in Z_p$ to compute $(g, g^{\alpha+c}, g^{(\alpha+c)^2}, \dots, g^{(\alpha+c)^q}, (g, g^r))$ by the above q -SDH+CDH problem instance.

Let $g_0 = g^{(\alpha+c)^q}$, then the inverted sequence of $(g, g^{\alpha+c}, g^{(\alpha+c)^2}, \dots, g^{(\alpha+c)^q})$ can be expressed as $(g_0, g_{01} = g_0^{\frac{1}{\alpha+c}} = g^{(\alpha+c)^{q-1}}, \dots, g_{0q} = g_0^{(\frac{1}{\alpha+c})^q} = g)$. When q is an odd number, we set $\rho = \frac{q-1}{2}$. When q is an even number, we set $\rho = \frac{q}{2}$.

1. if q is an odd number, we obtain $g_{0\rho} = g_0^{(\frac{1}{\alpha+c})^\rho}$ and $g_{0\rho+2} = g_0^{(\frac{1}{\alpha+c})^{\rho+2}}$.

Because the CDH problem is solvable, then we can obtain $g_0^{(\frac{1}{\alpha+c})^{2\rho+2}} = g_0^{(\frac{1}{\alpha+c})^{q+1}}$ by $(g_{0\rho}, g_{0\rho+2})$. Note that when q is an odd number, $2\rho+2 = q+1$.

$$g_0^{(\frac{1}{\alpha+c})^{2\rho+2}} = g_0^{(\frac{1}{\alpha+c})^{q+1}} = (g^{(\alpha+c)^q})^{(\frac{1}{\alpha+c})^{q+1}} = g^{\frac{1}{\alpha+c}}$$

2. if q is an even number, we obtain $g_{0\rho}$ and $g_{0\rho+1}$. Because the CDH problem is solvable, then we can obtain $g_0^{(\frac{1}{\alpha+c})^{2\rho+1}} = g_0^{(\frac{1}{\alpha+c})^{q+1}}$ by $(g_{0\rho}, g_{0\rho+1})$. Note that when q is an even number, $2\rho+1 = q+1$.

$$g_0^{(\frac{1}{\alpha+c})^{2\rho+1}} = g_0^{(\frac{1}{\alpha+c})^{q+1}} = (g^{(\alpha+c)^q})^{(\frac{1}{\alpha+c})^{q+1}} = g^{\frac{1}{\alpha+c}}$$

Given (g, g^r) , since the the CDH problem is solvable, then we can obtain $g^{\frac{r}{\alpha+c}}$. This denotes that q -SDH+CDH problem is also solvable. □

By the above discussion, we can obtain

Theorem 2. *The q -SDH problem \leq q -SDH+CDH problem \leq the the CDH problem.*

where symbol \leq denotes that the problem A is easier than the problem B to be solved.

3 Definition and Model of Self-certified Signature

3.1 Definition

A self-certificated signature scheme consist of four phases: **Setup**, **Key-Extract**, **Sign** and **Verify** as follows:

- Setup: This algorithm takes as input a security parameter 1^k and returns the master secret key msk of Certificate Authority (CA) and master public key P_{CA} . It also outputs a parameter **Param** which is shared in the system.
- Key-Extract: This algorithm takes as input the master secret msk , the master public key mpk , system parameter **Param** and a user's partial public key (ID, P) , where P denotes the partial public key of the user and is produced by the user's partial private key x_{ID} . It outputs a partial private key d_{ID} .
- Sign: This algorithm takes as input the master public key P_{CA} , system parameter **Param**, an identity ID , this identity's private key (x_{ID}, d_{ID}) , this identity's public key P and a message M . It outputs a signature δ .
- Verify: This algorithm takes as input the master public key P_{CA} , system parameter **Param**, an identity ID , this identity's public key P and a message/signature (M, δ) . It outputs *ture* if the signature is correct, or false otherwise.

3.2 Security Model of Self-certified Signature Scheme

A security definition of digital signature scheme was given by Goldwasser *et.al* [7]. A secure digital signature scheme must be against adaptive chosen message attacks, where an adversary is allowed to ask the signer to sign any message of its choice adaptively, i.e. he can adaptively choose it queries according to previous answers.

Obviously, our self-certified signature should satisfy security notion of digital signature, existential unforgeability against adaptive chosen message attack. According to the adversary's attack power, we divide the potential attackers into three kinds:

1. Type I: this type adversary \mathcal{A}_I is an ordinary adversary except a user and CA, it only has the public key of a user and CA.
2. Type II: this type adversary \mathcal{A}_{II} is a dishonest user, it has the partial secret key of a user besides the public key of a user and CA.
3. Type III: this type adversary \mathcal{A}_{III} is a malicious CA, it possesses the master secret of CA besides the public key of a user and CA.

One can find Type *I* and Type *II* are two powerful attacks. If a self-certificated signature scheme is secure against Type *II* adversary and Type *III* adversary, then the scheme is also secure against Type *I* adversary. The above classification makes the security model of self-certified signature clearer, therefore, we use this classification to redefine security modle of self-certificated signature.

Existential unforgeability against adaptive \mathcal{A}_{II} adversary

Roughly speaking, the existential unforgeability of a self-certified signature under a Type II attacker requires that it is difficult for a dishonest user to forge a valid signature if he does not obtain CA's witness on his public key. It is defined by using the following game between the challenger \mathcal{C} and a type II adversary \mathcal{A}_{II} .

- Setup: The challenger \mathcal{C} takes a security parameter k and runs ParaGen to obtain system parameters and a random public key P_{CA} of certificate authority.

Phase 1: The type II adversary issues the following series queries.

- Key-Extract queries: \mathcal{A}_{II} can request Key-Extract Oracle on (P_i, ID_i) , the challenger responds with the partial private key d_i corresponding to the public key (P_{CA}, ID_i, P_i) , where (ID_i, P_i) is the identity and his chosen public key of the adversary \mathcal{A}_{II}
- Sign queries: \mathcal{A}_{II} can request Signing Oracle on (P_i, ID_i, M_i) , the challenger outputs a signature δ such that $\text{true} \leftarrow \text{Verify}(M_i, \delta, Params, ID_i, P_i, P_{CA})$.
- Challenge: Once the Type II adversary decides that Phase 1 is over, it chooses its partial public key (ID^*, P^*) as his challenged public key.

Phase 2: The adversary issues the same queries as ones of phase 1.

- Output. Finally, Type II adversary outputs a new signature δ^* for a message M^* such that
 1. (ID^*, P^*) has not been requested as one of Key-Extract queries.
 2. (ID^*, P^*, M^*) has not been requested as one of Sign queries.
 3. δ^* is a valid signature under the public key (ID^*, P^*) .

Definition 5. We say a type II adversary \mathcal{A}_{II} can (t, q_e, q_s, ϵ) break a self-certificated signature scheme, if \mathcal{A}_{II} runs in time at most t , \mathcal{A}_{II} makes at most q_e key extract queries and at most q_s signing queries and the success probability is at least ϵ .

Existential unforgeability against adaptive \mathcal{A}_{III} adversary

as for \mathcal{A}_{III} adversary, the existential unforgeability of a self-certified signature under a Type III attacker requires that it is difficult for a malicious CA to forge a valid signature without a own private key of the user. It is defined by using the following game between the challenger \mathcal{C} and a type III adversary \mathcal{A}_{III} .

- Setup: The challenger \mathcal{C} takes a security parameter k and runs ParaGen to obtain system parameters and a random public-secret key pair (P_{CA}, s) of certificate authority, then send public parameters and public-secret key pair (P_{CA}, s) of of certificate authority to adversary III.
- Sign queries: \mathcal{A}_{III} can request Signing Oracle on (P_i, ID_i, M_i) , the challenger outputs a signature δ such that $\text{true} \leftarrow \text{Verify}(M_i, \delta, Params, ID_i, P_i, P_{CA})$.
- Output. Finally, Type III adversary outputs a new signature δ^* for a message M^* such that
 1. (ID^*, P^*, M^*) has not been requested as one of Sign queries.
 2. δ^* is a valid signature under the public key (ID^*, P^*) .

Definition 6. We say a type III adversary \mathcal{A}_{III} can (t, q_s, ϵ) break a self-certificated signature scheme, if \mathcal{A}_{III} runs in time at most t , \mathcal{A}_{III} , makes at most at most q_s signing queries and the success probability is at least ϵ .

4 Our Self-certified Signature Scheme

In the section, we will give our self-certified signature scheme. The main idea of the scheme is based on Okamoto’s signature [8] and the above q -SDH+CDH assumption. The scheme consists of four algorithms. The detail description is as follows:

[Setup]: Let \mathbb{G}_1 and \mathbb{G}_2 be two groups of prime order p and $g \in \mathbb{G}_1$ is a generator of group \mathbb{G}_1 , where p is a large prime. $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ is a bilinear map. Randomly choose $\alpha \in Z_p$ as the private key of CA (certificate authority) and compute $P_{CA} = g^\alpha$ as the public key of CA. $h_1, h_2 \in \mathbb{G}_1$ were randomly chosen and $H(\cdot), H_1(\cdot), H_0(\cdot)$ are three one-way hash functions which satisfy $H : \mathbb{G}_1^2 \times \{0, 1\}^* \rightarrow Z_p, H_1 : \mathbb{G}_1 \times \{0, 1\}^* \rightarrow Z_p$ and $H_0 : \mathbb{G}_1^2 \times \{0, 1\}^* \rightarrow \mathbb{G}_1$. The public parameters are $(g, h_1, h_2, e, H(\cdot), H_0(\cdot), H_1(\cdot), P_{CA})$.

[Key Generation]: When a user U_i with his identity ID_i intends to join this system, he first generates his public key. Therefore, he randomly selects $x_i \in Z_p$ as his private key and computes $P_i (= g^{x_i})$ as his partial public key. (Note that when P_i is the identity element of \mathbb{G}_1 , we need to reselect a $x_i \in Z_p$ to compute public key). Then the user U_i sends (ID_i, P_i) to certificate authority CA. After receiving (ID_i, P_i) , CA computes $h_0 = H_0(P_{CA}, ID_i, P_i)$ and $d_i = (h_0)^{\frac{1}{\alpha - H_1(P_i, ID_i)}}$, and sends d_i to the user. The user U_i can verify whether it holds by the equation $e(d_i, P_{CA} g^{-H_1(P_i, ID_i)}) = e(h_0, g)$, then the private key of the user U_i is (x_i, d_i) .

[Sign]: To sign a message M , the user with identity ID_i randomly chooses $s \in Z_p$ to compute $\delta = (\delta_1, \delta_2)$, where

$$\delta_1 = d_i^{x_i} \cdot (h_2 h_1^{H_1(ID_i, P_i)})^{sm}, \delta_2 = (P_{CA} g^{H_1(ID_i, P_i)})^s$$

and $m = H(\delta_2, M, P_i)$. The resultant signature on message M is (δ_1, δ_2)

[Verify]: Upon receiving the signature $\delta = (\delta_1, \delta_2)$ on message M , a verifier computes as follows:

1. compute $m = H(\delta_2, M, P_i)$ and $h_0 = H_0(P_{CA}, ID_i, P_i)$;
2. verify

$$e(P_{CA} g^{-H_1(ID_i, P_i)}, \delta_1) = e(h_0, P_i) e(\delta_2, (h_2 h_1^{H_1(ID_i, P_i)})^m)$$

5 Security Analysis

In the section, we show that the proposed scheme is correct and provably secure in the security model of our definition. If a signature $\delta = (\delta_1, \delta_2)$ is valid, then the signature δ must pass the verification equation. Because

$$\begin{aligned}
 & e(P_{CA}g^{-H_1(ID_i, P_i)}, \delta_1) \\
 &= e(P_{CA}g^{-H_1(ID_i, P_i)}, d_i^{x_i} \cdot (h_2h_1^{H_1(ID_i, P_i)})^{sm}) \\
 &= e(P_i, h_0)e(\delta_2, (h_2h_1^{H_1(ID_i, P_i)})^m)
 \end{aligned}$$

where $m = H(\delta_2, M, P_i)$ and $h_0 = H_0(P_{CA}, P_i, ID_i)$.

Theorem 3. *If there exists a type II adversary \mathcal{A}_{II} can (t, ϵ, q_e, q_s) -breaks the proposed scheme, then there exists another algorithm \mathcal{B} which can make use of \mathcal{A}_{II} to solve an instance of the q -SDH+CDH problem in \mathbb{G}_1 with the non-negligible probability.*

Proof. Assume \mathcal{A}_{II} is an adversary that (t, ϵ, q_e, q_s) -forges the proposed signature scheme. We will then construct algorithm \mathcal{B} that break the q -SDH+CDH assumption with (t', ϵ') . Hereafter, we often use $q = q_s + 1$. Let us recall q -SDH+CDH assumption: given a tuple $(g, g^\alpha, g^{\alpha^2}, \dots, g^{\alpha^q})$ and a random pair (g, g^r) , its goal is to compute $(c, g^{\frac{r}{\alpha+c}})$, where $c, r \in_R Z_p$ and r is unknown.

Setup: \mathcal{B} randomly chooses u, v, r_i ($i=1, \dots, q-1$) from Z_p . Let $f(x)$ be a polynomial of variable x such that $f(x) = \prod_{i=1}^{q-1} (x+r_i) = \sum_{i=0}^{q-1} \beta_i x^i$. Obviously, \mathcal{B} can efficiently compute $\beta_i \in Z_p$ ($i = 0, 1, \dots, q-1$) from all r_i . \mathcal{B} computes

$$g = \prod_{i=0}^{q-1} (g^{\alpha^i})^{\beta_i}, P_{CA} = \prod_{i=0}^{q-1} (g^{\alpha^{i+1}})^{\beta_i}$$

and

$$h_1 = g^u, h_2 = P_{CA}^u$$

Choose three one-way hash functions $H_0(\cdot) : \mathbb{G}_1^3 \rightarrow \mathbb{G}_1, H(\cdot) : \mathbb{G}_1^3 \rightarrow Z_p$ and $H_1(\cdot) : \mathbb{G}_1^2 \rightarrow Z_p$. Finally, \mathcal{B} publishes system parameters $(g, P_{CA}, H, H_0, H_1, h_1, h_2)$.

Phase 1:

H_1 - Oracle: When the adversary \mathcal{A}_{II} makes a query on (ID_i, P_i) to H_1 -oracle, if (ID_i, P_i) exists in H_1 -list, then H_{1_i} is returned. Otherwise, \mathcal{B} randomly chooses $i_1 \in \{1, \dots, q_1\}$ to return r_{i_1} to \mathcal{A} and adds $(ID_i, P_i, H_{1_i} = r_{i_1})$ in the H_1 -list.

H_0 - Oracle: When the adversary \mathcal{A}_{II} makes a query on (P_{CA}, ID_i, P_i) to H_0 -oracle, if (ID_i, P_i, b_i) exists in H_0 -list, then g^{b_i} is returned. Otherwise, \mathcal{B} randomly chooses $b_i \in Z_p$ to return g^{b_i} to \mathcal{A} and adds (ID_i, P_i, b_i) in the H_0 -list.

H - Oracle: When the adversary \mathcal{A}_{II} makes a query on $(\delta_{2_i}, ID_i, P_i)$ to H -oracle, if $(\delta_{2_i}, ID_i, P_i, m_i)$ exists in H -list, then m_i is returned. Otherwise, \mathcal{B} randomly chooses $m_i \in Z_p$ to return m_i to \mathcal{A} and adds $(\delta_{2_i}, ID_i, P_i, m_i)$ in the H -list.

Key-Extract Oracle: When the adversary \mathcal{A}_{II} makes a query on (ID_i, P_i) to key-extract oracle, \mathcal{B} checks whether (ID_i, P_i) exists in the H_1 -list and H_0 -list. If (ID_i, P_i) exists in the H_0 -list and H_1 -list, then \mathcal{B} computes

$$d_i = (g^{b_i})^{g_i(\alpha)} = h_0^{g_i(\alpha)} = h_0^{\frac{1}{\alpha - H_1(ID_i, P_i)}} \quad (1)$$

where $g_i(\alpha) = \prod_{i=1, i \neq i_1}^{q-1} (\alpha + r_i) = \frac{f(\alpha)}{\alpha + r_{i_1}}$.

If (ID_i, P_i) exists in the H_0 -list but doesn't exist in the H_1 -list, then \mathcal{B} randomly chooses $i_1 \in \{1, \dots, q_1\}$ and adds $(ID_i, P_i, H_{1_{i_1}} = r_{i_1})$ in the H_1 -list. Finally, \mathcal{B} returns d_i to \mathcal{A}_{II} according to the above equation (1).

If (ID_i, P_i) exists in the H_1 -list but doesn't exist in the H_0 -list, then \mathcal{B} randomly chooses $b_i \in Z_p$ and adds (ID_i, P_i, b_i) in the H_0 -list. Finally, \mathcal{B} returns d_i to \mathcal{A}_{II} according to the above equation (1).

Otherwise, \mathcal{B} randomly chooses $b_i \in Z_p$ and $i_1 \in \{1, \dots, q_1\}$, then add (ID_i, P_i, b_i) and $(ID_i, P_i, H_{1_{i_1}} = r_{i_1})$ in the H_0 -list and H_1 -list, respectively. Finally, \mathcal{B} returns d_i to \mathcal{A} according to the above equation (1).

Signing Oracle: Upon receiving a query to the signing oracle with (ID_i, P_i, M) , \mathcal{B} first checks whether (ID_i, P_i) exists in the H_0 -list or not. If it doesn't exist, \mathcal{B} randomly chooses $b_i \in Z_p$ and adds (ID_i, P_i, b_i) in the H_0 -list. Subsequently, \mathcal{B} simulates the reply to \mathcal{A}_{II} as follows:

1. randomly choose $k \in Z_p$ to compute $\delta_2 = P_{i_1}^{-\frac{b_i}{2H_1(ID_i, P_i)u \cdot k}}$
2. set $\delta_1 = P_{i_1}^{-\frac{b_i}{2H_1(ID_i, P_i)}}$
3. if (M, P_i, δ_2) had been queried in the H -list, then \mathcal{B} aborts it. Otherwise, \mathcal{B} sets $m = H(M, P_i, \delta_2) = k$ and adds it in the H -list.

And return (δ_1, δ_2) to the adversary \mathcal{A}_{II} . Clearly this is a valid signature under the public key (P_i, ID_i) and the distribution is exactly the same as that given the signing oracle.

In the following, we show that the returned signature is valid. Because

$$\begin{aligned} e(g, \delta_2^{-\frac{2H_1(ID_i, P_i)mu}{b_i}}) &= e(g, P_i) \\ &\Downarrow \\ e(g, \delta_2^{-2H_1(ID_i, P_i)mu}) &= e(g^{b_i}, P_i) = e(h_0, P_i) \\ &\Downarrow \\ e(P_{CA} \cdot (g)^{H_1(ID_i, P_i)}, \delta_1) &= e(h_0, P_i)e(\delta_2, h_2 \cdot h_1^{H_1(ID_i, P_i)})^m \end{aligned}$$

Challenge phase: Once the adversary \mathcal{A}_{II} decides that Phase1 is over, it randomly chooses $P^* \in \mathbb{G}_1$ as the challenged public key and sends it to \mathcal{B} .

Phase 2: The adversary \mathcal{A}_{II} issues the same query as the above queries .

Output: When \mathcal{A}_{II} outputs a forgery signature (δ_1^*, δ_2^*) on message M^* under the public key (P^*, ID^*) . If (δ_1^*, δ_2^*) is a valid signature on message M^* , and \mathcal{A}_{II} has neither made Key-Extract Oracle on (ID^*, P^*) nor a signing query on (ID^*, P^*, M^*) . \mathcal{B} checks whether (ID^*, P^*) exists in the H_0 -list. If it doesn't exist, then it aborts it. Otherwise, \mathcal{B} retrieves b^* from the H_0 -list and computes as follows:

$$\begin{aligned}
\delta_1^* &= (d^*)^{x^*} (h_2 h_1^{H_1(ID^*, P^*)})^{s^* m^*} \\
&= (h_0^*)^{\frac{x^*}{\alpha - H_1(ID^*, P^*)}} (h_2 h_1^{H_1(ID^*, P^*)})^{s^* m^*} \\
&= (P^*)^{\frac{h^*}{\alpha - H_1(ID^*, P^*)}} ((P_{CAg})^{u H_1(ID^*, P^*)})^{s^* m^*} \\
\delta_2^* &= (P_{CAg})^{H_1(ID^*, P^*)} s^*
\end{aligned}$$

Thus, \mathcal{B} can compute $(P^*)^{\frac{1}{\alpha - H_1(ID^*, P^*)}} = (\frac{\delta_1^*}{(\delta_2^*)^{um^*}})^{(b^*)^{-1}}$

Theorem 4. *If there exists a type III adversary \mathcal{A}_{III} can (t, ϵ, q_s) -break the above proposed scheme, then there exists an algorithm \mathcal{B} which can use the adversary \mathcal{A}_{III} to solve an instance of the CDH problem in \mathbb{G}_1 with the non-negligible probability.*

Proof. Due to the limited space, we omit it here.

Efficiency: Here, we will analyze efficiency of our proposed scheme by comparing our scheme with Shao's scheme [15] which was newly proposed in literature. For convenient comparison, we instantiate pairing-based scheme using Barreto-Naehrig curves [3] with 160-bit point representation. In elliptic curve, pairing operator and scalar multiplication operator are more expensive. Thus, we only consider the two operators. Let P_e denote pairing operations in group \mathbb{G}_1 , P_m be scalar multiplication operation in group \mathbb{G}_1 . According to the above table, we know that no pairing operator is included in the signing phase of our scheme. Thus, our scheme is efficient in terms of computation costs of signing and verification.

Table 1. Comparison of our proposed scheme with Shao's scheme

Scheme	Size	Verification	Generations
Shaoet.al scheme	$ p + \mathbb{G}_1 $	$2P_e + 2P_m$	$4P_m + P_e$
Our scheme	$2 \mathbb{G}_1 $	$3P_m + 3P_e$	$3P_m$

6 Conclusions

Self-certified cryptography aims at combining the advantages of identity-based and public key cryptography. Certificate is implicitly verified in self-certified public key system, it omits key escrow problem of ID-based cryptosystem and avoids complex certificate management of traditional PKI. Thus, self-certificated public key is a cryptosystem of promise. In the work, we first give format definition of security model for self-certificate signature scheme by classifying its adversary types. Then a novel signature scheme is proposed, and we also show that our scheme is secure in the random oracle model.

Acknowledgements

This work is supported by Natural Science Foundation of China (NO:60703044), the New Star Plan Project of Beijing Science and Technology (NO:2007B001),

the PHR, Program for New Century Excellent Talents in University(NCET-06-188), The Beijing Natural Science Foundation Programm and Scientific Research Key Program of Beijing Municipal Commission of Education (NO:KZ2008 10009005) and 973 Program (No:2007CB310700).

References

1. Boneh, D., Lynn, B., Shacham, H.: Short Signatures from the Weil Pairing. *Journal of Cryptology* 17(4), 297–319
2. Boneh, D., Boyen, X.: Short signatures without random oracles. In: Cachin, C., Camenisch, J.L. (eds.) *EUROCRYPT 2004*. LNCS, vol. 3027, pp. 56–73. Springer, Heidelberg (2004)
3. Boneh, D., Boyen, X.: Short signatures without random oracles. In: Cachin, C., Camenisch, J.L. (eds.) *EUROCRYPT 2004*. LNCS, vol. 3027, pp. 56–73. Springer, Heidelberg (2004)
4. Chang, Y., Chang, C., Huang, H.: Digital signature with message recovery using self-certificated public keys without trustworthy system authority. *Applied Mathematics and Computation* 161, 211–227 (2005)
5. Maurer, U.M.: Towards the equivalence of breaking the diffie-hellman protocol and computing discrete logarithms. In: Desmedt, Y.G. (ed.) *CRYPTO 1994*. LNCS, vol. 839, pp. 271–281. Springer, Heidelberg (1994)
6. Girault, M.: Self-certified public keys. In: Davies, D.W. (ed.) *EUROCRYPT 1991*. LNCS, vol. 547, pp. 490–497. Springer, Heidelberg (1991)
7. Goldwasser, S., Micali, S., Rivest, R.: A digital signature secure against adaptive chosen-message attacks. *SIAM J. Comput.* 17(2), 281–308
8. Okamoto, T.: Efficient blind and partially blind signatures without random oracles. In: Halevi, S., Rabin, T. (eds.) *TCC 2006*. LNCS, vol. 3876, pp. 80–99. Springer, Heidelberg (2006)
9. Peterson, H., Horster, P.: Self-certified keys- concepts and application. In: 3rd int. conference on communciation and multimedia security, pp. 102–116. Chapman & Hall, Sydney (1997)
10. Pointcheval, D., Stern, J.: Security proofs for signature schemes. In: Maurer, U.M. (ed.) *EUROCRYPT 1996*. LNCS, vol. 1070, pp. 387–398. Springer, Heidelberg (1996)
11. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakely, G.R., Chaum, D. (eds.) *CRYPTO 1984*. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
12. Saeednia, S.: A note on Girault’s self-certified mode. *Information Processing Letters* 86(6), 323–327 (2003)
13. Saeednia, S.: Identity-based and self-certificated key-exchange protocols. In: Mu, Y., Pieprzyk, J.P., Varadharajan, V. (eds.) *ACISP 1997*. LNCS, vol. 1270, pp. 47–53. Springer, Heidelberg (1997)
14. Sadeghi, A.-R., Steiner, M.: Assumptions related to discrete logarithms: Why subtleties make a real difference. In: Pfitzmann, B. (ed.) *EUROCRYPT 2001*. LNCS, vol. 2045, pp. 243–260. Springer, Heidelberg (2001)
15. Shao, Z.: Self-certified signature scheme from pairings. *The Journal of Systems and Software* 80, 388–395 (2007)
16. Tseng, Y., Jan, J., Chien, H.: Digital signature with message recovery using self-certificated public keys and its variants. *Applied Mathematics and Computation* 136, 203–214 (2003)
17. Zhang, J., Zou, W., Chen, D., Wang, Y.: On the Security of a Digital Signature with Message Recovery Using Self-certified Public Key. *Informatica* 29(3), 343–346

A Reversible Watermarking Scheme for 3D Meshes

Dan Wu¹ and Guozhao Wang²

¹ Department of Mathematics, China Jiliang University, Hangzhou, 310018, P.R. China
linlin_52@yahoo.com.cn

² Department of Mathematics, Zhejiang University, Hangzhou, 310027, P.R. China
wanggz@math.zju.edu.cn

Abstract. Reversible watermarking is suitable for hiding data in 3D meshes, because it can remove the embedding-induced distortions after extracting the hidden bits. This paper proposes a novel reversible watermarking scheme based on the difference expansion and the difference shifting. Since the 3D mesh includes a sequence of vertices, we embed the watermark into the vertex coordinates by modifying the differences between the adjacent vertex coordinates. The scheme can keep the mesh topology unchanged, and need not record the extra information. The experimental results show that the proposed scheme achieves good performance of the imperceptibility and high-capacity data embedding.

1 Introduction

In the past decades, digital watermarking has been widely used in areas such as ownership protection, content authentication, distribution tracking, and broadcast monitoring. In most cases, the cover media will experience some permanent distortion due to watermarking and cannot be inverted back to the original media. However, in some applications, especially in the medical, military and legal domain, even the imperceptible distortion introduced in the watermarking process is unacceptable. Under these circumstances, reversible watermarking is desired, which not only extracts the watermark, but also perfectly reconstructs the original host signal from the watermarked work. Many reversible watermarking algorithms have been proposed recently [1-6]. Tian [7] proposed a high capacity reversible data embedding algorithms, which was based on difference expansion, and the method had been extended by [8-10]. Ni [11] proposed an algorithm based on histogram shifting, which utilized the zero or the minimum points of the histogram of an image and slightly modified the pixel grayscale values to embed data into the image. Thodi [12] proposed a reversible watermarking technique called prediction error expansion, which better exploited the correlation inherent in the neighborhood of a pixel than the difference expansion scheme. However, most previous works focused on the image reversible watermark. Dittman [13] first proposed a reversible authentication scheme for 3D meshes. Lu [14] presented a reversible data hiding algorithm in the vector quantization (VQ) domain. Because the VQ compression is lossy the algorithm can only completely recover the compressed mesh. In [15], a reversible watermarking algorithm was proposed for 3D mesh models based on prediction error expansion. They predicted a vertex position by calculating the centroid of

its traversed neighbors, and then expanded the prediction error for data embedding. Despite the fact that the original mesh can be exactly recovered, the watermarked models are strongly distorted as compared with the original models.

In this paper, we apply difference expansion and difference shifting to embed the watermark in 3D meshes. A sequence of vertices is formed by adopting a mesh traversal strategy, and the difference between every two adjacent coordinates falls into three parts: one is shifted right if the difference is bigger than $K-1$; another is shifted left if the difference is smaller than $-K$; and the third part is used to embed the data (K is a positive integer controlled by the user). At the same time the distortion caused by expansion is uniformly separated into the two vertices, which keep the high quality of the models. Our algorithm needs not the location information, which improves the embedding capacity.

The rest of this paper is given as follows: In Section 2, we analyze the idea of difference expansion and difference shifting. A detailed description of our algorithm is provided in Section 3. Experimental results are presented in Section 4, and Section 5 concludes.

2 The Idea of Difference Expansion and Difference Shifting

There are two methods used to reversibly hide data in images: difference expansion [7] and histogram shifting [11]. Now we demonstrate the idea of difference expansion and difference shifting as follows.

If x_1 and x_2 are the gray values of a pixel-pair, then the integer-mean m and the difference d are defined as

$$\begin{cases} m = \text{floor}((x_1 + x_2) / 2) \\ d = x_1 - x_2 \end{cases} \quad (1)$$

Since this transformation is invertible, the gray levels x_1 and x_2 can be given

$$\begin{cases} x_1 = m + \text{floor}((d + 1) / 2) \\ x_2 = m - \text{floor}(d / 2) \end{cases} \quad (2)$$

Using the difference d , we can hide the bit b via the following equations

$$\begin{cases} d' = 2d + b & -K \leq d \leq K - 1 \\ d' = d + K & d \geq K \\ d' = d - K & d < -K \end{cases} \quad (3)$$

where K is the threshold controlled by the user.

The watermarked pixel x_1' and x_2' can be calculated by d' and m via (2). When $d \geq K$ or $d < -K$, we shift the difference d further away from the zero point, which is called the difference shifting, and leave $[K, 2K - 1]$ and $[-2K - 1, -K - 1]$ empty for difference expansion. When $-K \leq d \leq K - 1$,

we expand the difference d to embed the watermark bit b . As a result, the location map is unnecessary when extracting the information, which improves the embedding capacity.

3 Watermark Embedding and Extracting

The mesh geometry can be denoted by a tuple $\{V, F\}$, where $V = \{v_1, v_2, \dots, v_M\}$ is the set of vertex position defining the shape of the mesh in R^3 , and $F = \{f_1, f_2, \dots, f_N\}$ is the set of faces, as described in OBJ format.

Starting at a certain vertex \bar{v}_1 , we can get a sequence of vertices $\{\bar{v}_1, \bar{v}_2, \dots, \bar{v}_m\}$ ($m \leq M$) according to a certain mesh traversal order. As we all know, the sequence is different if the starting vertex is different. So we denote \bar{v}_1 as the secret key.

Normally, the vertex coordinates are floating point numbers. In order to perform difference expansion and difference shifting, all coordinates should be first transformed to integers. The integer coordinates (x, y, z) can be defined as follows:

$$(x, y, z) = \text{floor}((x_i, y_i, z_i) \times 10^p), \quad p \leq p_{\max}.$$

where (x_i, y_i, z_i) is the original coordinates with floating point numbers and p_{\max} is the maximum number of digits after the decimal point [16].

3.1 Watermark Embedding

The embedding process hides a binary message, which includes the hash of the original mesh for data authentication. The sequence of vertices of the mesh is denoted by $\{\bar{v}_1, \bar{v}_2, \dots, \bar{v}_m\}$, where $\bar{v}_i = \{\bar{x}_i, \bar{y}_i, \bar{z}_i\}$ is the i th vertex and $\{\bar{x}_i, \bar{y}_i, \bar{z}_i\}$ is the integer coordinates of \bar{v}_i . For the first pair (\bar{v}_1, \bar{v}_2) , the difference and the integer-mean of two vertices are calculated for x , y and z respectively

$$\begin{cases} dx = \bar{x}_1 - \bar{x}_2 \\ mx = \text{floor}\left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right), \end{cases} \tag{4}$$

$$\begin{cases} dy = \bar{y}_1 - \bar{y}_2 \\ my = \text{floor}\left(\frac{\bar{y}_1 + \bar{y}_2}{2}\right), \end{cases} \tag{5}$$

$$\begin{cases} dz = \bar{z}_1 - \bar{z}_2 \\ mz = \text{floor}\left(\frac{\bar{z}_1 + \bar{z}_2}{2}\right). \end{cases} \tag{6}$$

We just take x coordinates as an example since the situation is exactly the same as that in y and z coordinates. According to the difference, we can decide whether or not to embed information bit b via the equations

$$\begin{cases} dx' = 2dx + b & -K \leq dx \leq K - 1 \\ dx' = dx + K & dx \geq K \\ dx' = dx - K & dx < -K \end{cases} \quad (7)$$

Then the watermarked coordinates \bar{x}_1' and \bar{x}_2' can be calculated as follows

$$\begin{cases} \bar{x}_1' = mx + \text{floor}((dx'+1)/2) \\ \bar{x}_2' = mx - \text{floor}(dx'/2) \end{cases} \quad (8)$$

We get the new watermarked vertices (\bar{v}_1', \bar{v}_2') , transform the second pair (\bar{v}_2', \bar{v}_3') using (4), (7) and (8) to embed the watermark, and get the new watermarked vertices $(\bar{v}_2'', \bar{v}_3'')$. For the other pairs, the difference sequence and the integer mean sequence for x (y or z) coordinates can be obtained by the same procedure, and the watermark can be embedded into x (y or z) coordinates by difference expansion as above. After that, the watermarked vertices are $\{\bar{v}_1', \bar{v}_2'', \dots, \bar{v}_{m-1}'', \bar{v}_m'\}$, where $\{\bar{v}_1', \bar{v}_m'\}$ are modified only once, and $\{\bar{v}_2'', \dots, \bar{v}_{m-1}''\}$ twice. So the pure payload increases.

The maximum distortion between x_i and x_i' can be obtained:

$$\begin{cases} |x_i - x_i'| \leq \text{floor}(K/2) + 1, & i=1, m \\ |x_i - x_i'| \leq 2\text{floor}(K/2) + 2, & i=2, \dots, m-1 \end{cases} \quad (9)$$

Equation (9) shows that the distortion induced by data hiding can be controlled by K .

Different from Wu's algorithm, we embed the watermark into the vertex coordinates by modifying the differences between the adjacent vertex coordinates. The scheme can keep the mesh topology unchanged, and need not record the extra information.

3.2 Watermark Extracting and Original Data Recovering

Given the watermarked mesh, we can get the same sequence of vertices according to the starting vertex \bar{v}_1 and the same mesh traversal order. Then we extract the watermark and recover the original data in x coordinates as an example. The situation is exactly the same as that in y and z coordinates.

We denote the sequence of vertices as $\{\bar{v}_1', \bar{v}_2'', \dots, \bar{v}_{m-1}'', \bar{v}_m'\}$, and calculate the difference dx' and the integer-mean mx of the pair $(\bar{v}_{m-1}'', \bar{v}_m')$. When $-2K \leq dx' \leq 2K - 1$, the watermark can be obtained by extracting the lowest bits of difference dx' . In order to recover the original vertices, it is necessary to obtain the original difference dx via the following equations

$$\begin{cases} dx = \text{floor}(dx'/2) & -2K \leq dx' \leq 2K - 1 \\ dx = dx' - K & dx' \geq 2K \\ dx = dx' + K & dx' < -2K \end{cases} \quad (10)$$

Then the original x coordinates can be exactly recovered according to (2), using the restored difference dx and the integer-mean mx . We can extract the watermark and get the pair $(\bar{v}_{m-1}', \bar{v}_m')$, so \bar{v}_m is recovered. For the pair $(\bar{v}_{m-2}', \bar{v}_{m-1}')$, we can extract the watermark, get the pair $(\bar{v}_{m-2}', \bar{v}_{m-1}')$, and recover the vertex \bar{v}_{m-1} by the same procedure. We do the same procedure up to the first pair (\bar{v}_1', \bar{v}_2') , extract the watermark, and recover the vertex (\bar{v}_1, \bar{v}_2) . Then the original watermark is given as the extracted watermark is arranged in the contrary order.

4 Experimental Results

We have applied our proposed algorithm to four models: bunny, apple, eight and horse. Table 1 lists the number of vertices and the number of vertices after traversing of four models. Table 2 is the experimental results of our scheme which include the payload, the signal-noise-ratio (SNR, [15]) and the root mean square error (RMSE) of different models. The results show that the payload increases, the SNR decreases and RMSE increases with the threshold (K) increasing. Table 3 lists the payload, the SNR and RMSE of Wu's algorithm. According to Table 2 and Table 3, it can be seen that our algorithm embeds more watermarks than Wu's algorithm keeping the higher SNR and the lower RMSE.

Table 1. The number of vertices and the number of vertices after traversing three models

3D models	bunny	apple	ball	horse
the number of vertices (M)	34835	891	1202	48485
the number of vertices after traversing (m)	33938	762	1156	40872

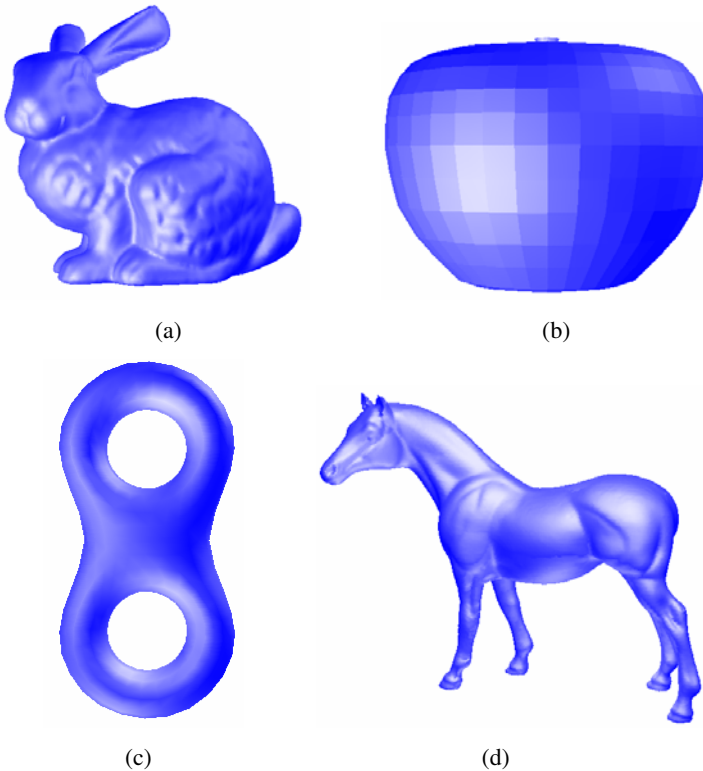
Table 2. Payload, SNR and RMSE of different models (Our proposed algorithm)

3D models	p	K	payload	SNR	RMSE
bunny	3	188	3880	68.1664	0.2393
	3	388	6838	61.8743	0.4944
	3	588	9035	58.2694	0.7494
	3	688	9998	56.91	0.8767
apple	4	2	134	44.1887	2.5112×10^{-4}
	4	3	165	40.3820	3.8694×10^{-4}
eight	4	2	107	69.4359	2.6246×10^{-4}
	4	3	126	65.6072	4.0797×10^{-4}
horse	5	9	22859	55.0140	9.8691×10^{-5}
	4	5	55538	40.5159	5.2260×10^{-4}
	4	8	73496	37.1807	7.6842×10^{-4}

Table 3. Payload, SNR and RMSE of different models ([15])

3D models	p	K	payload	SNR	RMSE
bunny	2	1388	4782	35.3594	0.2393
apple	4	85	41	17.1365	0.0052
eight	3	59	19	26.8759	0.0249
horse	5	99	8014	38.5179	6.1210×10^{-4}

The original models, the watermarked models using Wu's algorithm and the watermarked models using our algorithm are shown in Figure 1, 2 and 3. Compared with models in Figure 1, no visible distortions exist in Figure 3, and visible distortions exist in Figure 2. In the experiments, we can extract the correct watermarks from the watermarked models and exactly recover the original meshes.

**Fig. 1.** The original models (a) bunny (b) apple (c) eight (d) horse

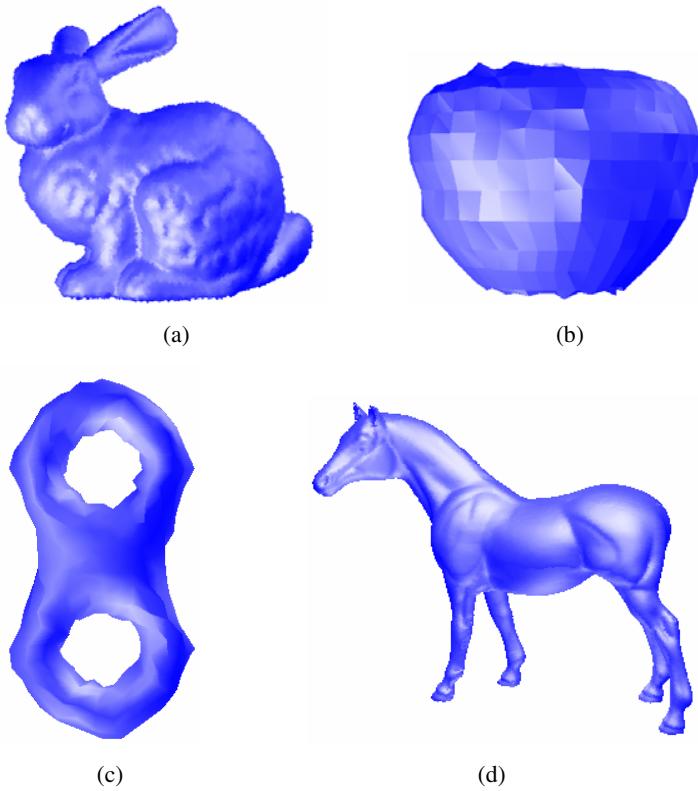


Fig. 2. The watermarked models ([15]) (a) the watermarked bunny after embedding 4782 bits, (b) the watermarked apple after embedding 41 bits, (c) the watermarked eight after embedding 19 bits, (d) the watermarked horse after embedding 8014 bits

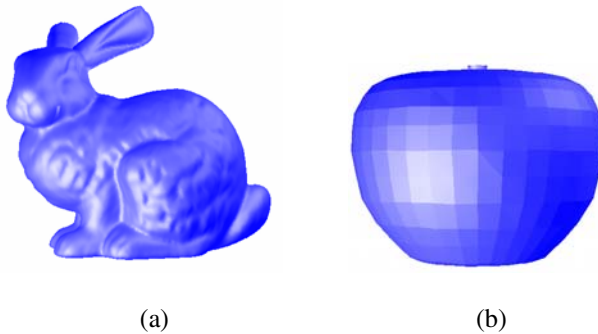
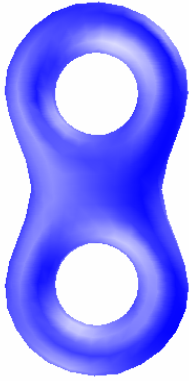
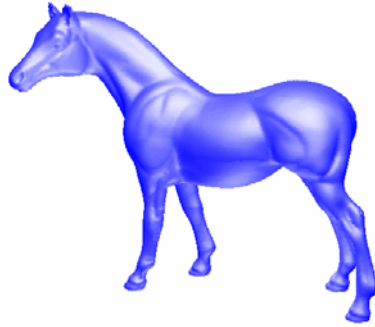


Fig. 3. The watermarked models (our algorithm) (a) the watermarked bunny after embedding 9998 bits, (b) the watermarked apple after embedding 134 bits, (c) the watermarked eight after embedding 126 bits, (d) the watermarked horse after embedding 22859 bits



(c)



(d)

Fig. 3. (continued)

5 Conclusion

In this paper, we have proposed a reversible watermarking algorithm of 3D meshes, which is based on the difference expansion and difference shifting. A sequence of vertices can be obtained by using a certain mesh traversal order. The difference between two vertices coordinates are expanded to embed the watermark without changing the mesh topology. The embedded watermark can be extracted from the watermarked mesh while the original mesh can be exactly recovered without any extra information. The distortion caused by the reversible watermark embedding is unnoticeable. So our reversible algorithm can be used for content authentication of 3D meshes.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China (No. 60773179 and No. 10601051) and the National Basic Research Program (973) of China (No. 2004CB318000) for financial support.

References

1. Honsinger, C.W., Jones, P.W., Rabbani, M., Stoffel, J.C.: Lossless recovery of an original image containing embedded data. US Patent, Docket No: 77102 (1999)
2. Goljan, M., Fridrich, J., Du, R.: Distortion-free data embedding for images. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 27–41. Springer, Heidelberg (2001)
3. Fridrich, J.: Lossless data embedding—new paradigm in digital watermarking. J. EURASIP Journal on Applied Signal Processing 2002(2), 185–196 (2002)
4. Celik, M.U., Sharma, G., Tekalp, A.M., Saber, E.: Reversible data hiding. In: Proceedings of International Conference on Image Processing, pp. 157–160 (2002)

5. Xuan, G., Zhu, J., Chen, J., Shi, Y.Q., Ni, Z.C., Su, W.: Distortionless data hiding based on integer wavelet transform. *J. Electronics Letters* 38(25), 1646–1648 (2003)
6. De Vleeschouwer, C., Delaigle, J.F., Macq, B.: Circular interpretation of bijective transformations in lossless watermarking for media asset management. *J. IEEE Transactions on Multimedia* 5(1), 97–105 (2003)
7. Tian, J.: Reversible data embedding using a difference expansion. *J. IEEE Transactions on Circuits and Systems for Video Technology* 13(8), 890–896 (2003)
8. Alattar, A.M.: Reversible watermark using the difference expansion of a generalized integer transform. *J. IEEE transactions on image processing* 13(8), 1147–1156 (2004)
9. Chang, C.C., Lu, T.C.: A difference expansion oriented data hiding scheme for restoring the original host images. *J. The Journal of Systems & Software* 79(12), 1754–1766 (2006)
10. Cho, J., Prost, R., Jung, H.: An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms. *J. IEEE Transactions on Signal Processing* 55(1), 142–155 (2007)
11. Ni, Z., Shi, Y.Q., Ansari, N., Su, W.: Reversible data hiding. *J. IEEE Transactions on Circuits and Systems for Video Technology* 16(3), 354–362 (2006)
12. Thodi, D.M., Rodriguez, J.J.: Expansion embedding techniques for reversible watermarking. *J. IEEE transactions on image processing* 16(3), 721–730 (2007)
13. Dittmann, J., Benedens, O.: Invertible Authentication for 3D Meshes. In: *Proceedings of SPIE, Security and Watermarking of Multimedia Contents*, vol. 5020, pp. 653–664 (2003)
14. Lu, Z.-M., Li, Z.: High capacity reversible data hiding for 3D meshes in the PVQ domain. In: Shi, Y.Q., Kim, H.-J., Katzenbeisser, S. (eds.) *IWDW 2007*. LNCS, vol. 5041, pp. 233–243. Springer, Heidelberg (2008)
15. Wu, H.T.: Reversible Watermarking of 3D Mesh Models by Prediction-error Expansion. In: *10th IEEE Workshop on Multimedia Signal Processing*, pp. 797–802. IEEE Press, Los Alamitos (2008)
16. Wang, X.T., Shao, C.Y., Xu, X.G., Niu, X.M.: Reversible Data-Hiding Scheme for 2-D Vector Maps Based on Difference Expansion. *J. IEEE Transactions on Information Forensics and Security* 2(3 Part 1), 311–320 (2007)

Neighbor-List Based Pairwise Key Management Scheme in Wireless Sensor Networks

Xing Zhang¹, Jingsha He², and Qian Wei¹

¹ College of Computer Science and Technology, Beijing University of Technology

² School of Software Engineering, Beijing University of Technology,
100124 Beijing, China

{zhang_xing, weiqian}@emails.bjut.edu.cn, jhe@bjut.edu.cn

Abstract. Through building the neighbor list, we propose effective pairwise key storage and update methods. The structure of the neighbor list is simple and can save the storage space for the node. By using the pseudo random function and the elliptic curve digital signature algorithm, we establish the pairwise keys decentralized completely and solve the new node joining problem. Analysis shows that neighbor-list based pairwise key management scheme is energy efficient.

1 Introduction

Wireless sensor networks (WSNs) have become one of the most interesting and promising areas over the past few years. However, the constrained capacity of wireless sensor nodes such as limitation in computation power, memory and battery lifetime increases the insecurity of wireless sensor networks. As a part of the basic requirement for security, key management plays a central role in encryption and authentication.

All pairwise key management schemes proposed so far cannot address how to store and update the pairwise keys. Moreover, there are still a lot of deficiencies in current key management schemes, such as the authentication problem of node identity, the security problem of new node joining, and so forth [1, 2]. Further, due to the limited battery lifetime, security mechanisms for sensor networks must be energy efficient.

In this paper, we use the neighbor list to help nodes manage more efficiently the pairwise keys and the sequence numbers so as to guarantee the confidentiality, authenticity, integrity and freshness of data transfer. The pairwise key is set up only through some broadcast information obtained during the network initialization stage. Meanwhile, no other message is further exchanged, and therefore the communication overhead is very small. The pairwise keys are completely decentralized, and hence some nodes' compromise will not affect the other non-compromised pairwise keys. In order to solve the new node joining problem, we propose a composite mechanism based on the elliptic curve digital signature algorithm to entirely deal with the situation and will not cost the nodes too much resource.

The rest of this paper is organized as follows. We present the neighbor-list based pairwise key management scheme in detail in Section 2. In Section 3 and 4, we

analyze the security and performance of the key management scheme. Finally we conclude the paper in Section 5.

2 Neighbor-List Based Pairwise Key Management Scheme

Before deployment, every node is predistributed a pseudo-random function f [3] and a master key K_I . Through the pseudo-random function f and the master key K_I , every node can compute the individual key of any node, say A

$$K_a = f(K_I, ID_a). \quad (1)$$

2.1 Neighbor List Building Phase

After deployment, each sensor node will broadcast a network joining message so as to determine the neighboring relationships in the network. To prevent two or more nodes attempting to transmit at the same time in the network initialization phase, each node implements a widely used binary exponential backoff algorithm to avoid collisions before sending the network joining packet. Additionally, in order to prohibit an adversary from working out all pairwise keys after obtaining K_I as it does in LEAP protocol [4], the pairwise key computed through the pseudo-random function should be updated in time.

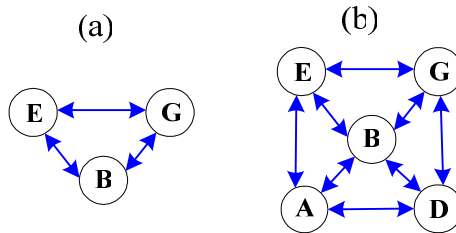


Fig. 1. (a) Initial phase (b) New node joining phase

As Fig. 1(a) shows, we take example for node E and G to illustrate the pairwise key setup procedure

$$\begin{aligned} E &\rightarrow *: join \parallel ID_e \parallel SN_e \parallel C(K_e, [ID_e \parallel SN_e]) \\ G &\rightarrow *: join \parallel ID_g \parallel SN_g \parallel C(K_g, [ID_g \parallel SN_g]). \end{aligned} \quad (2)$$

$$\begin{aligned} K_{eg} &= f(K_e \oplus K_g, SN_e \oplus SN_g) \\ SN_{eg} &= ID_e \oplus ID_g. \end{aligned} \quad (3)$$

where $join$ denotes the message type; \parallel is the concatenation operator; $C(K, M)$ is used to compute the message M 's message authentication code (MAC) with key K . According to equation (1), every receiving node can compute the sender's individual key (e.g. K_e and K_g). After verifying the joining packet correct, through equation (3) the neighboring nodes can get same pairwise key (e.g. K_{eg}) and sequence number (e.g.

SN_{eg}). In our scheme, the pairwise key is not stored with plaintext in the neighbor list to prevent the adversary getting the key after compromising the node. Before storage, a node’s pairwise key needs to be circled plus its individual keystored, say E

$$K'_{eg} = K_{eg} \oplus K_e . \tag{4}$$

where we use a simple method to encrypt the pairwise key and readers may choose a complicated way according to the higher security requirement. Additionally, the SN field and the Key Lifetime field are unified into one field for saving the node’s memory, as shown in Fig. 2. The sequence number SN increases by 1 after each message sent. When the SN reaches the predefined threshold, it will start from 1 and at the same time the pairwise key will be also updated. The SN is public and only used to ensure the packet fresh, so it is not necessarily secret. After the pairwise keys with its all neighbors are computed, a node will delete its neighbors’ individual keys (e.g. K_e and K_g), sequence numbers (e.g. SN_e and SN_g), the pseudo-random function f and master key K_I for improving security and saving storage space.

2 bytes	8 bytes	2 bytes
Node ID	Pairwise Key	Sequence Number

Fig. 2. The neighbor list

2.2 Neighbor List Maintaining Phase

To save power, a sensor node may employ early rejection by turning off its radio after determining the message is false or malicious. Upon the receiver getting a packet, it first sees whether the sender of the packet is in its neighbor list; then it sees whether the SN is as same as that in its neighbor list; then it recomposes the message authentication code and compares the computing result with the received MAC ; finally, it decrypts the message. Hence, the usage of the neighbor list not only improves the security of the data transfer but also reduces the node’s computational and communication overhead.

2.2.1 New Node Joining

We employ Elliptic Curve Digital Signature Algorithm (ECDSA) [5] to authenticate the node’s identity without compromising energy efficiency. As Fig. 1(b) shows, we suppose B is the existing node, A and D are new joining nodes. When the new node A sets up the pairwise keys with its neighbors, it broadcasts a joining message

$$A \rightarrow *: join \parallel ID_a \parallel P_a \parallel T_a \parallel SN_a \parallel C_a \parallel c_a . \tag{5}$$

P and s denote the public key and the private key respectively. The trusted authority TA’s public key P_T is pre-distributed to every node before deployment, and however, its private key s_T is not stored in the WSNs, so adversaries have no opportunity to get s_T by directly attacking the trusted authority. T_a is the timestamp before which a node should join the network. If the current time $t < T_a$, then node A will be regarded as a new node, otherwise the replica of a compromised node. In the paper, we assume the

sensor nodes are stationary. The handling of the compromising nodes sees Section 2.2.2. The signature $\langle C_a, c_a \rangle$ was calculated by the TA according to ECDSA [5]. The calculating process is as follows

$$\begin{aligned} C_a &= r_a G = (x_{ca}, y_{ca}) \\ c_a &= r_a^{-1} (H(ID_a \parallel P_a \parallel T_a) + s_T x_{ca}) \pmod{n}. \end{aligned} \quad (6)$$

where r_a is a random number; G is the generator in the cyclic group of points over the elliptic curve and has an order n of at least 160 bits; H is a hash function that can translate a binary sequence into an integer.

The receiving nodes, say D and B, calculate

$$V_a = c_a^{-1} (H(ID_a \parallel P_a \parallel T_a) G + c_a^{-1} x_{ca} P_T). \quad (7)$$

where $P_T = s_T G$, $P_a = s_a G$. If $V_a = C_a$, then A's neighbors can make sure that A is legitimate node. The above equation's verification procedure is as follows

$$V_a = c_a^{-1} r_a r_a^{-1} (H(ID_a \parallel P_a \parallel T_a) + x_{ca} s_T) G = c_a^{-1} r_a c_a G = C_a$$

Following the same procedure, the receivers can also verify identities of the node D after hearing the broadcast joining messages from D. The pairwise key negotiations fall into the following two situations: between two new nodes, between the new and old nodes.

New nodes are preloaded with the pseudo-random function f and a new master key K'_j . The pairwise key negotiation between two new nodes refers to equation (3), where $K_a = f(K'_j, ID_a)$ and $K_d = f(K'_j, ID_d)$. Consequently, the new node A and D establish the pairwise key without sending any other message again.

Below we take example for node A and B to demonstrate the pairwise key negotiation between the new and old nodes. After authenticating identities of the new node A and D, the old node B generates randomly a sequence number SN_b and broadcasts a response message with TA's signature $\langle C_b, c_b \rangle$

$$B \rightarrow *: \text{reply} \parallel ID_b \parallel P_b \parallel T_b \parallel SN_b \parallel C_b \parallel c_b. \quad (8)$$

According to Diffie-Hellman protocol [6], the node A and B calculate respectively the pairwise key with its own private key and the opposite's public key. The computing result is identical, because $K_{ab} = s_a P_b = s_a s_b G = s_b P_a$. In addition, their sequence number derives from $SN_{ab} = SN_a \oplus SN_b$.

Besides the broadcast identity message, no other messages need to be exchanged in the procedure of the pairwise key setup. Therefore, the communication overhead is low in the new node joining process.

Although the public-key algorithm is much more computationally expensive, it is easier to manage and more resilient to node compromise than symmetric key technology. Under the same security level, the smaller key size of ECC (elliptic curve cryptography) offers advantages of faster computational efficiency, as well as memory, energy, and bandwidth savings, thus ECC can be better suitable for resource constrained sensor nodes.

2.2.2 Pairwise Key Update and Revocation

To defend against cryptanalysis and to prevent an adversary from decrypting all the previously messages after compromising a sensor node, all pairwise keys must be

updated periodically. When a pairwise key's lifetime (i.e. the SN) reaches its predefined threshold, two neighboring nodes need to perform the neighbor list upgrade process. The node whose ID is smaller launches the updating process as shown below, and the neighbor lists will be revised accordingly.

$$\begin{aligned}
 & \text{If } ID_S < ID_R \\
 & \text{Unicast: } update \parallel ID_S \parallel ID_R \parallel SN \parallel E(K, [K' \parallel ID_S \parallel ID_R \parallel SN]) \quad (9) \\
 & \text{Acknowledgment: } ID_S \parallel ID_R \parallel SN+1 \parallel C(K', [ID_S \parallel ID_R \parallel SN+1]).
 \end{aligned}$$

where *update* denotes the message type; ID_S and ID_R are the identifiers of the sender and the receiver respectively; $E(K, M)$ is used to encipher the message M with key K ; K and K' are the old and new pairwise key respectively. Additionally, in order to achieve semantic security, the initialization vector IV is introduced into the encryption mechanism. One implication of semantic security is that encrypting the same plaintext two times should give two different cipher texts. Here we use a specially formatted 8 byte IV whose structure is $ID_R \parallel ID_S \parallel SN$, and cipher block chaining (CBC) to implement the encryption mechanism.

When some sensor nodes are captured by adversaries and become insider attackers, we must segregate them from the network to avoid that they eavesdrop on the information transmitted in the network and launch active attacks to the network. We assume that the intrusion detection scheme in the network can discover these compromised nodes, and can inform the neighbors of compromised nodes to revoke the pairwise keys related to compromised nodes from their neighbor lists in time.

3 Security Analysis

In the paper, we achieve semantic security by using a specially formatted 8 byte IV and cipher block chaining (CBC) to implement our encryption mechanism. Moreover, the pairwise key can be updated periodically, so our scheme can effectively fight off the cryptanalysis.

Almost all key management schemes use the nonce to implement replay protection. A nonce is randomly generated for every time and cannot be always saved in a node, so an adversary can later replay the message encrypted or verified with the nonce. While the SN used in our scheme is stored in each node's neighbor list, an adversary cannot replay or forge the message encrypted or verified with the specific SN. Thus, compared with the nonce, the SN does better in counteracting the replay attack and the node identity forgery attack.

Through equation (3), we can conclude that all pairwise keys in our scheme are decentralized and are basically different. As thus they have not a centralized rule any more and cannot be computed uniformly. Even though the adversary knows the pseudo-random function f and the master key K_t , it cannot get non-compromised pairwise keys, because each node's first SN in equation (3) has already been erased after establishing the pairwise key. Additionally, we employ ECDSA to authenticate the new joining node's identity. Even though an adversary compromises a node, it cannot get the private key and certificate of non-compromised nodes. The adversary cannot even inject the compromised node to other places of the network, because when the current time t is greater than the node's timestamp T , the node will be regarded as the replica of a compromised node.

4 Performance Evaluation

4.1 Computational Cost

In our scheme, the computational overhead mainly occurs in the node identity verification and the pairwise key setup. In initial phase, we need to verify the node's identity by computing a *MAC* and to set up the pairwise key by running a pseudo random function. Due to the computational efficiency of pseudo random functions, the computational overhead related to them is negligible. Although we exploit ECC algorithm in the new node joining phase, ECC is computationally efficient, especially for assembly language implementations, which makes a 160-bit point multiplication of ECC requires only 0.81s. The new node identity authentication needs two point multiplication operations, whereas the pairwise key setup between the new and old nodes needs only one point multiplication operation. Additionally, by using the SN instead of the nonce, the scheme eliminates the computational overhead for generating the nonce with each sending packet.

4.2 Communication Cost

In the pairwise key setup procedure, except the broadcast joining or replying message, no other message is exchanged between neighbors, and therefore the communication overhead in our scheme is very low. According to equation (2), (5), (8), (9), the messages sent by every node can be authenticated in our scheme, but in contrast other schemes [1, 2] cannot do so. If the message cannot be authenticated by the recipient, then the adversary can freely inject malicious false packets into the network. Moreover, the recipients have to feed back the response packets and wait for authenticating the sender, so communication overhead in our scheme is much less than that in other schemes. Furthermore, we know that communication overhead is over three orders of magnitude greater than computation overhead in energy consumption.

4.3 Storage Requirement

The master key K_I is deleted after establishing the pairwise key. Each node needs only to store its individual key, the pairwise keys with its all neighbors, its public key P and private key s , and the trusted authority's public key P_T . For saving the node's memory, the SN field and the Key Lifetime field are unified into one field in the neighbor list. As Fig. 2 shows, each entry of the neighbor list only occupies every node's 12-byte storage space. We suppose a node has averagely m neighbors, and then the required space by its neighbor list is $12m$ bytes. If $m=15$, then the consumed storage space is only 180 bytes.

Overall, we conclude our scheme is scalable and efficient in computation, communication and storage.

5 Conclusion

We propose a neighbor-list based pairwise key management scheme which combines the benefits of symmetric and asymmetric cryptography. Through the neighbor list,

we can efficiently establish and maintain the pairwise key and the sequence number to guarantee the confidentiality, authenticity, integrity and freshness of data transfer. Additionally, the neighbor list can help the sensor node to employ early rejection to save power.

References

1. Wang, Y., Attebury, G., Ramamurthy, B.: A Survey of Security Issues in Wireless Sensor Networks. *IEEE Communications Surveys & Tutorials* 8(2), 2–22 (2006)
2. Zhou, Y., Fang, Y., Zhang, Y.: Securing Wireless Sensor Networks: A Survey. *IEEE Communications Surveys & Tutorials* 10(3), 6–28 (2008)
3. Goldreich, O., Goldwasser, S., Micali, S.: How to Construct Random Functions. *Journal of the ACM* 33(4), 210–217 (1986)
4. Zhu, S., Setia, S., Jajodia, S.: LEAP: Efficient Security Mechanism for Large-Scale Distributed Sensor Networks. In: *Proc. 10th ACM Conf. Computer and Commun. Security (CCS 2003)*, Washington, DC (October 2003)
5. Vanstone, S.: Responses to NIST's Proposal. *Communications of the ACM* 35(July), 50–52 (1992)
6. Diffie, W., Hellman, M.E.: New Directions in Cryptography. *IEEE Transactions on Information Theory* 22(6), 644–654 (1976)

Author Index

- Anderson, John 1
Ang, Rebecca P. 183, 195
Antunovic, Michael 385
Ashman, Helen 385
- Bachimont, Bruno 226
Bottini, Thomas 226
Bradshaw, Jeffrey M. 2
Brahmi, Zaki 347
Buzzanca, Armando 288
- Cao, Cun Gen 160
Castellano, Giovanna 288
Čereković, Aleksandra 7
Chang, Chew Hung 171
Chang, Hsin-I 250
Chang, Yu-Teng 238, 250, 311, 357
Chatterjea, Kalyani 171
Chen, Chen 397
Chen, Hua 501
Chen, Ping-Chang 238, 357
Chen, Yuquan 136
Cheng, Yiyang 376
Chua, Alton Y.K. 183, 195, 478
Costa, Rosa M. 335
- Delopoulos, Anastasios 126
Diao, Lu Hong 160
- Falelakis, Manolis 126
Fanelli, Anna Maria 288
Felfernig, Alexander 69
Furukawa, Takuya 7
- Gago, Izabela Salotti Braga 335
Gammoudi, Mohamed Mohsen 347
Geng, Qin 501
Goh, Dion Hoe-Lian 171, 183, 195, 478
- Han, Ning 397
Han, Shu-Huei 357
He, Jingsha 522
Hu, Changzhen 409
Hu, Xiaohong 370
Huang, Hsuan-Hung 7
- Huang, Zhisheng 418
Hui-bing, Zhang 263
- Ilahi, Mounira 347
- Jing-wei, Zhang 263
Jung, Hanmin 104
- Kamei, Koji 19
Karydas, Lazaros 126
Kim, Do-Wan 104
Kim, Thi Nhu Quynh 171
Komatsu, Takanori 19
Kotoulas, Spyros 430
- Lee, Chei Sian 183, 195
Lee, Mi-Kyoung 104
Lei, Song 114
Li, Liping 467
Li, Mi 207
Li, Peiqiang 430
Li, Wenbin 376
Li, Xin 467
Li, Yuan 397
Lim, Ee-Peng 171
Liu, Chang 93
Liu, Hui 136
Liu, Jiming 323
Liu, Jixue 148
Liu, Lei 160
Liu, TaiFeng 376
Lo, Chih-Yao 238, 250, 311, 357
Lu, Shengfu 207
- Ma, Xinming 370
Malarvizhi, Nandagopal 490
Mandl, Monika 69
Merckel, Loic 31
Mogles, Nataliya M. 54
Morizet-Mahoudeaux, Pierre 226
- Nakano, Yukiko 7
Nguyen, Quang Minh 171
Nishida, Toyooki 7, 19, 31
Niu, Jianwei 93

- Ohmoto, Yoshimasa 19, 42
 Ohno, Takehiko 42
 Okada, Shogo 19
 Okadome, Takeshi 19
- Pandžić, Igor S. 7
 Pironti, Marco 299
 Pisano, Paola 299
- Qiu, Hongjun 323
- Razikin, Khasfariyati 171, 478
 Reder, Lynne 4
 Remondino, Marco 299
 Ren, Jiadong 409
- Schubert, Monika 69
 Shahriar, Md. Sumon 148
 Shi, Jijun 281
 Shi, Lei 370
 Shi, Yunhui 273, 281
 Shi, Zhongzhi 5
 Smith, Gavin 385
 Song, Yangyang 207
 Sumi, Yasuyuki 19
 Sun, Aixin 171
 Sun, Jingyu 442
 Sun, Xiaowei 273
- Takahashi, Satoshi 217
 Tanaka, Yuzuru 454
 Theng, Yin-Leng 171
 Treur, Jan 54
- Ueda, Kazuhiro 19, 42
 Urbani, Jacopo 430
 Uthariaraj, V. Rhymend 490
 Utsumi, Hisao 454
- van der Mee, Andy 54
 van Harmelen, Frank 3
- Wang, Guozhao 513
 Wang, Haibo 217
 Wang, Hao 93
 Wang, Kunsheng 409
 Wang, Ruizhi 442
 Wang, Yan 418
 Wang, Yongji 81
 Wei, Liang 114
 Wei, Qian 522
 Wen, Wen 281
 Werneck, Vera M.B. 335
 Wu, Dan 513
 Wu, Di 409
 Wu, Hu 81
 Wu, Jinglong 217
- Xu, Yong 19
- Yamaoka, Yuji 7
 Yan, Shu Ying 160
 Yao, Chunlian 397
 Yin, Baocai 273, 281
 You, Beom-Jong 104
 Yu, Kun 93
 Yu, Xueli 442
- Zeng, Yi 418, 430
 Zhang, Jianhong 501
 Zhang, Jun 467
 Zhang, Li 370
 Zhang, Sen 160
 Zhang, Xindong 376
 Zhang, Xing 522
 Zhong, Ning 207, 323, 376, 418, 430, 442
 Zhou, Zhi-Hua 6
 Zhou, Zhoufan 454