Ngoc Thanh Nguyen
Geun Sik Jo
Robert J. Howlett
Lakhmi C. Jain (Eds.)

LNAI 4953

Agent and Multi-Agent Systems: Technologies and Applications

Second KES International Symposium, KES-AMSTA 2008
Incheon, Korea, March 2008
Proceedings

Springer

Lecture Notes in Artificial Intelligence 4953

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Ngoc Thanh Nguyen   Geun Sik Jo
Robert J. Howlett   Lakhmi C. Jain (Eds.)

Agent and Multi-Agent Systems: Technologies and Applications

Second KES International Symposium, KES-AMSTA 2008
Incheon, Korea, March 26-28, 2008
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Ngoc Thanh Nguyen
Wroclaw University of Technology, Institute of Information Science and Engineering
Str. Janiszewskiego 11/17, 50-370 Wroclaw, Poland
E-mail: ngoc-thanh.nguyen@pwr.wroc.pl

Geun Sik Jo
Inha University, School of Computer Science and Information Engineering
#253 YongHyun-dong Nam-Ku, Incheon, Korea
E-mail: gsjo@inha.ac.kr

Robert J. Howlett
University of Brighton, School of Environment and Technology
Brighton, BN2 4GJ, UK
E-mail: r.j.howlett@bton.ac.uk

Lakhmi C. Jain
University of South Australia, School of Electrical and Information Engineering
Mawson Lakes, South Australia 5095, Australia
E-mail: lakhmi.jain@unisa.edu.au

Library of Congress Control Number: 2008922056

CR Subject Classification (1998): I.2, H.4, H.3, J.1, H.5, K.6, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-78581-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-78581-1 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12238828 06/3180 5 4 3 2 1 0

Preface

Following from the very successful First KES Symposium on Agent and Multi-Agent Systems – Technologies and Applications (KES-AMSTA 2007), held in Wroclaw, Poland, 31 May–1 June 2007, the second event in the KES-AMSTA symposium series (KES-AMSTA 2008) was held in Incheon, Korea, March 26–28, 2008. The symposium was organized by the School of Computer and Information Engineering, Inha University, KES International and the KES Focus Group on Agent and Multi-agent Systems. The KES-AMSTA Symposium Series is a sub-series of the KES Conference Series.

The aim of the symposium was to provide an international forum for scientific research into the technologies and applications of agent and multi-agent systems. Agent and multi-agent systems are related to the modern software which has long been recognized as a promising technology for constructing autonomous, complex and intelligent systems. A key development in the field of agent and multi-agent systems has been the specification of agent communication languages and formalization of ontologies. Agent communication languages are intended to provide standard declarative mechanisms for agents to communicate knowledge and make requests of each other, whereas ontologies are intended for conceptualization of the knowledge domain.

The symposium attracted a very large number of scientists and practitioners who submitted their papers for nine main tracks concerning the methodology and applications of agent and multi-agent systems, a doctoral track and two special sessions. From the submissions for KES-AMSTA 2008 coming from more than 20 countries throughout the world, only 91 papers were selected for presentation and inclusion in the proceedings. Many of the papers have been reviewed in a double blind process. The Program Committee defined the following main tracks: *Methodological Aspects of Agent Systems; Agent-Oriented Web Applications; Ontology Management; Multi-Agent Resource Allocation; Cooperation, Coordination, and Teamwork; Agents for Network Management; Multi-Agent Learning; Mobile Agents and Agents in Applications*. Jointly with the main tracks of the symposium there was also the doctoral track and the following two special sessions: *Intelligent and Secure Agent for Digital Content Management* and *Self-Organization in Multi-Agent Systems*.

We would like to thank the invited speakers – J.K. Lee (Korea), L.C. Jain (Australia), and D.S. Huang (China) – for their interesting and informative talks of world-class standard.

Special thanks go to one of the Program Co-chairs, Jason J. Jung (Korea), for his great help in the organization work. We thank the Program Co-chairs for their efforts in managing the reviews and the paper selection process. Thanks are due to the Program Committee and the Board of Reviewers, essential for reviewing the papers to ensure the high quality. We thank the members of the Local Organizing Committee, the Doctoral Track Chair, Publicity Chair and Special Session Chairs. We acknowledge with gratitude the efforts of the KES Focus Group on Agent and Multi-Agent Systems for coordinating the organization of the symposium. We extend

cordial thanks to the KES Secretariat for the support with central administration and the registration process. Special thanks go to Franz I.S. Ko and his team for very effective publicity action. Finally, we thank authors, presenters and delegates for their contribution to a successful event.

Thanks are also due to the many experts who contributed to making the event a success.

January 2008

Ngoc Thanh Nguyen
Geun Sik Jo
Robert J. Howlett
Lakhmi C. Jain

Symposium Organization

Honorary Chair

Jae Kyu Lee,
College of Business,
Korea Advanced Institute of Science and Technology, Korea

General Chair

Geun Sik Jo
School of Computer and Information Engineering
Inha University, Korea

Program Chair

Ngoc Thanh Nguyen
Institute of Information Science and Engineering
Wroclaw University of Technology, Poland

Executive Chair

Robert J. Howlett
Centre for SMART Systems
School of Environment and Technology
University of Brighton, UK

Program Co-chairs

Jason J. Jung
Department of Computer Engineering
Yeungnam University, Korea

Janusz Sobecki
Institute of Applied Informatics
Wroclaw University of Technology, Poland

Longbing Cao
Faculty of Information Technology
University of Technology Sydney, Australia

Local Organizing Committee Chair

Jong-Sik Lee
School of Computer and Information Engineering
Inha University, Korea

Doctoral Track Chair

Radoslaw Katarzyniak
Institute of Applied Informatics
Wroclaw University of Technology, Poland

Publicity Chair

Sang-Gil Kang,
School of Computer and Information Engineering,
Inha University, Korea

Special Session Chairs

1. *Intelligent and Secure Agent for Digital Content Management*
   Geuk Lee
   Hannam University, Republic of Korea

2. *Self-Organization in Multi-Agent Systems*
   Dariusz Krol
   Wroclaw University of Technology, Poland, and

   Gabriel Ciobanu
   University of Iasi, Romania

Invited Speakers

1. Jae Kyu Lee,
   Korea Advanced Institute of Science and Technology, Korea

2. Lakhmi C. Jain
   University of South Australia, Australia

3. De-Shuang Huang
   Chinese Academy of Sciences, China

International Program Committee

Costin Badica	University of Craiova, Romania
María-Victoria Belmonte	Universidad de Málaga, Spain
Maria Bielikova	Slovak University of Technology, Slovakia
Tru Cao	Ho Chi Minh City University of Technology, Vietnam
Noëlle Carbonell	Nancy Université and LORIA, France
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan

Paul Davidsson	Blekinge Institute of Technology, Sweden
Grzegorz Dobrowolski	AGH - University of Science and Technology, Poland
Bogdan Gabrys	Bournemouth University, UK
Vladimir Gorodetsky	Russian Academy of Sciences, Russia
Mirsad Hadzikadic	University of North Carolina at Charlotte, USA
Anne Hakansson	Uppsala University, Sweden
Sun-Gwan Han	Gyeong-in National University of Education, Korea
Sung-Kook Han	Wonkwang University, Korea
Ronald L. Hartung	Franklin University, USA
Heikki Helin	TeliaSonera, Finland
Sachio Hirokaw	Kyushu University, Japan
Adam Jatowt	Kyoto University, Japan
Gordan Jezic	University of Zagreb, Croatia
Jong-Jin Jung	Daejin University, Korea
Oleg Karsaev	St. Petersburg Institute for Informatics and Automation, Russia
Radoslaw Katarzyniak	Wroclaw University of Technology, Poland
Przemyslaw Kazienko	Wroclaw University of Technology, Poland
Chonggun Kim	Yeungnam University, Korea
Hong-Gee Kim	Seoul National University, Korea
In-Young Ko	Information and Communications University, Korea
Dariusz Krol	Wroclaw University of Technology, Poland
Daniel Kudenko	University of York Heslington, UK
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Gu-In Kwon	Inha University, Korea
Raymond Lau	Systems City University of Hong Kong, China
Eun-Ser Lee	University of Soonsil, Korea
Huey-Ming Lee	Chinese Culture University, Taiwan
Jaeho Lee	The University of Seoul, Korea
Kyoung-Jun Lee	Kyunghee University, Korea
Zne-Jung Lee	Huafan University, Taiwan
Pawan Lingras	Saint Mary's University, Canada
Tokuro Matsuo	Yamagata University, Japan
Wee Keong Ng	Nanyang Technological University, Singapore

Board of Additional Reviewers

Gui Jung Kim
Sijung Kim
In-Young Ko
Bogdan Trawinski
Adelinde Uhrmacher
Mirko Violi

ES-AMSTA Symposium Series and Focus Group on Agent and Multi-Agent Systems Chair

Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland

KES International Conference Series Chairs

Robert J. Howlett, University of Brighton, UK
Lakhmi C. Jain, University of South Australia, Australia

Table of Contents

Main Track 1: Methodological Aspects of Agent Systems

A Method for Validating and Discovering Associations between Multi-level Emergent Behaviours in Agent-Based Simulations 1
Chih-Chun Chen, Sylvia B. Nagl, and Christopher D. Clack

Background Sensing Control for Planning Agents Working in the Real World 11
Hisashi Hayashi, Seiji Tokura, Fumio Ozaki, and Miwako Doi

Real-Time Neural Networks Application of Micro-Electroforming for Different Geometry Forms 21
Sheau-Wen Shiah, Pai-Yu Chang, Tzeng-Yuan Heh, and Po-Hung Lin

On Drawing Conclusions in Presence of Inconsistent Data 32
Sylvia Encheva and Sharil Tumin

Basic Concepts of Knowledge-Based Image Understanding 42
R. Tadeusiewicz and P.S. Szczepaniak

KOSPI Time Series Analysis Using Neural Network with Weighted Fuzzy Membership Functions 53
Sang-Hong Lee and Joon S. Lim

Facilitating MAS Complete Life Cycle through the Protégé-Prometheus Approach 63
Marina V. Sokolova and Antonio Fernández-Caballero

Holonic Multi-agent System Model for Fuzzy Automatic Speech / Speaker Recognition 73
J.J. Valencia-Jiménez and Antonio Fernández-Caballero

Social Semantic Cloud of Tag: Semantic Model for Social Tagging 83
Hak-Lae Kim, John G. Breslin, Sung-Kwon Yang, and Hong-Gee Kim

Realization of an Intelligent Frog Call Identification Agent 93
Chenn-Jung Huang, Yi-Ju Yang, Dian-Xiu Yang, You-Jia Chen, and Hsiang-Yu Wei

Towards a Distributed Fuzzy Decision Making System 103
Kevin K.F. Yuen and H.C.W. Lau

A Color Set Web-Based Agent System for 2-Dimension Emotion Image Space 113
Pei-Lin Chen and Jen Yen

Approach to Solving Security Problems Using Meta-Agents in Multi Agent System 122
Esmiralda Moradian and Anne Håkansson

A Software Engineering Process for BDI Agents 132
Aaron Hector, Frans Henskens, and Michael Hannaford

Design Pattern Based Development Methodology and Support Tool for Multi Agent System 142
Hyunsang Youn and Eunseok Lee

Moving Object Detection and Classification Using Neural Network 152
M. Ali Akber Dewan, M. Julius Hossain, and Oksam Chae

Data Integration in a System with Agents' Models 162
Grażyna Brzykcy

Main Track 2: Agent-Oriented Web Applications

A Comparative Study of Two Short Text Semantic Similarity Measures 172
James O'Shea, Zuhair Bandar, Keeley Crockett, and David McLean

A User Interface for a Jini-Based Appliances in Ubiquitous Smart Homes 182
Seong Joon Lee, Won Tae Kim, and Kwang Seon Ahn

A Study of Meta-Search Agent Based on Tags and Ontological Approach for Improving Web Searches 192
Jae-Youn Jung, Hyeon-Cheol Zin, and Chonggun Kim

An Agent Using Matrix for Backward Path Search on MANET 203
SungSoo Lee, ZhiPeng Liu, and ChongGun Kim

A Personalized URL Re-ranking Methodology Using User's Browsing Behavior 212
Harshit Kumar, Sungjoon Park, and Sanggil Kang

Library Services as Multi Agent System 222
Toshiro Minami

Main Track 3: Ontology Management

Ontology Methodology Based Context Awareness Using System Entity Structure for Network Analysis 232
Taekyu Kim, Young-Shin Han, Tae-young Kim, and Jong-Sik Lee

A Scalable Framework for Distributed Ontologies 242
Ngot Phu Bui, Young-Tack Park, and TaeChoong Chung

Bridging Semantic Gap in Distant Communication: Ontology-Based Approach 252
Nilar Aye, Fumio Hattori, and Kazuhiro Kuwabara

Using Meta-agents to Reason with Multiple Ontologies 261
Ronald Hartung and Anne Håkansson

Adaptive Information Provisioning in an Agent-Based Virtual Organization—Ontologies in the System 271
Michal Szymczak, Grzegorz Frackowiak, Maciej Gawinecki, Maria Ganzha, Marcin Paprzycki, Myon-Woong Park, Yo-Sub Han, and Y.T. Sohn

Main Track 4: Multi-agent Resource Allocation

Resource and Remembering Influences on Acquaintance Networks 281
Chung-Yuan Huang, Chia-Ying Cheng, and Chuen-Tsai Sun

A Formalization for Distributed Cooperative Sensor Resource Allocation 292
Toshihiro Matsui and Hiroshi Matsuo

Ant Colony Programming with the Candidate List 302
Mariusz Boryczka

A Rough Set Approach on Supply Chain Dynamic Performance Measurement 312
Pei Zheng and Kin Keung Lai

Analyzing Knowledge Exchanges in Hybrid MAS GIS Decision Support Systems, Toward a New DSS Architecture 323
D. Urbani and M. Delhom

Multi-agent System for Custom Relationship Management with SVMs Tool 333
Yanshan Xiao, Bo Liu, Dan Luo, and Longbing Cao

Main Track 5: Cooperation, Coordination, and Teamwork

A Model for Fuzzy Grounding of Modal Conjunctions in Artificial Cognitive Agents 341
Radosław Piotr Katarzyniak, Ngoc Thanh Nguyen, and Lakhmi C. Jain

An Integrated Argumentation Environment for Arguing Agents 351
Taro Fukumoto, Syuusuke Kuribara, and Hajime Sawamura

Agent Collaboration for Multiple Trading Strategy Integration 361
Longbing Cao, Dan Luo, Yanshan Xiao, and Zhigang Zheng

Strategy Acquisition on Multi-issue Negotiation without Estimating Opponent's Preference 371
Shohei Yoshikawa, Yoshiaki Yasumura, and Kuniaki Uehara

Predictor Agent for Online Auctions 381
Deborah Lim, Patricia Anthony, and Chong Mun Ho

Discovering Behavior Patterns in Muti-agent Teams 391
Fernando Ramos and Huberto Ayanegui

Fuzzy Logic for Cooperative Robot Communication 401
Dingyun Zhu and Tom Gedeon

An Approach to Event-Driven Algorithm for Intelligent Agents in Multi-agent Systems 411
Anne Håkansson and Ronald Hartung

Agvs Distributed Control Subject to Imprecise Operation Times 421
Grzegorz Bocewicz, Robert Wójcik, and Zbigniew Banaszak

On the Multi-threading Approach of Efficient Multi-agent Methodology for Modelling Cellular Communications Bandwidth Management 431
P.M. Papazoglou, D.A. Karras, and R.C. Papademetriou

Main Track 6: Agents for Network Management

Design of the Architecture for Agent of the Personality Security in the Ubiquitous Environment 444
Eun-Ser Lee and Sang Ho Lee

The Multi Agent System Solutions for Wireless Sensor Network Applications 454
Khin Haymar Saw Hla, YoungSik Choi, and Jong Sou Park

A Study of Implanted and Wearable Body Sensor Networks 464
Sana Ullah, Henry Higgin, M. Arif Siddiqui, and Kyung Sup Kwak

Complexity of Road Networks as Agents' Environments: Analyzing Relevance between Agent Evaluations and Their Environments 474
Kazunori Iwata, Nobuhiro Ito, Youhei Kaneda, and Naohiro Ishii

Resource Limitations, Transmission Costs and Critical Thresholds in Scale-Free Networks 485
Chung-Yuan Huang, Chia-Ying Cheng, and Chuen-Tsai Sun

Main Track 7: Multi-agent Learning

Temporal Difference Learning and Simulated Annealing for Optimal Control: A Case Study 495
Jinsong Leng, Beulah M. Sathyaraj, and Lakhmi Jain

Separating Learning as an Aspect in Malaca Agents 505
M. Amor, L. Fuentes, and J.A. Valenzuela

Learning Semantic Web from E-Tourism 516
Waralak V. Siricharoen

Incremental Biomedical Ontology Change Management through Learning Agents 526
Arash Shaban-Nejad and Volker Haarslev

An A-Team Approach to Learning Classifiers from Distributed Data Sources 536
Ireneusz Czarnowski, Piotr Jędrzejowicz, and Izabela Wierzbowska

The User Preference Learning for Multi-agent Based on Neural Network in Ubiquitous Computing Environment 547
Eungyeong Kim, Hyogun Yoon, Malrey Lee, and Thomas M. Gatton

Main Track 8: Mobile Agents

Performance Evaluation of Mobile Grid Services 557
Sze-Wing Wong and Kam-Wing Ng

Are You Satisfied with Your Recommendation Service?: Discovering Social Networks for Personalized Mobile Services 567
Jason J. Jung, Kono Kim, Hojin Lee, and Seongrae Park

SETNR/A: An Agent-Based Secure Payment Protocol for Mobile Commerce 574
Chung-Ming Ou and C.R. Ou

A Key Management Scheme Combined with Intrusion Detection for Mobile Ad Hoc Networks 584
Yingfang Fu, Jingsha He, Liangyu Luan, Guorui Li, and Wang Rong

Agent Based Formation and Enactment of Adaptable Inter-organizational Workflows in Virtual Organizations 594
Habin Lee, Hyung Jun Ahn, Hanseup Kim, John Shepherdson, and Sung Joo Park

Main Track 9: Agents in Applications

The Impact of Information Sharing Mechanism to Geographic Market Formation 604
Xingang Xia, Hua Sun, Guo Liu, and Zhangang Han

Agent System for Online Ticket Resale 614
Jae Hyung Cho, Hyun Soo Kim, Hyung Rim Choi, and Jae Un Jung

Application of Ubiquitous Computing Technology to the Steel-Plate Piling Process of Ship Construction 624
Hoon Oh and Jeong Seok Heo

The Choquet Integral Analytic Hierarchy Process for Radwaste Repository Site Selection in Taiwan 634
Chen Lin

Contribution of Hypercontours to Multiagent Automatic Image Analysis 647
P.S. Szczepaniak

Bayesian Communication Leading to Nash Equilibrium Through Robust Messages - p-Belief System Case - 654
Takashi Matsuhisa

Manager-Agent Framework of a Medical Device Communication for u-Healthcare Services with USN 664
Yung Bok Kim and Jae-Jo Lee

An Intelligent Multi Agent Design in Healthcare Management System 674
Dharmendra Sharma and Fariba Shadabi

A Middleware Platform Based on Multi-Agents for u-Healthcare Services with Sensor Networks 683
Yung Bok Kim, Chul-Su Kim, and Jun Wook Lee

Doctoral Track

Measuring Similarity of Observations Made by Artificial Cognitive Agents 693
Grzegorz Popek and Radosław Piotr Katarzyniak

An Algorithm for Agent Knowledge Integration Using Conjunctive and Disjunctive Structures 703
Trong Hieu Tran and Ngoc Thanh Nguyen

A Note on Root Choice for Parallel Processing of Tree Decompositions 713
Yueping Li and Yunting Lu

An Agent's Activities Are Controlled by His Priorities 723
Huiliang Zhang, Shell Ying Huang, and Yuming Chang

Mining Based Decision Support Multi-agent System for Personalized e-Healthcare Service 733
Eunyoung Kang, Hee Yong Youn, and Ungmo Kim

A Multi-agent Mechanism in Machine Learning Approach to Anti-virus System 743
Minh Nhat Quang Truong and Trong Nghia Hoang

Data-Mining Model Based on Multi-agent for the Intelligent Distributed Framework 753
Romeo Mark A. Mateo, Insook Yoon, and Jaewan Lee

Recalling the Embodied Meaning of Modal Conjunctions in Artificial Cognitive Agents 763
Wojciech Lorkiewicz and Radosław Piotr Katarzyniak

Performance Evaluation of Multiagent Systems: Communication Criterion 773
Faten Ben Hmida, Wided Lejouad Chaari, and Moncef Tagina

Using Multiple Models to Imitate the YMCA 783
Axel Tidemann

Special Session: Self-organization in Multi-agent Systems

A Comparison of Performance-Evaluating Strategies for Data Exchange in Multi-agent System 793
Dariusz Król and Michał Zelmozer

The Emergence and Collapse of the Self Monitoring Center in Multi-agent Systems 803
Asaki Nishikawa

The Design of Model Checking Agent for SMS Management System 813
Ali Selamat and Siti Dianah Abdul Bujang

An Intelligent Agent-Based System for Multilingual Financial News Digest 822
James N.K. Liu and M.K. Ho

An Agent-Based Diabetic Patient Simulation 832
Sara Ghoreishi Nejad, Robert Martens, and Raman Paranjape

A Multi-agent System for Computer Network Security Monitoring 842
Agnieszka Prusiewicz

Special Session: Intelligent and Secure Agent for Digital Content Management

Vulnerabilities in a Remote Agent Authentication Scheme Using Smart Cards 850
Youngsook Lee, Junghyun Nam, and Dongho Won

A Security Enhancement of the E0 Cipher in Bluetooth System 858
HyeongRag Kim, HoonJae Lee, and SangJae Moon

Study of Optimal Traffic Information Using Agents Techniques 868
You-Sik Hong, CheonShik Kim, and Geuk Lee

A Recommendation Using Item Quality 878
Sung-hoon Cho, Moo-hun Lee, Bong-Hoi Kim, and Eui-in Choi

Secure Storage and Communication in J2ME Based Lightweight Multi-Agent Systems 887
Syed Muhammad Ali Shah, Naseer Gul, Hafiz Farooq Ahmad, and Rami Bahsoon

A New Key Agreement Protocol Based on Chaotic Maps 897
Eun-Jun Yoon and Kee-Young Yoo

Author Index 907

A Method for Validating and Discovering Associations between Multi-level Emergent Behaviours in Agent-Based Simulations

Chih-Chun Chen[1], Sylvia B. Nagl[2], and Christopher D. Clack[3]

[1] Department of Computer Science, University College London
[2] Department of Oncology and Biochemistry, University College London
[3] Department of Computer Science, University College London

Abstract. Agent-based models (ABM) and their simulations have been used to study complex systems with interacting entities and to model multi-agent systems. Simulations are used to explore the dynamic consequences of these models. In many cases, the behaviours that are of interest are emergent ones that arise as a result of interactions between agents rather than the actions of any individual agent. In this paper, we propose a formalism for describing emergent behaviours at any level of abstraction based on the idea that event types can be defined that characterise sets of behavioural 'motifs'. This provides the basis for a method for studying the associations between multi-level behaviours in simulations. There are two categories of hypotheses that we seek to address with respect to an ABM and its simulations:

– Hypotheses concerned with associations between emergent behaviours defined at various levels of abstraction.
– Hypotheses concerned with the links between parameter sensitivity / initial conditions and emergent behaviours e.g. the ABM is sensitive to a parameter x because x predisposes the system or part of the system to exhibit a particular (emergent) behaviour.

Keywords: agent-based modelling, emergence, complex systems, multi-agent systems.

1 Introduction

Agent-based modelling (ABM) and simulation is a widely used technique for studying studying complex systems. Each agent in the simulation represents an instance of an entity that is being modelled, while different agent types in the ABM can be seen as representing different 'species' of entities. Agent types have specifications which determine the behaviour of an agent given its own state and/or the state of its environment, where an agent's environment can consist of other agents, objects (e.g. coordination artifacts), and/or continuous media. Complex systems are those in which entities interact in a non-trivial fashion

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 1–10, 2008.

so that behaviours at different levels of abstraction can influence one another. The usual reductionist cause-effect model therefore can not be applied, since 'causality' can be at different levels. This has been formalised in information-theoretic interpretations of complexity and emergence e.g. [12], [2], [14], where dynamics at lower resolutions have high predictive efficiency so as to make higher resolution states statistically redundant.

In many cases, ABM and their simulations are used to investigate the behaviours that emerge at the whole system or global level given the set of rules defined at the agent level in the agent types. By observing that a particular system level behaviour emerges, we can say that the set of rules in the ABM are sufficient to generate the emergent behaviour. Sensitivity of the ABM to certain parameters can also be investigated with multiple simulations, where non-linear differences between simulations with different parameters reveal sensitivity.

However, hypotheses about the mechanisms underlying a particular system behaviour are often expressed in natural language and validated through visualising the simulation and observing the interactions. We seek to formalise this process so that such hypotheses can be formally expressed and validated computationally rather than by human observation.

At the same time, Systems Sciences e.g. Systems Biology seek to integrate several levels of abstraction in a single model. While there are formalisms for specifying hierarchically organised systems e.g. P-Systems [9], hierarchical petri nets [5], there has been little work on using such formalisms to represent spontaneously arising hierarchies (those which are emergent). This requires a more general theory, which we provide here using our formalism. By being able to formally describe behaviours at different levels of abstraction e.g. different resolutions, we can then detect these behaviours in simulation and validate hypotheses about the relationships between these multi-level behaviours (with system-level behaviours being at the highest level of abstraction). The methdology we propose treats each simulation as an individual in the population of simulations that can be generated by an ABM and uses sammping across simulations to determine associations between behaviours at different levels. Furthermore, such associations can give us insight into other properties of an ABM such as sensitivity to particular parameters.

The paper will be organised as follws:

- Section 2 describes our conception of models and simulations.
- Section 3 briefly introduces the theory of emergence on which our work is based, referring to recent work in complexity science to formalise emergent properties.
- Section 4 introduces the complex event formalism for describing multi-level behaviours in multi-agent simulations and proposes a methodology for determining associations between behaviours at different levels of abstraction based on this formalism.
- Section 5 summarises and concludes the paper.

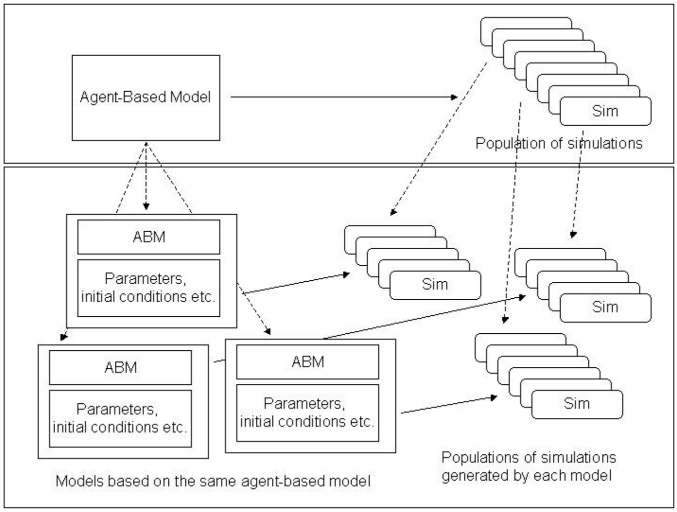


Fig. 1. An agent-based model (ABM) generates a population of simulations, Sim_{Total}. Different parametisations/initial conditions etc. of the ABM can be regarded as different models $X, Y...$, each of which generate a population of simulations $Sim_X, Sim_Y ... \in Sim_{Total}$, which are subsets of the population of simulations generated by the ABM.

2 Models, Simulations, Agents and Agent Types as Functions

Central to our approach is the idea that an agent-based model (ABM) is a function which, given a set of arguments (e.g. parameter values, initial agent positions and states), returns a simulation. Simulation parameters can be treated as constraints that define subtypes of the ABM (i.e. more specific models). These ideas are illustrated in Figure 1.

We define an agent-based model as a function which takes the arguments: $Agents$, a set of agents and $Config$, a configuration defining the initial conditions (e.g. where each agent is situated, global and local variables etc.), and returns the simulation:

$$ABM \;\; Agents \;\; Config = Sim$$

where $Agents$ and $Config$ are themselves defined as functions, each agent (instance) $a_0, ..., a_n$ in $Agents$ is a function, and each member of $Config$ is a function $c_0, ..., c_n$ that determines the initial state of an agent and its position with respect to the other agents. Environmental objects, shared data spaces, global and local variables etc. are all members of $Agents$ in this formalisation.

ABMs can also be parametised so that given a set of parameters P, there are a set of simulations that satisfy the model i.e. there is a function $Model$ for which:

$$Model \;\; ABM \;\; P(Sim)$$

returns true if the simulation belongs to the set of simulations that can be generated by $ABM \times P$ and false otherwise.

The ABM is made up of a set of functions called agent types $A_0, ..., A_n$ that determine which agents (instances) can be generated by the model so that each A_i returns an agent when given a c_i:

$$A \ \ c = a$$

An agent type is a function defining the behaviour of an agent given its own state Q_a and the state of its environment or neighbourhood Q_e (which might istelf be made up of other agents' states). Usually, this is expressed as a set of rules or constraints. At a particular level of abstraction, an agent's behaviour can be viewed as an atomic state transition or event. We call these events *simple events* (se):

$$A \ \ Q \ \ O = se,$$

where

$$Q = Q_a \times Q_e,$$

and O is a function defining the level of abstraction at which the event is observed.

3 Emergence and Emergent Properties

The term 'emergence' is often used in complex systems discourse to refer to phenomena where a property or behaviour arises at some higher level of abstraction from the actions and interactions between lower level entities. However, defining the term has proved difficult, since complex systems are investigated from different perspectives and those working in different disciplines do not always agree on what counts as emergence [3]. However, we can distinguish two main categories of emergence theories:

- Information theoretic interpretations (see [13] for a review), which formalise the fact that subsets of a system's states can be statistically significantly related to one another due to the system's internal dynamics. For example, if a certain set of states is present in a time series, we may be able to predict the occurrence of some other set of states in the future.
- Definitions from the study of multi-agent systems and simulations (e.g. [1], [4], [6]), which tend to characterise emergence in terms of the specified/unspecified aspects of the system. Unspecified properties and behaviours arise from interaction between agents rather than as a consequence of a single agent's actions; these are termed emergent.

The latter of these requires that as well as being at a higher level of abstraction, emergent properties are also those that are in some way 'greater than the sum of their parts'. This has been formalised in language / grammar-based emergence definitions e.g. [4], [6] by requiring that the 'whole' language can not be generated by the same grammar that generates the 'parts' language. For example, Kubik defines a 'whole' language ($L(\bigcup_i P_i)$) and a 'sum of parts' language ($L(P_i)$), where P_i is the set of rules. Emergence corresponds to the case where an array is in $L(\bigcup_i P_i)$ (the whole language) but not in $L(P_i)$ (the sum of parts language):

$$L(\bigcup_i P_i) \supset superposition(L(P_i)) \tag{1}$$

Whereas $L(P_i)$ is derived by putting all the parts together and deriving configurations for every part separately and then putting results together, while $L(\bigcup_i P_i)$ is obtained by putting all parts together and then deriving configurations.

It is also worth clarifying the term 'level of abstraction', which is derived largely from information-theoretic view that different dynamics can emerge from observing a system at different scopes and resolution (as defined in [11]).

The **scope** of a representation of the system is the 'set of components within the boundary between the associated system and its environment' [11][1] (at a given resolution). The scope S of a temporally extended system can be considered to be made up of its temporal scope S_τ, which defines the set of moments of time over which the system is represented and a spatial scope S_x, which defines the set of components whose states are being considered.

Resolution is the number of states that can be distinguished: i.e., given the same scope, a higher-resolution (finer) representation will be able to distinguish a greater number of possible states. Again, there is both a spatial aspect R_x and a temporal aspect R_τ which together define the overall resolution R.

A level of abstraction is then a function of the scope and resolution, where a higher level of abstraction M has a greater scope and/or a lower resolution than a lower level of abstraction μ (see equations (2), (3) and (4)):

$$R_M \leq R_\mu \tag{2}$$

$$S_M \geq S_\mu \tag{3}$$

$$(R_M, S_M) \neq (R_\mu, S_\mu) \tag{4}$$

To summarise, we stipulate that as well as either having lower resolutions, greater scopes, or both, emergent behaviours are those that have not been explicitly specified.

4 Discovering Associations between Behaviours at Different Levels

In the complex event formalism, a behaviour corresponds to an event, where an event is a state transition defined at some level of abstraction. A distinction is drawn between the origin of an event (the set of rules from which it results)

[1] The system's environment is considered to be outside the scope of representation.

and its state transitions. As well as being able to describe state transitions at different levels of abstraction, it is also possible to describe the origins of events at different levels i.e. we can ask for a given event, how many rules were executed to give that event.

This section first introduces the formalism and then describes a method for identifying, quantifying and then determining associations between behaviours at different levels.

4.1 Complex Events

In the ABM and multi-agent systems context, certain events occur due to a firing of a state transition rule (these are also known as agent 'actions'). We call these *simple events* SE. The state transitions that result from simple events can be described at different levels of abstraction i.e. have different scopes. For example, a simple event (from a single rule) where the colour and size of an agent both change can be further decomposed into two simple events - one with the colour state change, the other with the size state change.

A complex event CE is a set of interrelated events, where relationships between events can be defined in any dimensions e.g. time, space:

$$CE \;::\; SE \mid CE_1 \bowtie CE_2 \tag{5}$$

$\bowtie$ denotes the fact that CE_2 satisfies a set of location constraints with respect to CE_1. Conceptually, complex events are a configuration of simple events where the configuration can be defined in a hyperspace that includes time, physical space and any other dimensions. Location constraints define relationships between complex events, which can be represented as a coloured multi-graph, where the coloured nodes stand for event types and coloured edges stand for the different relationship types (sets of location constraints) existing between events. A simulation, therefore, is itself a (usually very big) complex event.

The state transitions of simple events can be described at different levels of abstraction. For example, a rule that causes state changes in components a, b and c can cause simple events $(q_a, q_b, q_c) \rightarrow (q'_a, q'_b, q'_c)$, $(q_a, q_b) \rightarrow (q'_a, q'_b)$..., $q_a \rightarrow q'_a$...etc. We call this the scope of the event. Two simple events e_1 and e_2 in a system are said to be of the same type if (a) e_1 and e_2 result from the same agent rule and (b) the scope of e_1 is identical to the scope of e_2 i.e. for every component in which a state change occurs in e_1, there is a component of the same type in which the same type of state change occurs in e_2. Two complex events CE_1 and CE_2 are said to be of the same type if, for each constituent event $e1$ in CE_1 there is exactly one event $e2$ in CE_2 satisfying the same location constraints, and $e1$ and $e2$ are events of the same type[2]. The scope of a complex event's state transition (the components in which state transitions occur) should be distinguished from the event scope, which is the minimal set of simple events of which the complex event is composed, i.e. the number of rule firings.

[2] See [7] for a formal definition of types.

To specify a complex event type, we need to specify the types for each of the constituent events and the location constraints that hold between them. Depending on the semantic and expressivity requirements of different applications, it is possible to use formal expressions from different formal systems such as temporal logics to specify the location constraints (good reviews can be found in [8] and [10]). Complex event types can differ in specificity, with a fully determined complex event type CET_{Full} being defined as one whose constituent events are in a fully determined configuration. A partially determined complex event type CET_{Part} is then one with a partially determined configuration and therefore defines a set of complex events with fully determined configurations.

$$CET_{Part} = \{CET_{Full}\} \tag{6}$$

So a complex event type can be defined as either a simple event type, two complex event types with a defined relationship, or a set of complex event types:

$$CET \;\; :: \;\; SET \;\mid\; CET_1 \bowtie CET_2 \;\mid\; \{CET\} \tag{7}$$

The dimensions in which configurations are not fully specified lower the resolution of the complex event, with weaker constraints (greater ranges of possible values) implying a lower resolution in that dimension. More generally, the greater the number of complex event types with fully determined configurations that a complex event type contains, the lower its resolution. Given our definition of an emergent behaviour as one that (as well as being defined at a greater scope or lower resolution) is not explicitly specified in the ABM, we can say that the set of complex events that correspond to emergent behaviours are those that are *not simple events*. Similarly, the complex event types that represent emergent behaviours are those that are *not simple event types*, i.e.:

$$CET_{Emergent} = CET - SET \tag{8}$$

Having briefly introduced the formalism and its relation to emergent behaviours, we now describe how it can be used in the context of multi-agent simulations to determine correlations between behaviours at different levels.

4.2 Correlation Studies to Determine Associations

By determining the degrees to which each complex event type (representing a behaviour at a particular level of abstraction) is observed in simulation, we can calculate the correlations between the complex event types and determine associations between behaviours at different levels. This can be done across all simulations or across classes of simulations that are themselves classified according to some behaviour exhibited (since association may itself be determined by another behaviour that is currently unknown). In other words, we are interested in what makes one simulation of a model exhibit a different behaviour to another simulation of the same model. To determine associations across simulations, the method can be summarised as follows:

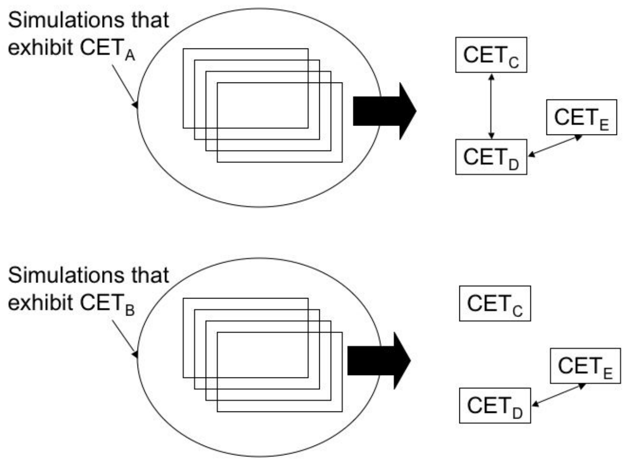


Fig. 2. Simulations with particular behavioural differences may also differ in the associations existing between behaviours at different levels. Double arrows indicate a positive correlation whereas a line connection indicates a negative correlation. In simulations that exhibit CET_A, CET_D is positively correlated with both CET_C and CET_E whereas in simulations that exhibit CET_B, CET_C is only positively correlated with CET_E while CET_E and CET_C are negatively correlated with each other.

1. Specify complex event types $CET_0, CET_1, ..., CET_n$ that correspond to behaviours at different levels of abstraction.
2. Run x simulations of the agent-based model and determine the occurrence levels of each of the complex event types for each simulation $p_{CET0}, p_{CET1}, ...,$ p_{CETn}. This gives x sets of n scores. The simplest occurrence level measure p_{CETi} would be counting the number of times each complex event type occurs through the simulation, but other measures might be more suitable depending on the model and its aims.
3. Calculate the correlations between the p_{CETi} scores across the simulations i.e. $r_{CET0CET1}, r_{CET0CET2}, ...r_{CETnCETn-1}$ and determine significance.

It is also possible that simulations that exhibit different behaviours also have different correlation patterns. To investigate this possibility, the same method can be applied but instead of grouping all simulations together when calculating the r values, we can have different sets of r values corresponding to the sets of simulations exhibiting different behaviours (see Figure 2).

We should emphasise here that associations between behaviours at different levels across simulations do not always indicate causal relationships; nor do they translate into 'laws'. Instead, each simulation should be seen as a unique individual which may or may not have the same associations. For example, even if there is a positive correlation between CET_A and CET_B across simulations, it does not follow that a given simulation with a high degree of CET_A will exhibit a correspondingly high degree of CET_B (this holds for all correlation relationships e.g. we can not infer from the fact that high intelligence scores in childhood IQ tests tend to be associated with higher income that a particular individual with high intelligence scores will definitely have a higher income).

5 Summary and Conclusions

In this paper we have introduced a method for studying associations between emergent behaviours at different levels of abstraction in agent-based simulations. This is based on a population view of agent-based simulations, where each simulation is treated as an individual but can also be multiply classified as belonging to more specific models e.g. having the same parameter settings, initial conditions, or according to the features/behaviours it exhibits.

The complex event formalism allows us to specify the emergent behaviours that we wish to identify in a simulation so that they can be detected computationally. This formalism, together with our treatment of simulations as multiply classifiable indivdiuals, allows us to formulate and validate two catgeories of hypotheses:

- Hypotheses concerned with associations between emergent behaviours defined at various levels of abstraction.
- Hypotheses concerned with the links between parameter sensitivity / initial conditions and emergent behaviours e.g. the ABM is sensitive to a parameter x because x predisposes the system or part of the system to exhibit a particular (emergent) behaviour.

We are also able to analyse behaviours in terms of the rules that give rise to them. A distinction is drawn between:

- Simple events, which represent behaviours arising purely from the single execution of an agent rule and are hence explicitly specified.
- Complex events, which represent behaviours arising from execution of more than one agent rule (these can be from the same agent) i.e. the interactions between rules.

This distinction is based on established definitions of emergence in the field of multi-agent systems engineering. The decomposition of complex events gives us a means of determining the rule-based mechanisms underlying an emergent behaviour, something that would be of considerable benefit in the design and engineering of multi-agent systems.

References

1. Bonabeau, E., Dessalles, J.L.: Detection and emergence. Intellectica 2(25), 85–94 (1997)
2. Crutchfield, J.P., Feldman, D.P.: Regularities unseen, randomness observed: Levels of entropy convergence. Chaos 13(1), 25–54 (2003)

3. Deguet, J., Demazeau, Y., Magnin, L.: Elements about the emergence issue: A survey of emergence definitions. ComPlexUs 3, 24–31 (2006)
4. Demazeau, Y.: Steps towards multi-agent oriented programming. In: First International Workshop on Multi Agent Systems, Boston, Mass. (1997)
5. Fehling, R.: A concept of hierarchical petri nets with building blocks. Lectore Notes in Computer Science; Advances in Petri Nets 1993 674, 148–168 (1991)
6. Kubik, A.: Toward a formalization of emergence. Artificial Life 9, 41–66 (2003)
7. Lawvere, F.W., Schamiel, S.H.: Conceptual mathematics: a first introduction to categories. Cambridge University Press, Cambridge (1997)
8. Moller, F., Birtwistle, G.: Logics for concurrency: Structure versus automata. LNCS, Vol. 1043, Springer, Heidelberg (1996)
9. Paun, G.: From cells to computers: computing with membranes (p systems). Biosystems 59, 139–158 (2001)
10. Rouff, C.A., Hinchey, M.G., Rash, J., Truszkowski, W., Gordon-Spears, D. (eds.): Agent technology from a formal perspective. Springer-Verlag, London (2005)
11. Ryan, A.: Emergence is coupled to scope, not level. Nonlinear Sciences (2006)
12. Shalizi, C.: Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata. PhD thesis, University of Michigan (2001)
13. Shalizi, C.R.: Methods and Techniques of Complex Systems Science: An Overview, chapter Methods and Techniques of Complex Systems Science: An Overview, pp. 33–114. Springer, New York (2006)
14. Shalizi, C.R., Shalizi, K.L.: Blind construction of optimal nonlinear recursive predictors for discrete sequences. In: Chickering, M., Halpern, J.J. (eds.) Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference, AUAI Press (2004)

Background Sensing Control for Planning Agents Working in the Real World

Hisashi Hayashi, Seiji Tokura, Fumio Ozaki, and Miwako Doi

Corporate Research & Development Center, Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582 Japan

Abstract. Online planning agents can adapt to the dynamic environment by continuously modifying plans during the plan execution. However, robotics agents working in the real environment cannot obtain necessary information from the outer world by simply turning all sensors on while executing a plan. This paper shows a new background sensing control method by which planning agents can effectively observe the real environment and obtain important information when necessary during the plan execution.

1 Introduction

One of the important functions of autonomous agents is to create intentions and commit to them. Even if autonomous agents receive a command from the user or another agent, they decide whether or not to execute the goal, and sometimes they do not commit to it. Moreover, even if the goal is not given, autonomous agents create goals based on the situation.

Most BDI (Belief, Desire, and Intention) agents [2,3,9,11,12], which are based on PRS [7], conduct reactive rule-based reasoning to create their intentions. These agents use rules of the form:

if *conditions* then *commands*.

In these rules, it is possible to write conditions about the belief (or working memory). In addition, it is sometimes possible to write the conditions about the events that trigger the rules. These rules are evaluated when an event happens or the belief is changed. When the condition of a rule is satisfied, the agent executes the commands following the rule. These commands include actions that change the environment, creation of subgoals, and belief updates. Because these BDI agents immediately execute the ready-made commands, they are regarded as reactive autonomous agents.

On the other hand, given the goal, deliberative agents make plans before acting. In particular, online planning agents [1,8,15] can adapt to the dynamic environment by continuously modifying the plan during the execution. Online planning is effective to achieve the goal in the dynamic environment when the simple reaction is insufficient and deliberation is necessary. By creating goals through if-then rules and achieving the goals via online planning, it is possible

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 11–20, 2008.

to make deliberative autonomous agents [5,16]. However, no matter how good the online planners are, they are useless if they cannot observe the world. In the agent simulation [4,10], it might be easy to observe the environment. In the real world, the agent cannot recognize important events by simply turning all the sensors on. They have to actively sense the world and analyze the situation. If the agent tries to observe an event, it might be impossible to detect other events or to execute some actions simultaneously. Therefore, it is important to decide what event the agent needs to recognize. This paper shows a background sensing control method for autonomous planning agents working in the real world.

The rest of this paper is organized as follows. Section 2 informally explains our background sensing control using the mobile robot scenario. Section 3 presents our agent architecture. Section 4 shows how to implement the mobile robot scenario. Section 5 reports the experimental results of the mobile robot scenario. Section 6 discusses related work. Section 7 is the conclusion.

2 Scenario

This section informally explains how our background sensing control works using a mobile robot scenario. Figure 1 shows the map of the environment where the robot moves from one node to the next node along an arc. The robot has a camera, and it can localize itself by looking at the marker which is placed at a node. After finding the marker of a node, the robot can move to the node.

Suppose that the robot is initially at `node10`. When the robot needs to go to `node4`, the initial route would be:

`node10` → (`arc10`) → `node8` → (`arc8`) → `node7` → (`arc7`) → `node6` → (`arc5`) → `node4`

While moving to `node10` along `arc10`, suppose that there is an obstacle. If the robot turns the sonar sensors on, the robot can go around the obstacle until the robot finds no obstacle towards `node8`.

If the obstacle is small, the robot can easily avoid it and go to `node8`. However, what if the obstacle is big? In this example, while avoiding the obstacle, the robot gets closer to `node7`. If the robot searches markers while avoiding the obstacle, it might find the marker of `node7`. In that case, the robot can directly move to `node7`. Because the distance of the route becomes shorter, it is better to change the route and go directly to `node7` without going through `node8`.

Normally, the robot does not try to find a marker while avoiding an obstacle. Or the robot might try to find the marker of the next destination (`node8`). In either case, the robot cannot find the marker of `node7` while avoiding the obstacle. Therefore, it is important to set the target of background observation according to the situation while executing a plan.

3 Agent Architecture

Figure 2 shows the overall agent architecture. The agent is composed of two subagents: the event-handling agent and the planning agent. Both of these subagents

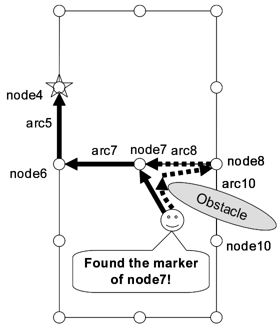


Fig. 1. The Map

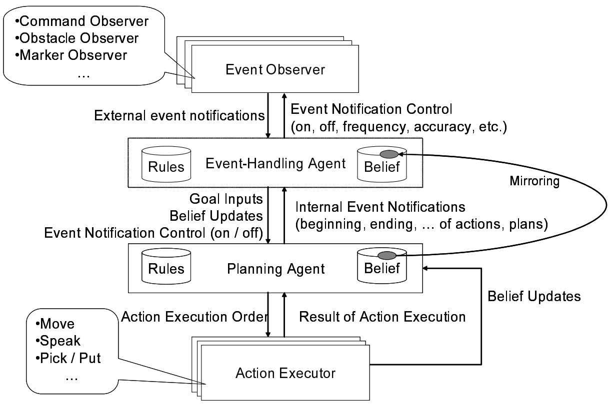


Fig. 2. The Agent Architecture

have their own beliefs that are described by fluents (time-varying predicates) and horn-clauses. The event-handling agent uses its belief for reactive reasoning, and the planning agent uses its belief for planning. Some fluents in the belief of the planning agent are mirrored to the belief of the event-handling agent. Although these two agents have separate beliefs, it is possible to use one common belief instead. Event observers and action executors are respectively plugged into the event-handling agent and the planning agent. Event observers declare the events they can observe to the event-handling agent. Similarly, action executors declare the actions they can execute to the planning agent.

The event-handling agent handles two kinds of events: external events and internal events. Event observers notify external events to the event-handling agent. The planning agent notifies internal events to the event-handling agent when beginning, ending, suspending, or resuming an action or a goal. The event-handling agent can specify the timing of event notifications, types of events to notify, and other conditions of event notifications.

When the event-handling agent receives an event, it conducts rule-based reasoning using the event rules of the form:

$\texttt{eventRule}(E, [L_1, ..., L_n], [C_1, ..., C_m])$

where E is an event, each L_i $(0 \leq i \leq n)$ is either a fluent or the negation $not(F)$ of the fluent F, and each L_i $(1 \leq i \leq m)$ is a command. Because we implement the interpreter in Prolog, the event rules are written as Prolog facts, and unification is available as usual. This event rule means that

> When the event E is notified, if each of the conditions $L_1, ..., L_n$ $(n \geq 0)$ hold, then sequentially execute the commands from C_1 to C_m $(m \geq 1)$.

When the conditions of more than one event rule hold, the agent executes only the commands specified by the first rule. The commands include belief updates of the event-handling agent, belief updates of the planning agent, goal inputs to the planning agent, and event observation controls such as switch-on /off of event observations, frequency of event notifications, accuracy of observation, and other condition settings. From the viewpoint of background sensing control, event observation controls are the most important commands.

The planning agent called Dynagent makes plans using its belief and rules for planning. When planning, Dynagent searches for the best plan comparing the costs of alternative plans. In order to execute an action in a plan, it selects an appropriate action executor and tells the action executor to execute the action. During the plan execution, every time an action is executed or the belief is updated, it dynamically modifies the plan. The details of Dynagent are described in [8] and this paper provides no further explanation of the planning algorithm.

4 Scenario Implementation

This section shows how to write the event rules to implement the scenario in Section 2. Initially, in the belief of the event-handling agent, only the following fluent is recorded:

```
planningAgent(planningAgent)
```

This means that the name of the planning agent is `planningAgent`. When initializing, the event-handling agent creates the following event:

```
initialize
```

This event triggers the following event rule:

```
eventRule(initialize,[planningAgent(PA)],[
  switchOn(beginAction(_),[]),
  switchOn(resumeAction(_),[]),
  switchOn(succeedAction(_),[]),
  switchOn(suspendAction(_),[]),
  subscribeFluentFrom(foundMarker(_),PA)]).
```

By executing these `switchOn` commands, the event-handling agent tells the planning agent to notify events when beginning, ending, suspending, and resuming actions. By executing the `subscribeFluentFrom` command, any fluent that is unifiable with:

```
foundMarker(_)
```

will be mirrored from the belief of the planning agent to the belief of the event-handling agent.

The robot is initially at `node10`. Suppose that an event observer receives the command to go to `node4` from the user. Then the event observer notifies the following event to the event-handling agent:

```
command(goto(node4)).
```

This event triggers the following event rule:

```
eventRule(command(Goal),[planningAgent(PA)],[
  assertGoal(PA,Goal)]).
```

Following this event rule, the event-handling agent inputs the following goal to the planning agent:

```
goto(node4)
```

To achieve this goal, the planning agent makes the following plan:

```
[gotoNextNode(arc10,node8,2,2),
 gotoNextNode(arc8,node7,1,2),
 gotoNextNode(arc7,node6,0,2),
 gotoNextNode(arc5,node4,0,3)]
```

When the planning agent tells an action executor to execute the first action to go to `node8` (the coordination of `node8` is (2,2)) along `arc10`, the planning agent also notifies the following internal event to the event-handling agent:

```
beginAction(gotoNextNode(arc10,node8,2,2))
```

This event triggers the following rule:

```
eventRule(beginAction(gotoNextNode(_,Node,X,Y)),[],[
  assertFluent(self,goingToNode(Node,X,Y)),
  switchOn(objectFound(Node,X,Y),[accuracy(1,meter)])]).
```

Based on this rule, the event-handling agent records the following fluent to its belief:

```
goingToNode(node8,2,2)
```

This fluent means that the robot is moving to `node8`. Then, the event-handling agent tells an event observer to notify the following event when it detects an obstacle towards `node8` within 1 meter.

```
objectFound(node8,2,2)
```

While moving to `node8`, suppose that the robot finds an obstacle ahead within 1 meter by means of sonar sensors. Then, the event observer notifies the following event to the event-handling agent:

```
objectFound(node8,2,2)
```

This event is handled by the following event rule:

```
eventRule(objectFound(Node,X,Y),[goingToNode(Node,X,Y)],[
  assertFluent(self,objectAhead(Node,X,Y)),
  switchOff(objectFound(Node,X,Y)),
  switchOn(objectNotFound(Node,X,Y),[accuracy(1,meter)]),
  switchOn(markerFound(_),[ignore(Node,X,Y)])]).
```

Follwoing this event rule, the event-handling agent adds the following fluent to its belief:

```
objectAhead(node8,2,2)
```

Then, the event-handling agent tells the event observer to stop observing the following event:

```
objectFound(node8,2,2)
```

Instead, the event-handling agent tells the event observer to notify the following event when the obstacle is cleared towards `node8`:

```
objectNotFound(node8,2,2)
```

Using the sonar sensors, the robot reactively avoids the obstacle ahead by moving around the obstacle. While avoiding the obstacle, the robot tentatively moves away from the arc. Therefore, it might find a marker unexpectedly. According to the above event rule, when the robot starts avoiding the obstacle, the event-handling agent tells the event observer to search markers except the marker of the next node.

Suppose that the marker of `node7` is detected while avoiding an obstacle on `arc10`. Then, the event observer notifies the following event to the event-handling agent:

```
markerFound(node7)
```

This event is handled by the following event rule:

```
eventRule(markerFound(Node),[
  not(foundMarker(Node)),
  goingToNode(Node2,X,Y),
  eval(diff(Node,Node2)),
  planningAgent(PA)
],[
  assertFluent(PA,foundMarker(Node)),
  switchOn(markerNotFound(Node),[]),
  checkPlan(PA)]).
```

Following this event rule, the event-handling agent adds the following fluent to the belief of the planning agent to record that the marker of `node7` is found:

```
foundMarker(node7)
```

Note that this fluent is also mirrored to the belief of the event-handling agent. Then, the event-handling agent tells the event observer to notify the following event when the marker of `node7` is not recognized:

```
markerNotFound(node7)
```

By executing `checkPlan(planningAgent)`, the event-handling agent allows the planning agent to change the plan even in the middle of the action execution. Considering the costs of plans, the planning agent changes the plan to the following plan:

```
[gotoNode(node7,1,2),
 gotoNextNode(arc7,node6,0,2),
 gotoNextNode(arc5,node4,0,3)]
```

Because the first action of this plan is different from the current action being executed (`gotoNextNode(arc10,node8,2,2)`), the planning agent tells the action executor to stop the current action execution, and notifies the following internal event to the event-handling agent:

```
suspendAction(gotoNextNode(arc10,node8,2,2))
```

This event triggers the following event rule:

```
eventRule(suspendAction(gotoNextNode(_,Node,X,Y)),[
  planningAgent(PA)
],[
  retractFluent(self,goingToNode(Node,X,Y)),
  retractFluent(self,objectAhead(Node,X,Y)),
  switchOff(markerFound(_)),
  switchOff(objectFound(Node,X,Y)),
  switchOff(objectNotFound(Node,X,Y))]).
```

Following this event rule, the event-handling agent removes the following fluents from its belief:

```
goingToNode(node8,2,2)
objectAhead(node8,2,2)
```

Then, it tells the event observers to stop observing markers and obstacles. Following the new plan, the planning agent tells the action executor to go to `node7`, and notifies the following internal event to the event-handling agent:

```
beginAction(gotoNode(node7,1,2))
```

This event is handled by the following event rule:

```
eventRule(beginAction(gotoNode(Node,X,Y)),[],[
  assertFluent(self,goingToNode(Node,X,Y)),
  switchOn(objectFound(Node,X,Y),[accuracy(1,meter)])]).
```

Following this event rule, the event-handling agent adds the following fluent to its belief to record that the robot is moving to `node7` which is located at (1,2):

```
goingToNode(node7,1,2)
```

Then, it tells the event observer to notify the following event when finding an obstacle towards `node7` within 1 meter:

```
objectFound(node7,1,2)
```

The planning agent continues to execute the new plan in this way.

5 Experiments

This section shows the results of experiments based on the scenario explained in the previous sections. We used the mobile robot called ApriAttenda [17]. The planning agent has the map shown in Figure 1 in its belief. However, as shown in Figure 3, the real environment where the robot moves is smaller because of the limited space. Initially, the robot is at `node10`. The destination of the robot is `node7`. At `node7`, we placed a marker the robot can recognize. The initial route is the following:

`node10` → (`arc10`) → `node8` → (`arc8`) → `node7`

In pattern (a) in Figure 3, there is no obstacle, and the robot moves as planned. In pattern (b) in Figure 3, there is an obstacle on `arc10`, and the robot detects the marker of `node7` while avoiding the obstacle and goes directly to `node7`. In pattern (c) in Figure 3, there is an obstacle on `arc10`, but the robot does not change the route while avoiding the obstacle because the robot is blindfolded and does not detect the marker of `node7`.

We compared the time to reach the goal in these three patterns. Each pattern was tested five times. The results are shown in Figure 4. As a result, in pattern (b), the robot arrived at the destination the fastest. Pattern (b) is 1.5 times as fast as pattern (a) and 2.6 times as fast as pattern (c). These results show the effect of our background sensing control. Note that even if the robot finds a marker, the planning agent will not change the plan unless the new route is better in terms of distance.

6 Related Work

As discussed in [13], the simplest way to monitor the plan execution is to replan in the case of action failure. In this method, the agent needs to check only the

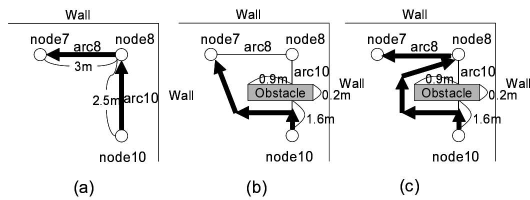
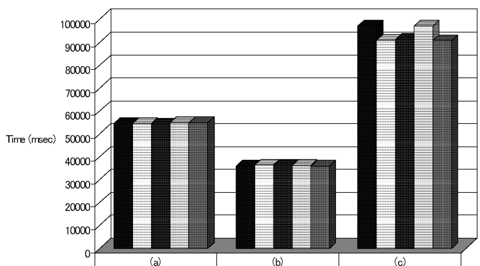


Fig. 3. Experimental Environments



	(a)	(b)	(c)
1	54469	36047	97204
2	54510	36406	90890
3	54328	36141	91204
4	54828	36250	97266
5	54844	35969	90766

Fig. 4. Experimental Results

result of action execution. However, in this method, the agent might continue to execute an invalid plan until the agent actually fails to execute the action.

A more sophisticated way of plan monitoring is to check the preconditions of actions in the plan. In order to check the validity of preconditions, PLANEX [6] uses a triangle table, and many partial-order planners use causal links which were first introduced in NONLIN [14]. In [13], it is pointed out that the first partial-order planner that smoothly integrates planning and execution is IPEM [1]. In these plan monitoring methods, it is possible to check the validity of a plan. However, it is impossible to observe the events which lead to the creation of new valid plans.

There are some agent systems that combine reactive reasoning of BDI agents and planning. Cypress [16] combines the BDI agent of PRS-CL and the HTN planner of SIPE-2. Another agent system [5] combines the BDI agent of Jack and the HTN planner of SHOP. However, these agents cannot adjust event observation according to the plan execution.

7 Conclusions

This paper has presented a new background sensing control method for planning agents working in the real world. By this method, it is possible to recognize useful events from the outer world, changing the target of background observation according to the situation and the plan execution. It is the event-handler that dynamically specifies the events to observe. Therefore, our background sensing control method can be applied to other online planners although we used Dynagent. We have also confirmed by means of experiments that our background sensing control method is effective for the creation of new better plans.

References

1. Ambros-Ingerson, J., Steel, S.: Integrating planning, execution and monitoring. In: AAAI 1988, pp. 735–740 (1988)
2. Braubach, L., Pokahr, A., Lamersdorf, W.: Jadex: A BDI-agent system combining middleware and reasoning. In: Software Agent-Based Applications, Platforms and Development Kits, Birkhiiuser Book, pp. 143–168 (2005)
3. Dastani, M., van Riemsdijk, B., Dignnm, F., Meyer, J.-J.: A programming language for cognitive agents: Goal directed 3APL. In: Dastani, M., Dix, J., El Fallah-Seghrouchni, A. (eds.) PROMAS 2003. LNCS (LNAI), vol. 3067, pp. 111–130. Springer, Heidelberg (2004)
4. Deguchi, H., Tanuma, H., Shimizu, T.: SOARS: Spot oriented agent role simulator - design and agent based dynamical system. In: AESCS 2004, pp. 49–56 (2004)
5. deSilva, L., Padgham, L.: Planning on demand in BDI systems. In: ICAPS 2005 (2005)
6. Pikes, R., Hart, P., Nilsson, N.: Learning and executing generalized robot plans. Artificial Intelligence 3(4), 251–288 (1972)
7. Georgeff, M., Ingrand, F.: Decision-making in an embedded reasoning system. In: IJCAI 1989, pp. 972–978 (1989)
8. Hayashi, H., Tokura, S., Hasegawa, T., Ozaki, F.: Dynagent: An incremental forward-chaining HTN planning agent in dynamic domains. In: Baldoni, M., Endriss, U., Omicini, A., Torroni, P. (eds.) DALT 2005. LNCS (LNAI), vol. 3904, pp. 171–187. Springer, Heidelberg (2006)
9. Howden, N., Rönnquist, R., Hodgson, A., Lucas, A.: JACK: Intelligent agents - summary of an agent infrastructure. In: IAMSMAS 2001(2001)
10. Ishida, T.: Q: A scenario description language for interactive agents. IEEE Computer 35(11), 54–59 (2002)
11. Morley, D., Myers, K.: The SPARK agent framework. In: AAMAS 2004, pp. 712–719 (2004)
12. Rao, A.: AgentSpeak(L): BDI agents speak out in a logical computable language. In: Perram, J., Van de Velde, W. (eds.) MAAMAW 1996. LNCS, vol. 1038, pp. 42–55. Springer, Heidelberg (1996)
13. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs (1995)
14. Tate, A.: Generating project networks. In: LICAI 1977, pp. 888–893 (1977)
15. Wilkins, D.: Practical Planning. Morgan Kaufmann, San Francisco (1988)
16. Wilkins, D., Myers, K., Lowrance, J., Wesley, L.: Planning and reacting in uncertain and dynamic environments. Journal of Experimental and Theoretical Al 7(1), 197–227 (1995)
17. Yoshimi, T., Nishiyama, M., Sonoma, T., Nakamoto, H., Tokura, S., Sato, H., Ozaki, F., Matsuhira, N., Mizoguchi, H.: Development of a person following robot with vision based target detection. In: IROS 2006, pp. 5286–5291 (2006)

Real-Time Neural Networks Application of Micro-Electroforming for Different Geometry Forms

Sheau-Wen Shiah[1], Pai-Yu Chang[2], Tzeng-Yuan Heh[1], and Po-Hung Lin[1]

[1] Chung Cheng Institute of Technology, National Defense University, No. 190, Sanyuan 1st St., Tahsi, Taoyuan, Taiwan 335. R.O.C.
[2] Fortune Institute of Technology, No. 125-8, Ciwun Rd., Cishan Township, Kaohsiung, Taiwan 842. R.O.C.
Sheau-Wen Shiah, Pai-Yu Chang, Tzeng-Yuan Heh, Po-Hung Lin, white_lph@yahoo.com.tw

Abstract. In this study, the approach of using neural networks is implemented for demonstrating its effectiveness in the real-time application of micro-electroform on the different geometry forms. Three back-propagation neural networks are established via the training process with the numerical database to predict the distributions of Sh/Shmax、Cf/Cfmax and I/Imax. Comparisons of the predictions with the test target vectors indicate that the averaged root-mean-squared errors from three back-propagation neural networks are well within 4.15 %. The trained neural networks can verify the prediction capability for agent technology. Then, to fabricate the microstructure of higher surface accurate, higher hardness, lower residual stress and can be duplicated perfectly. Nevertheless, the instant knowledge of micro-electrforming characteristics is practically needed for many industrial agents technology applications.

Keywords: Neural Networks, Real-Time, Micro-Electroforming.

1 Introduction

Neural networks are biologically inspired computing systems with the capability of recognizing patterns through experience, rather than from deduction or recollection of rules. Owing to their flexibility and their ability to perform instantaneous calculations and to develop generalization, neural networks have been applied well beyond the established realism of pattern recognition [1、2]. However, very limited research has been done to instantly predict the micro-electrforming characteristics. Therefore, the purpose of this study is to build up three back-propagation neural networks, trained with the numerical database, for demonstrating their usefulness in the real-time determination of agent technology. Micro-electroforming includes various behaviors of mechanics, chemistry, electricity, and so on these behaviors mutually affected one another during electroforming. First, the effect of internal stress, see relevant research by Yang and Kang [3], indicated that the cladding material induces the phenomena of twist and deformation, also used the second cathode to reduce the internal stress in the Ni-cladding process and increase the uniformity of the Ni-cladding material. Zeller

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 21–31, 2008.

and Lsndau [4] examined micro-electroforming and used pulse current to increase the tension and yielding stress for the alloy of Ni-P. Moreover, Chan et al. [5] used reverse pulse current technology and reduced the internal stress for Ni material. Furthermore, the effect of Current density, DC source is frequently used for micro-electroforming. However, the larger internal stress in the cladding layer causes the surface to be uneven. Using pulse current can boost cladding layer density, increase adhesive force, and reduce internal stress. Lee et al. [6] and Yin et al. [7] used the pulse current for the micro-electroforming of Ni-W alloy, and determined the degree to which the wolfram material is increased by the pulse current. To determine the stirring effect, Andricacos et al. [8] used the ring disk technology for depositing Ni-Fe film in the cathode electrode, increased the convection flow effect, obtained much alloy ion mass diffusion and reduced the proliferation thickness. Furthermore, to measure the current density and the polarization curve on the ring disk. Regarding the influence of electrolyte concentration, cladding layer ability to resist erosion is affected by the electrolyte concentration PH value. Yamasaki et al. [9] noted that chloride ammonium concentration in electrolyte seriously influenced the Ni-W alloys of the cladding layer. When chloride ammonium 0.5 mol/liter was taken, the optimum wolfram content in cladding layer was obtained, while an excessive chloride ammonium concentration limited the wolfram content of the cladding layer. Additionally, selecting different electrolyte also can influence the micro-electroforming module. Gould and Lopez [10] conducted sandwich structure electroforming on SiOx, and found that the gold electrolyte is more stable than aluminum electrolyte.

For the numerical analysis and the discussion of the physical aspects of micro-electroforming are sourced in the last 20 years. Hessami and Tobias [11] simulated the co-deposit of Ni-Fe alloys, and established diffusion and convection equations. Pesco and Cheh [12] demonstrated that electrolyte mass was only affected by diffusion flow for cavities with a high aspect ratio. However, both diffusion mass flow effect and convection mass flow effect existed when the electrolyte flow was laminar. Kondo and Fukui [13] described the width of the cavity more 200 μm, the effect of convection mass transfer governs the behavior of the electrolyte. However, when cavity width is less 100 μm, the effect of diffusion mass transfer is to dominate. Georgiadou [14] took electrolyte 0.002 mol/liter CuSO4+ 0.1 mol/liter H2SO4, and calculated the current density only for the cathode surface. Moreover, Duchanoy and Lapicque [15] displayed that the effect of current density altered the angle of the holes, and observed that the lower Reynolds number condition had a more uniform distribution-cladding layer than the higher Reynolds number condition. About the measurement research by Beck et al. [16] pointed out that the larger hardness values for Ni-Si alloys of cladding layer have obtained.

2 Neural Network

A neural network is agent technology, which employs artificial neurons linked mutually to simulate a biological neural network. An artificial neuron, considering as a simple imitation of a physical neuron, can acquire information from the environment or neurons in the upstream layer and dispatch its calculation result back to the environment or to neurons in the downstream layer. Among a variety of neural network models, back-propagation networks are most widely used in the areas of diagnosis and prediction mechanisms. As a supervised learning network, the back-propagation network is first trained with the training sets, consisting of known input and output vectors. It learns the nonlinear relationship between input and output vectors based on an existing database from theoretical computations and/or experimental measurements. This nonlinear relationship can be expressed in the form of connected weights between nodes and the bias of each node. In this work, the back-propagation-learning rule is used to adjust the weights and biases of networks for minimizing the sum squared error of the network. When certain inexperienced input vectors are presented, trained back-propagation networks can readily predict reasonable output vectors for those new cases.

In this study, three back-propagation networks were adopted to predict the radial profiles of micro-electroforming. The training sets are obtained from numerical database results. To raise the convergence speed, both the momentum term and the adaptive learning rate are adopted in the training process. To appraise the training quality of these back-propagation networks, the predicted micro-electroforming properties are compared with target vectors. The training quality can be then determined using sum-squared error (SEE) or averaged root-mean-squared error (RMS), defined as follows.

$$SSE = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(T_{ij} - P_{ij}\right)^2 \tag{1}$$

$$RMS = \frac{1}{m}\cdot\frac{1}{n}\cdot\sqrt{SSE} \tag{2}$$

Where m, n, T, P stand for the number of training pairs, the number of output nodes, the target value, and the predicted value by the associated neural network.

3 Theoretical Analysis

As figure 1, illustrates two-dimensional micro-cavity of micro-electroforming. The contact angle θ was defined between the upstream photo resist wall side and the cathode surface, $\theta=90^o$ represents as the normal etched micro-cavity, $\theta=120^o$, 60^o individually represents as internal concave under cut etched structure and external protrude under cut etched structure. The linear velocity profile was adopted at the inlet and outlet side boundary, and the velocity, concentration and voltage all are constant at the free electrolyte surface. Both the concentration and voltage gradient orthogonal to the inert surface of photo resist are zero. The surface concentration of copper ion on the cathode is zero because of the condition of diffusion control, and the no slip condition for velocity equal zero were used on the photo resist.

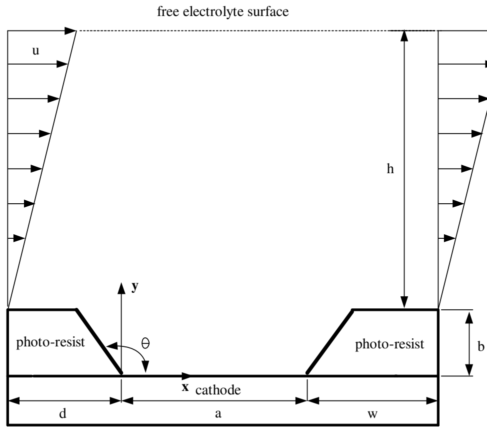


Fig. 1. Physical configuration and coordinate system

3.1 Assumptions

To simplify the computed model, in this article we used the following assumptions: Considering the working fluid is electrolyte 0.6 mol/liter CuSO4 + 1.85 mol/liter H2SO4, which the electrolyte with constant (temperature- independent) properties. The temperature of copper electro-deposition bath is 20^{o}C. A uniform chemical reaction rate is processing for the electrolyte in the micro-electroforming of U-type micro-cavity, and the micro-electroforming process is steady state and laminar flow. The effect of heat diffusion and the compressibility effect in the Momentum equation are disregarded.

3.2 Governing Equations

The electrochemistry reaction of micro-electroforming is to deposit the metal ion or alloy ion for the microstructure. We relate the governing equations or the experience equations in micro-electroforming process to satisfy the behaviors that are the electrode fluid flow, electrode mass flow and electrode voltage flow.

Electrode continuity equation: $\nabla \cdot (\rho U) = 0$ (3)

U= (u, v), velocity components in the X, Y directions, and ρ is electrolyte density.

Electrode momentum equation:
$$(U \cdot \nabla)U = \frac{1}{\rho}\nabla P + \mu \nabla^2 U \tag{4}$$

P is flow local pressure, μ is electrolyte viscosity.
Convection diffusion equation of ion concentration:

$$\nabla \cdot \rho UC = \nabla \cdot J \tag{5}$$

Voltage distributionφmust satisfy Lap lace's equation:

$$\nabla^2 \phi = 0 \tag{6}$$

Using Ohm's law to compute the current density:

$$i = -k\nabla\phi \tag{7}$$

Let i be current density and k be electrolyte conductance coefficient.

3.3 Calculation of Parameters

To determine the domain of evaluation, the fully developed velocity boundary layer thickness δ, can be represented approximately by Schmidt number Sc following from Incropera and DeWitt:

$$\delta / \delta_c \approx Sc^{1/3} = (\nu / D)^{1/3} \tag{8}$$

whereδc is the concentration diffusion layer thickness, υ is the dynamic viscosity coefficient and D is the mass diffusion coefficient. The value of Sc = 64 was calculated by υ, D. Therefore, δc=b and δ=h=4b were shown in the figure 1, the both concentration and velocity of free electrolyte face height were determined. Furthermore, the values of inlet and outlet velocity gradient in the calculated domain are proportional to Peclet number Pe, and which is defined by Kondo and Fukui as

$$Pe = [\delta^2 (du/dy)_{y=0}]/D \tag{9}$$

The bulk concentration Cb=0.5374 mol/liter, incorporated the inlet uniform concentration in the calculated domain, was composed of the 0.6 mol/liter CuSO4 solute and 1.85 mol/liter H2SO4 solution. The value of free electrolyte face voltage ϕ_∞ = 0.002 V was calculated by the following equation

$$\phi_\infty /(\delta + \delta_c) \approx \phi_a / \zeta \tag{10}$$

which the experiment operating voltage of anode ϕ_a is about 2V, and the experiment operating distance between anode with cathode ζ is 5 cm.

The mass fraction Sh/Shmax is evaluated on the basic of cathode surface distribution:

$$\begin{aligned} Sh/Sh_{\max} &= (ka/D)/(ka/D)_{\max} \\ &= k/k_{\max} = (\partial c/\partial y)_{y=0} / [(\partial c/\partial y)_{y=0}]_{\max} \end{aligned} \tag{11}$$

where k is the mass transfer coefficient near the cathode surface, which was shown as

$$k = \lfloor D(\partial c/\partial y)_{y=0} / \Delta C \rfloor \tag{12}$$

The fraction of friction coefficient Cf/Cfmax is evaluated by the shear stress distribution on the hole of the U-type micro-cavity.

$$\frac{C_f}{C_{f_{\max}}} = \frac{\tau_s}{\rho V^2/2} / \left(\frac{\tau_s}{\rho V^2/2}\right)_{\max} \tag{13}$$

The fraction of current density $I/I_{\max}$ also is evaluated by the voltage distribution on the hole of the U-type micro-cavity:

$$I/I_{\max} = \nabla\phi / \nabla\phi_{\max} \tag{14}$$

4 Results and Discussions

In the training stage of back-propagation neural networks, the setting of neural network parameters is a training set contains training pairs. The initial values of the weights and biases of the networks are generated randomly between –1.0 and +1.0. Figure 2 displays typical convergence histories of RMS errors during the training process. In the beginning of the event, the initial RMS errors of three networks are exceedingly high and drop at a much faster rate (within the first 10 iterations) than those in the later phase the event. This is because the initial guesses for these networks are well away from their converged values. As the training error of the Sh/Shmax network decreases to a certain low level, it cannot be further reduced with increasing number of iterations. It normally takes 12,000 iterations to complete the training for each network. On the basis of training pairs, the averaged RMS errors for Sh/Shmax, Cf/Cfmax, and I/Imax networks are 4.15%, 3.28%, and 2.53%, indicating that all three back-propagation networks have been trained fairly well. The higher training error of the Sh/Shmax network can be attributed to the constraint of inherently complicated Sh/Shmax profiles arousing from the streamline for different geometry phenomenon.

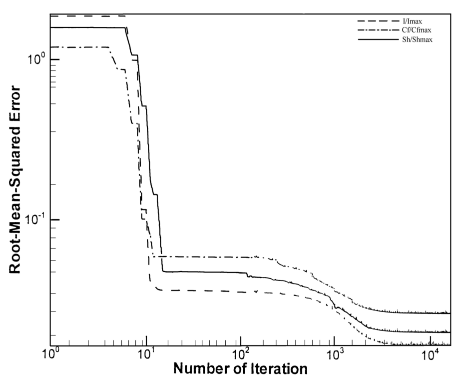


Fig. 2. Convergence histories of root-mean-squared errors during training phase

Figure 3 illustrates the streamlines of micro-cavity micro-electroform for Cb=0.5374 mol/liter, ϕ_∞=0.002 V, Pe=1.31, contact angle θ=60°, 90° and 120°. In figure 3(a) has only convection flow, for contact angle θ=60°, streamline 0.0411 denotes the penetration flow which involves the electrolyte flowing from the upstream side, through the center of the cavity, then flowing to the downstream side. Due to reduce the hole width of micro-cavity, to cause the suppression of mass diffusion effect, and the enhancement of mass convection effect near the hole of micro-cavity, then forms only a single vortex at the cavity center. When contact angle θ=90°, electrolyte transport to the cathode by the diffusion flow and convection flow, the nonzero velocity gradient in the horizontal direction near the vertical cavity photo resist wall induced counter-rotating vortices appeared at the corners of the cavity. The streamline contours as transition of vortices at both corners at the contact angle θ=90°

for micro-cavity, this phenomenon agrees with the numerical results of Kondo and Fukui, so, the streamline contours as transition of vortices at both corners and of the single vortex at the center exist for cavity contact angle of 60^{o} to 90^{o}, for the contact angle $\theta=90^{o}$ has more strength convection flow and has penetration streamline 0.0462 than contact angle $\theta=60^{o}$ has. In figure 3(c), contact angle $\theta=120^{o}$ obtained the most penetration streamline 0.0472 between contact angle $\theta=60^{o}$, 90^{o} and 120^{o}, streamline contours not have any counter-rotating vortices at the corners near the up and down stream sides of the micro-cavity.

All conditions are taken as shown in figure 3, the velocity gradient-formed isostrain-rate contours are shown in figure 4, which reveals a symmetrical distribution around the mid-cavity. For contact angle $\theta=60^{o}$ figure 4(a), owing to the transition effect at the corner of the cavity, forming a phenomenon of the highest isostrain-rate density profile and most velocity unbalance between contact angle $\theta=90^{o}$ and 120^{o}. However, figure 4(c), contact angle $\theta= 120^{o}$, shows the smoother isoshear-rate contours and obtained more velocity balance profile.

The mass fraction Sh/Shmax is assessed on the cathode surface, as illustrated in figure 5. The present results of contact angle $\theta=90^{o}$ display excellent agreement with the Kondo and Fukui's data. The deviations of the root-mean-squared were below 2.13 %. Consequently, the numerical data of Kondo and Fukui were carried out implemented to ensure the accuracy of the present and neural network results. For the case of contact angle $\theta=90^{o}$, the copper is transported to the cathode via diffusion because that the copper ion is higher on the upstream side. Figure 3(b) shows that when the vortex captures the higher copper ion, the local mass transfer to the cathode is clearly increased. This result leads to the maximum flux value occurring at the upstream side of X=0. When the penetration flow is formed by higher copper ion concentration, the fresh bulk electrolyte first passes through the bottom surface on the upstream side, and then moves along the cathode surface to the downstream side. During this process, the copper ions are consumes and thus decreases. Upon reaching the downstream side, the local mass transfer to the cathode is again enhanced to a degree by diffusion from the vortex. Consequently, a smaller increase in the flux is obtained at the downstream side of X=0.9~1.0. For the other case of contact angle $\theta=120^{o}$ has the similar results with $\theta=90^{o}$, Sh/Sh_{max} decreasing from the upstream side, and will be minimized at X=0.7, upon reaching the downstream side, the local mass transfer to the cathode is again enhanced to a degree by diffusion effect. For the other case of contact angle $\theta=60^{o}$, a weaker diffusion effect appeared at the both upstream side and the downstream side induced a slight local mass transfer. However, X=0.3~0.8 occurs the hump due to most strength convection flow. The above process is primarily controlled by both convection effect of the penetration flow and the vortices effect at the corners. Two observations can be made based on this case. First, the diffusion-driven transport to the cathode is weak over convection. And second, a local resistance of mass transfer to the cathode is introduced to avoid letting the copper ion be captured by the recalculation vortices.

Figure 6 shows the dimensionless friction coefficient fraction Cf/Cfmax at the cavity hole for Cb= 0.5374 mol/liter, ϕ_{∞}=0.002 V, Pe=1.31, and for cavities with different contact angle, namely $\theta=60^{o}$, 90^{o}, 120^{o}. These curves were all sagging. The contact angle of cavity $\theta=60^{o}$ has a stronger convection effect and induced a higher friction coefficient than did $\theta=90^{o}$, 120^{o}. However, the contact angle of cavity $\theta=60^{o}$, the single vortex occupied the entire micro-cavity, causing stronger convection near the hole of micro-cavity, and thus a higher friction coefficient value was obtained for the contact angle of cavity $\theta=60^{o}$ than the value for $\theta=90^{o}$, 120^{o}.

Figure 7 shows the dimensionless current density fraction I/Imax at the hole for same conditions, such as: bulk concentration Cb, free electrolyte face voltage and Pe value as shown in figure 7, and for cavities with different contact angle. All current

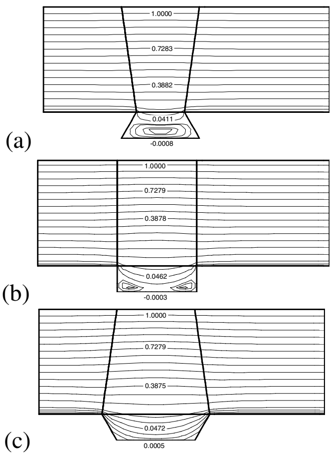


Fig. 3. Illustrates the streamlines of micro-cavity for contact angle

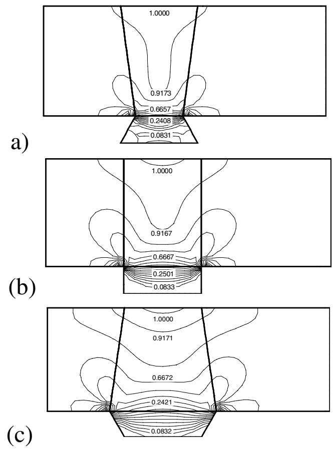


Fig. 4. Illustrates the isostrain-rate of micro-cavity for contact angle

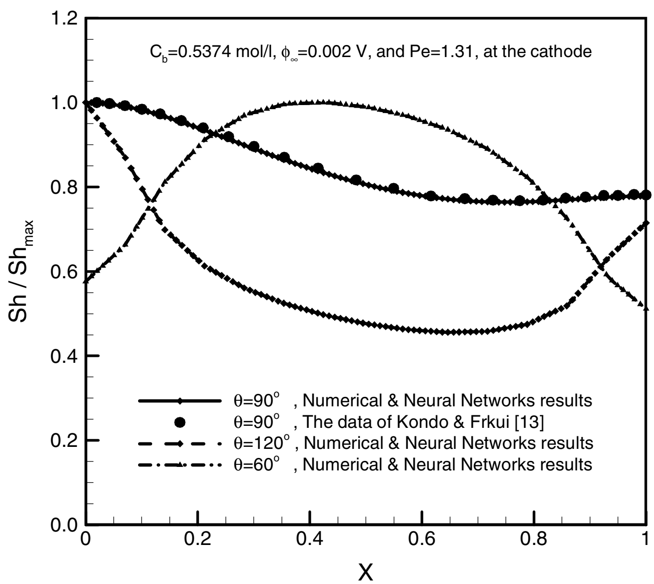


Fig. 5. Mass fraction Sh/Shmax versus X for contact angle

density curves also displayed sagging condition as the dimensionless friction coefficient fraction shown in figure 7. The contact angle of cavity θ=60° obtained higher current density value via smaller cavity hole surface, and the cathode angle θ=120° displayed lower current density by way of larger cavity hole surface. However, the contact angle of cavity θ=90°, the vortices at both corners of micro-cavity, causing stronger diffusion flow effect and induced higher current density fraction than the value for contact angle of cavity θ=120° at X=0.275~0.725. And we can forecast that more flatness current density distribution for cathode angle θ=60° will display more flatness electroform layer than did θ=90°, 120° at the hole of cavity.

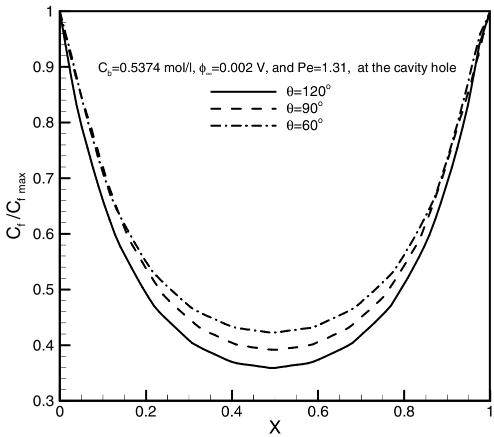


Fig. 6. Friction coefficient fraction Cf/Cfmax versus X for contact angle

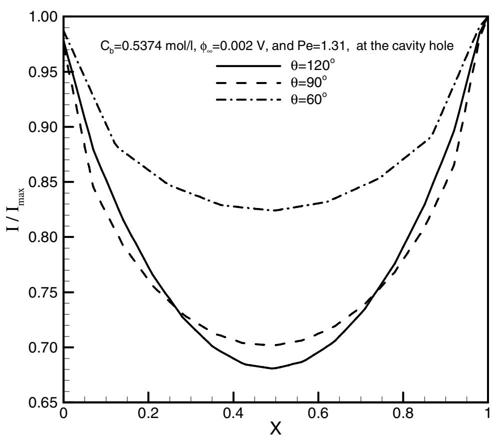


Fig. 7. Current density fraction I/Imax versus X for contact angle

5 Concluding Remarks

Without putting huge efforts of intensive computations or lengthy measurements, the present study validates the successful practice of using neural networks in real-time determination for micro-electroforming. Three back-propagation networks are adopted to predict the radial profiles of micro-electroforming. The training set, acquired from numerical database. For the training stage, the averaged RMS errors for Sh/Shmax, Cf/Cfmax, and I/Imax networks are 4.15%, 3.28%, and 2.53%. The trained neural networks can verify the prediction capability for agent technology.

Based on the geometric effect of cathode angle, the diffusion effect was induced for the cathode angle $\theta=60^o$, and appears only a single vortex at the cavity center, causing stronger convection near the micro-cavity hole. Consequently, the streamline contours as transition of vortices at both corners and of the single vortex at the center exist for cathode angle of 90^o to 120^o.

The mass fraction Sh/Sh_{max} is assessed on the cathode surface for different parameters for different geometry cathode angle. When cathode angle $\theta=90^o$ similar to cavity width a=30 μm. The results all demonstrate excellent agreement with Kondo's data. As the dimensionless friction coefficient ratio C_f/C_{fmax}, the cathode angle 120^o had a stronger convection effect and induced higher friction coefficient than did $\theta=90^o$, 60^o. As the dimensionless current density ratio I/I_{max}, the cathode angle 60^o obtained larger current density with cathode surface than did $\theta=90^o$, 120^o.

References

1. Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison-Wesley, New York (1991)
2. Demuth, H., Beale, M.: Neural Network Toolbox User's Guidance. The Math Works Inc., Natick, Massachusetts (1993)
3. Yang, H., Kang, S.W.: Improvement of Thickness Uniformity in Nickel Electroforming for the LIGA Process. J. Machine Tools and Manufacture 40, 1065–1072 (2000)
4. Zeller, R.L., Lsndau, U.: Electrochemically Active Surface Area. Voltammetric Charge Correlation for Ruthenium and Iridium Dioxide Electrodes. J. Electrochem. 138(4), 489–494 (2001)
5. Chan, K.C., Tan, H.J.: Numerical Analysis of an Inside-out Tube Inversion Process. J. of Materials Processing Technology 66, 130–136 (1997)
6. Lee, S.L., Lee, Y.F., Chang, M.H., Lin, J.C.: Pulse Plating Effects During Ni-W Electrode position. Corrosion Prevention & Control, 71-76 (1999)
7. Yin, K.M., Jan, S.L., Lee, C.C.: Current Pulse with Reverse Plating of Nickel-Iron Alloys in a Sulphate Bath. Surface and Coatings Technology 88, 219–225 (1996)
8. Andricacos, P.C., Tabib, J., Romankiw, L.T.: Stripping Voltammeter of Nickel-Iron Film Electrodeposited on Platinum Using a Rotating Ring-Disk Electrode. J. Electrochem. 135(5), 1172–1174 (1988)
9. Yamasaki, T., Schlobmacher, P., Ehrlich, K., Ogino, Y.: Formation of Amorphous Electrodeposited Ni-W Alloys and Their Nanocrystallization. Nanostructured Materials 10(3), 375–388 (1998)
10. Gould, R.D., Lopez, M.G.: Electrical Conductivity and Dynamics of Electroforming in Al-SiOx-Al Thin Film Sandwich Structure. Thin Solid Films 433, 315–320 (2003)

11. Hessami, S., Tobias, C.W.: A Mathematical Model for Anomalous Co deposition of Nickel-Iron on a Rotating Disk Electrode. J. Electrochem. 136, 3611–3616 (1989)
12. Pesco, A.M., Cheh, H.Y.: The Current Distribution within Plated Through-Holes: II. The Effect of Periodic Electrolysis. J. Electrochem. 136(2), 408–414 (1989)
13. Kondo, K., Fukui, K., Uno, K., Shinohara, K.: Shape Evolution of Electrodeposited Copper Bumps. J. Electrochem. 143(6), 1880–1886 (1996)
14. Georgiadou, M.: Modeling Current Density Distribution in Electrochemical Systems. Electrochemical Acta, 48 (2003) 4089-4095.
15. Duchanoy, C., Lapicque, F.: Current Distributions on Patterned Electrodes in Laminar Flow. Chemical Engineering Science 55, 1115–1126 (2000)
16. Beck, U., Smith, D.T., Reiners, G., Dapkunas, S.J.: Mechanical Properties of SiO2 and Si3N4 Coatings: A BAM/NIST Co-Operative Project. Thin Solid Films 332, 164–171 (1998)

On Drawing Conclusions in Presence of Inconsistent Data

Sylvia Encheva[1] and Sharil Tumin[2]

[1] Stord/Haugesund University College, Bjørnsonsg. 45, 5528 Haugesund, Norway
sbe@hsh.no
[2] University of Bergen, IT-Dept., P. O. Box 7800, 5020 Bergen, Norway
edpst@it.uib.no

Abstract. Most automated tests assessing students understanding of a concept contain one question requiring application of that concept. To attain a higher level of certainty in the evaluation process we propose a test with three different questions requiring application of one concept. Such a test is intended to facilitate the self-assessment process and can be suggested to students after a concept has been introduced. Lattice theory and higher-order logic are further applied for presenting a structure that can serve as a building block of an intelligent tutoring system.

Keywords: five-valued logic, automated tests.

1 Introduction

A lot of research has been done on evaluating students' knowledge during the last few decades. Considerable part of that research aims at developing automated tests providing immediate feedback. A common drawback of such tests is that they employ Boolean logic in the process of decision making. Thus if the response does not appear to be necessarily true, the system selects false. While Boolean logic appears to be sufficient for most everyday reasoning, it is certainly unable to provide meaningful conclusions in presence of inconsistent and/or incomplete input [5], [8]. This problem can be resolved by applying many-valued logic.

Another interesting issue regarding assessment of knowledge is related to the number of questions one should ask while establishing students' level of mastering of a particular skill. We propose short tests that can be incorporated in an intelligent tutoring system, where each test contains four questions requiring a particular skill application. Such a test can be suggested to students after a skill has been obtained.

Introducing more than one question helps to attain a higher level of certainty in the process of decision making. At the same time the number of possible answer combinations increases rapidly. In practice it appears to be quite difficult and timeconsuming to find a proper response to each answer combination. In order to overcome this obstacle we propose grouping of all possible answer combinations into meaningful sets first. Application of many-valued logic, for drawing conclusions and providing recommendations, is then suggested.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 32–41, 2008.

The rest of the paper is organized as follows. Related work, basic terms and concepts are presented in Section 2. The management model is described in Section 3. The paper ends with a description of the system in Section 4 and a conclusion in Section 5.

2 Background

Let P be a non-empty ordered set. If $sup\{x, y\}$ and $inf\{x, y\}$ exist for all $x, y \in P$, then P is called a *lattice*, [2]. In a lattice illustrating partial ordering of knowledge values, the logical conjunction is identified with the meet operation and the logical disjunction with the join operation.

A three-valued logic, known as Kleene's logic is developed in [9] and has three truth values, truth, unknown and false, where unknown indicates a state of partial vagueness. The semantic characterization of a four-valued logic for

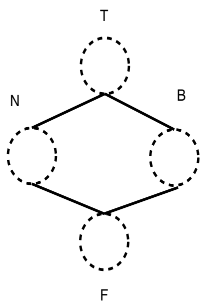


Fig. 1. Logic lattice for the truth values in Belnap's logic

expressing practical deductive processes is presented in [1]. The Belnap's logic has four truth values true (T), false (F), both (B), none (N). The meaning of these values can be described as follows: an atomic sentence is stated to be true only (T), an atomic sentence is stated to be false only (F), an atomic sentence is stated to be both true and false, for instance, by different sources, or in different points of time (B), and an atomic sentences status is unknown. That is, neither true, nor false (N). A logic lattice for the four truth values is shown on Fig. 1. A truth table for the ontological operation $\wedge$ is presented in Table. 1.

Nested line diagrams are used for visualizing large concept lattices, emphasizing sub-structures and regularities, and combining conceptual scales, [14]. A

Table 1. Truth table for the ontological operation $\wedge$ in Belnap's logic

$\vee$	T	B	N	F
T	T	B	N	F
B	B	B	F	F
N	N	F	N	F
F	F	F	F	F

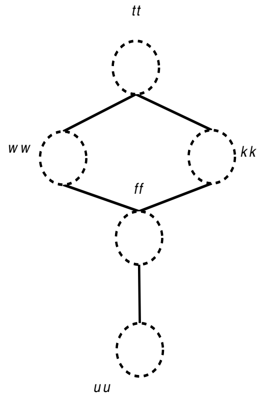


Fig. 2. Lattice of the five-valued logic

nested line diagram consists of an outer line diagram, which contains in each node inner diagrams.

The five-valued logic introduced in [3] is based on the following truth values: *uu* - unknown or undefined, *kk* - possibly known but consistent, *ff* - false, *tt* - true, *ww* - inconsistent.

A truth table for the ontological operation ∨ is presented in Table 2.

Table 2. Truth table for the ontological operation ∨

∨	uu	kk	ff	tt	ww
uu	uu	uu	ff	uu	uu
kk	uu	kk	ff	kk	ww
ff	ff	ff	ff	ff	ff
tt	uu	kk	ff	tt	ww
ww	uu	ww	ff	ii	ww

A level-based instruction model is proposed in [11]. A model for student knowledge diagnosis through adaptive testing is presented in [6]. An approach for integrating intelligent agents, user models, and automatic content categorization in a virtual environment is presented in [12].

The Questionmark system [10] applies multiple response questions where a set of options is presented following a question stem. The final outcome is in a binary form, i.e. correct or incorrect because the system is based on Boolean logic [4], [13].

3 Establishing Level of Understanding

For establishing the current level of understanding of a concept we propose a multiple choice test with three questions. A student can take the test several times after the concept has been introduced and once at the end of the semester.

The ‘summarized’ result of the student’s responses after the concept has been introduced is obtained applying operation for truth values in a five value logic, [3]. Since the operation is both associative and commutative we can present both to the student and to the lecturer the result of each test and the ‘summarized’ result of all tests taken by the student. All possible responses are grouped in four sets as illustrated on Fig. 7. Each of the four sets contains five of the twenty possible answer combinations. Any five answer combinations within one set are arranged in a lattice like the one shown on Fig. 2. Thus all the twenty possible answer combinations are reduced to four cases, that correspond to the truth values in Belnap’s logic The final outcome is a result of the operation between the ‘summarized’ result of all tests taken after the concept has been introduced and the test at the end of the semester. The corresponding truth values assigned to those two results are from the Belnap’s logic, [1]. Every answer in a test can be correct c, incorrect i or partially correct p. The case where no answer is provided is denoted by n. Thus we obtain the following twenty answer combinations: *ccc* - three correct answers, *ccp* - two correct answers and one partially correct answer, *ccn* - two correct answers and one unanswered question, *cci* - two correct answers and one incorrect answer, *cpp* - one correct answer and two partially correct answers, *cnn* - one correct answer and two unanswered questions, *cii* - one correct answer and two incorrect answers, *cpn* - one correct answer, one partially correct answer, one unanswered question, *cpi* - one correct answer, one partially correct answer, and one incorrect answer, *cni* - one correct answer, one unanswered question, and one incorrect answer, *ppp* - three partially answers, *ppn* - two partially answers and one unanswered question, *ppi* - two partially answers and one incorrect answer, *pnn* - one partially answer and two unanswered questions, *pii* - one partially answer and two incorrect answers, *pni* - one partially answer, one unanswered question and one incorrect answer, *nnn* - three unanswered questions, *nni* - two unanswered questions and one incorrect answer, *nii* - one unanswered question and two incorrect answers, *iii* - three incorrect answers.

Inner lattices - all answer combinations are grouped in the four lattices with respect to the five truth values, [3].

Any two nodes, in Fig. 3, Fig. 4, Fig. 5, and Fig. 6 connected by an edge differ in one answer only. Going upwards from one level to another in the lattices on Fig. 3, Fig. 4, Fig. 5, and Fig. 6 increases the level of knowledge.

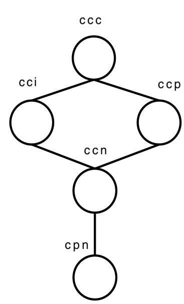


Fig. 3. Answer combinations related to the truth value *tt*

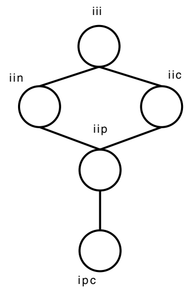


Fig. 4. Answer combinations related to the truth value *ww*

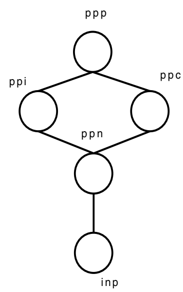


Fig. 5. Answer combinations related to the truth value *pp*

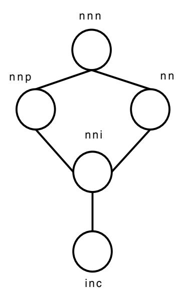


Fig. 6. Answer combinations related to the truth value *nn*

Outer lattice the lattice with the four truth values of Belnap's logic from Fig. 1.

Nested lattice - the previously described twenty answer combinations are arranged in a nested lattice, Fig. 7. This lattice visualizes the learning progress of each student.

The 'summarized' result of tests taken after a concept is introduced can move inside one of the four outer circles and which implies changes of the understanding status within one of the four truth values. If it moves from one of the outer circles to another than it implies changes of the understanding status between two of the four truth values.

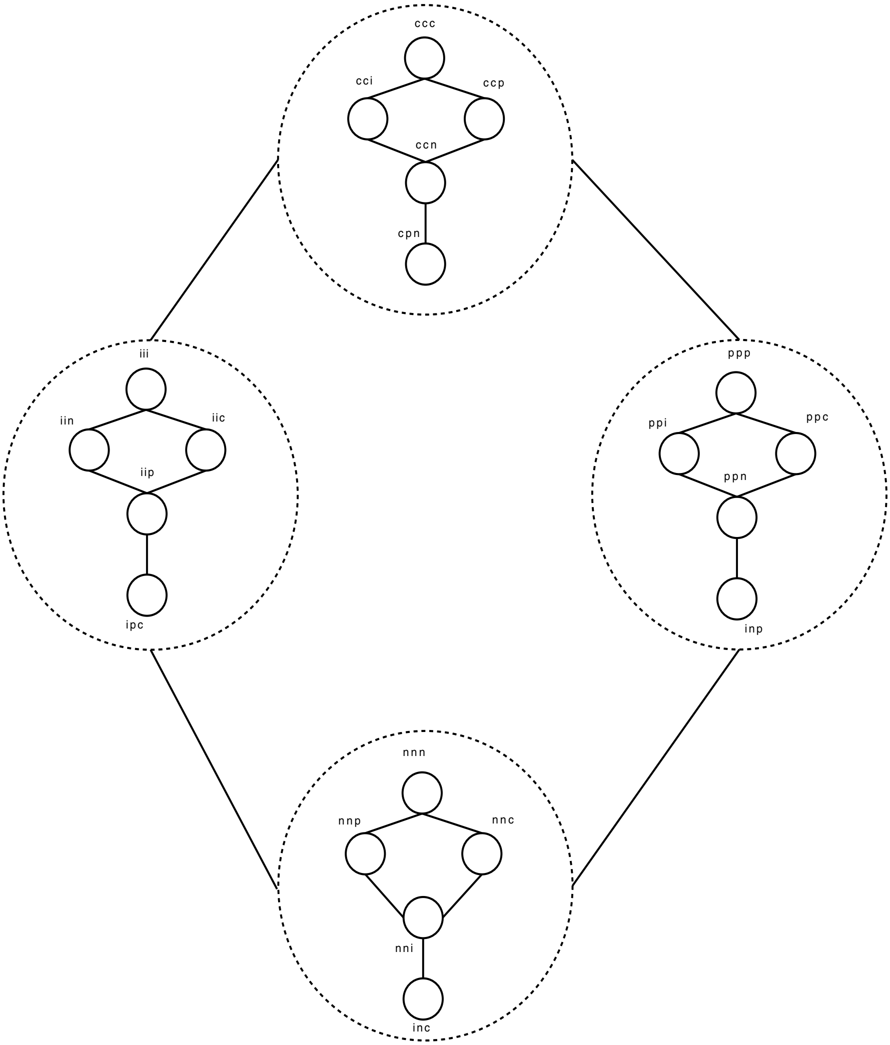


Fig. 7. Nested lattice

On Fig. 8 we show in details how the twenty answer combinations from Fig. 7 are related. Any two nodes except *ccc, iii, ppp*, and *nnn* in the lattice on Fig. 7 connected by an edge differ in exactly one answer.

The four nodes *ccc, iii, ppp*, and *nnn* are connected since they correspond to the four truth values in Belnap's logic.

4 System Implementation

The three tests' questions of a particular assessment are randomly taken from a pool of questions related to a particular skill. Within the options following

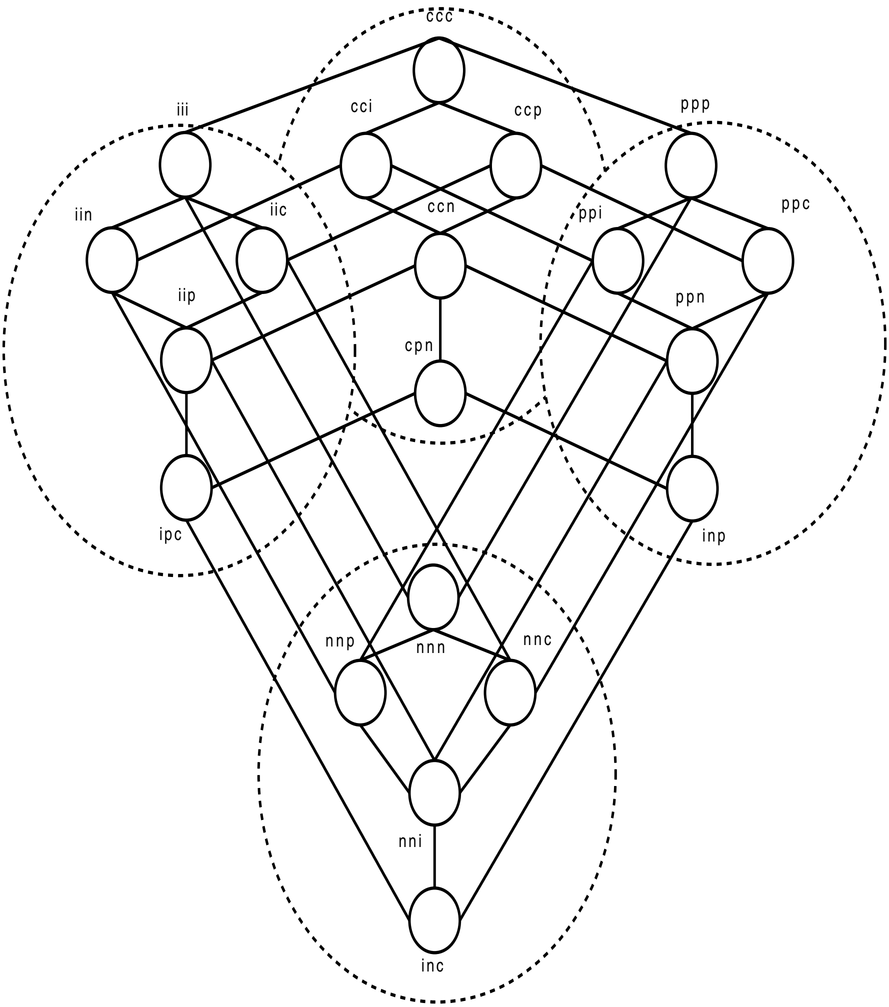


Fig. 8. Relations among corresponding truth values

each question, there is only one correct answer. Each question can have more than one partially correct and/or incorrect answers related to it. Thus for each test with three options the system chooses randomly one partially correct out of the set with partially correct answers and one incorrect answer out of the set with incorrect answers. A student taking a test can only choose one of the presented options. It is up to the test designer to state which one is correct, partially correct or incorrect. An assessment is dynamically build at run time each time a student is accessing the assessment part of the Web site.

The system shown on Fig. 9 can be implemented as a typical three-tiers Web application server architecture. The presentation layer is handled by an Apache Web server. The logic layer is written in Python. The data layer is implemented

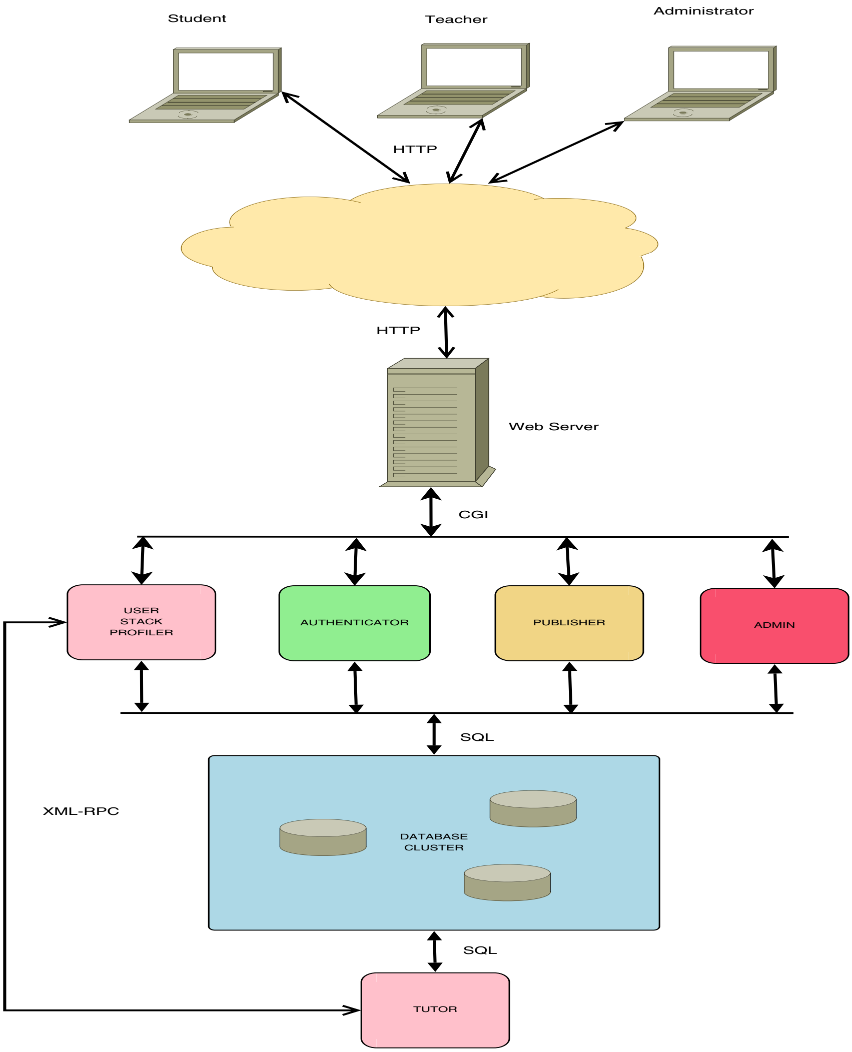


Fig. 9. System architecture

using SQLite database engine. Python provides a programming environment for implementing script-based handler for dynamic content, data integration and users' software agents. The back end SQLite databases are used to store both static and dynamic data. Apache is a modular Web server that can incorporate a high level scripting language as a module such as f. ex. mod_python. Using mod_python, python interpreter becomes a part of the Web server. SQLite is a small footprint, zero-administration and serverless database system. SQLite stores persistence data into files. SQLite thus provides a database platform for multiple databases. The separation of these units makes it possible to modularly design and implement the system as loosely couple independent sub-systems. Communication framework based on XML-RPC is used to connect the Web application middleware and the intelligent assessment/diagnostic system together. The system is composed of sub-systems, all of which can be written in Python. The dynamic document publisher sub-system compiles a page to be presented to the user. How the page will look is defined in a template object in relation to the user response, current state variables and activities history. Documents and template objects are read from a common documents database. Tests questions data is read from a common tests database.

The authenticator sub-system authenticates a user during login and creates initial session contact in the system if the user provides correct credentials. This sub-system also provides user authorization during an active user's session and is also responsible for session cleanup at user's log off. The stack profiler sub-system keeps track of user activities history in a stack like data structure in the user database. Each event, like for example response/result of a test or a change of learning flows after following a hint given by the system, is stored in this database. This sub-system provides the percepts to the intelligent tutorial sub-system. The stack profiler receives student's test responses and calculates the final result of the test. The stack profiler provides a user with immediate feed-back.

Analysis of the user's profile is given to the tutor sub- system. This subsystem runs in different processes but operates on the same user database. The tutor sub-system provides a student with intelligent diagnostics and the optimal study flow for that particular student based on her current profile. The hints and diagnostics are written into the user database. The stack profiler which is accessible by the user's browser reads the most recent messages from the tutor sub-system whenever hints or diagnostics are requested by the user. Using a rule-based method the tutor sub-system provides the student with the best study advice to facilitate students' progress in mastering new skills. Since the system is built on different sub-systems the independent sub-system can Be maintained and improved separately. These subsystems are operationally bounded with each other by the information stored in the multiple databases.

5 Conclusion

This paper is devoted to assessing students understanding of a concept applying automated testing procedure. Responses to three questions addressing one particular concept are used in the decision making process. This contributes to attaining higher level of certainty in the provided feedback. Employing many-valued logic for drawing a conclusion after a test is taken consecutively allows real use of student's previous tests results. Furthermore applying many-valued logic captures the complexity of students' responses much better than Boolean logic.

References

1. Belnap, N.J.: A useful four.valued logic. In: Dunn, J.M., Epstain, G. (eds.) Modern uses of multiple-valued logic, pp. 8–37. D. Reidel Publishing Co., Dordrecht (1977)
2. Davey, B.A., Priestley, H.A.: Introduction to lattices and order. Cambridge University Press, Cambridge (2005)
3. Ferreira, U.: A Five-valued Logic and a System. Journal of Computer Science and Technology 4(3), 134–140 (2004)
4. Goodstein, R.L.: Boolean Algebra. Dover Publications (2007)
5. Gradel, E., Otto, M., Rosen, E.: Undecidability results on two-variable logics. Archive of Mathematical Logic 38, 313–354 (1999)
6. Guzmàn, E., Conejo, R.: A model for student knowledge diagnosis through adaptive testing. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 12–21. Springer, Heidelberg (2004)
7. Huffman, D., Goldberg, F., Michlin, M.: Using computers to create constructivist environments: impact on pedagogy and achievement. Journal of Computers in mathematics and science teaching 22(2), 151–168 (2003)
8. Immerman, N., Rabinovich, A., Reps, T., Sagiv, M., Yorsh, G.: The boundery between decidability and undecidability of transitive closure logics. In: Marcinkowski, J., Tarlecki, A. (eds.) CSL 2004. LNCS, vol. 3210, Springer, Heidelberg (2004)
9. Kleene, S.: Introduction to Metamathematics. D. Van Nostrand Co., Inc., New York (1952)
10. http://www.leeds.ac.uk/perception/v4_mrq.html
11. Park, C., Kim, M.: Development of a Level-Based Instruction Model in Web-Based Education. LNCS (LNAI), vol. 3190, pp. 215–221. Springer, Heidelberg (2003)
12. Santos, C.T., Osòrio, F.S.: Integrating intelligent agents, user models, and automatic content categorization in virtual environment. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 128–139. Springer, Heidelberg (2004)
13. Whitesitt, J.E.: Boolean Algebra and Its Applications. Dover Publications (1995)
14. Wille, R.: Concept lattices and conceptual knowledge systems. Computers Math. Applic. 23(6-9), 493–515 (1992)

Basic Concepts of Knowledge-Based Image Understanding

R. Tadeusiewicz[1] and P.S. Szczepaniak[2]

[1] AGH University of Technology,
al. Mickiewicza 30, 30-059 Krakow, Poland
[2] Institute of Computer Science, Technical University of Lodz,
ul. Wolczanska 215, 90-924 Lodz, Poland

Abstract. In the paper, the main paradigm of image understanding as well as possible way for practical machine realization in relatively simple situations is presented. The notion 'simple situations' reflects more our humility with respect to the complication of human perception process than the form of objects to be recognized and interpreted. Crucial for our approach are formalization of human knowledge about class of images to be automatically interpreted and realization of cognitive resonance while the particular method put at work is the active contour approach.

Keywords: Image understanding, machine learning, active contours.

1 Introduction to Image Understanding

So far, *image understanding* is assumed to be human specific ability to explain and interpret content of image on the basis of image perception, object recognition, knowledge, experience, reasoning, culture, association, context analysis, etc. Consequently, it seems to be a complicated and not unique process. Let us consider two images given on Fig. 1a and 1b.

The first problem we have in automatic image analysis is recognition of objects; here the main two objects are: man and vehicle. On the Fig. 1b an additional problem in object recognition exists: do we have a boat or a bicycle? On the basis of surroundings, we conclude that this must be a kind of bicycle, called ricksha. Additional observation can be that in both cases "man is sitting on the vehicle". Even assuming that computer is able to follow this recognition process using its specific methods of image analysis, the answer to the questions "what common story is contained on both images" seems to be impossible for computers if human understanding mechanism is not implemented. Additional problem is that the explaining story "waiting for pick-up guest" is most probable but not the only possible.

Another, simpler example - Figure 2. Using current methods of image analysis computer is able to perform segmentation and to extract objects which are Chinese and English letters. But this is only the first important step in understanding of the image content. The next step is to explain the meaning of both texts. And finally, we maybe can reason that one text explains the meaning of

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 42–52, 2008.

a)
b)

Fig. 1. Ricksha or ”waiting for pick-up guest”

Fig. 2. Information for pedestrians

the second textual row and vice versa. This final remark may be understood as the story of image 2. However, if the meaning of both rows is different on account of something (assuming we know only one language) then our story is false.

2 Automatic Image Understanding basic Approach

The approach of *automatic understanding of images* described in the book [1] is dedicated to medical applications [2,3,4,5]. However, its generality allows to extract some fundamental issues.

Trying to explain what automatic understanding is and how we can force the computer to understand the image content we must demonstrate the fundamental difference between a formal description of an image and the content meaning of the image, which can be discovered by an intelligent entity capable of understanding the profound sense of the image in question. The fundamental features of automatic image understanding are:

1. Imitation of the human way in image analysis and in reasoning about the content. The qualified professional and well defined field of interest are preferred here. Example medical images of certain organ and restricted class of variations in image content.
2. Linguistic description of the image content using a special kind of image description language.
3. The image content linguistic description constructed in this manner constitutes the basis for understanding of image merit content.

The most important difference between all traditional methods of automatic image processing and the new paradigm for image understanding is that there is a feed forward data flow in the traditional methods while in the new paradigm [1] there are two-directional interactions between signals (features) extracted from the image analysis and expectations resulting from the knowledge of image content as given by experts. In Fig. 3 the traditional chart representing image processing for recognition purposes is presented.

Unlike in this simple scheme representing classical recognition, in the course of image understanding we always have a two-directional flow of information (Fig. 4).

In both figures we can see that when we use the traditional pattern recognition paradigm (Fig. 3), all processes of image analysis are based on a feed-forward scheme (one-directional flow of signals). On the contrary, when we apply automatic understanding of the image (Fig. 4), the total input data stream (all features obtained as a result of an analysis of the image under consideration) must

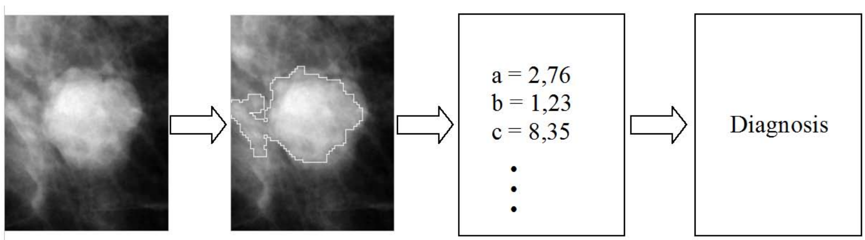


Fig. 3. Classical automatic pattern recognition process

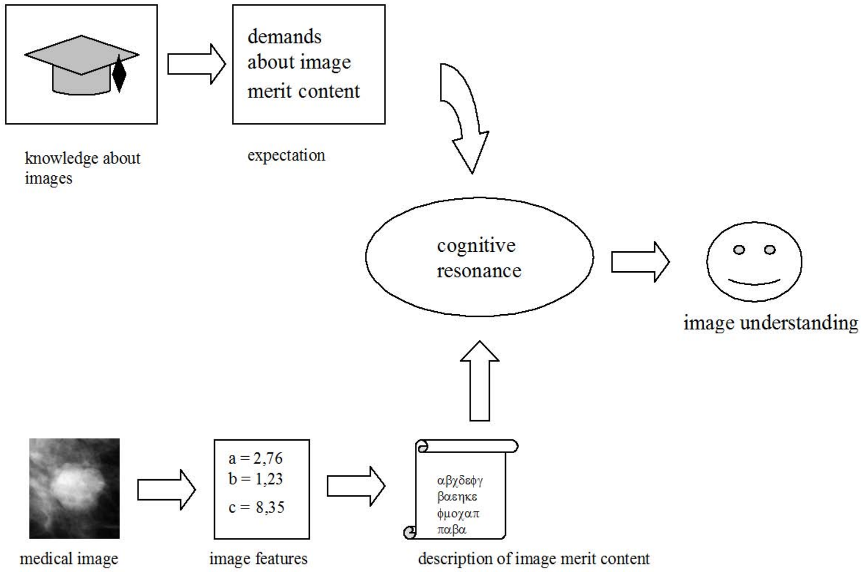


Fig. 4. The main paradigm of image understanding

be compared with the stream of *demands* generated by a dedicated *source of knowledge*. The demands are always connected with a specific (selected) hypothesis of the image content semantic interpretation. As a result, we can emphasize that the proposed 'demands' are a kind of postulates, describing (basing on the knowledge about the image contents) the desired values of some (selected) features of the image. The selected parameters of the image under consideration must have desired values when some assumption about semantic interpretation of the image content is to be validated as true. The fact that the parameters of the input image are different can be interpreted as a partial falsification of one of possible hypotheses about the meaning of the image content, however, it still cannot be considered the final solution.

Due to this specific model of inference we call our mechanism the '*cognitive resonance*'. This name is appropriate for our ideas because during the comparison process of the features calculated for the input image and the demands generated by the source of knowledge we can observe an amplification of some hypotheses (about the meaning of the image content) while other (competitive) hypotheses weaken. It is very similar to the interferential image formed during a mutual activity of two wave sources: at some points in space waves can add to one another, in other points there are in opposite phases and the final result is near zero.

Such a structure of the system for image understanding corresponds to one of the very well known models of the natural human visual perception, referred to as 'knowledge based perception'. The human eye cannot recognize an object if the brain has no template for it. This holds true even when the brain knows the object, but shown in another view, which means that other signals are coming to the visual cortex. Indeed, natural perception is not just the processing of visual signals received by eyes. It is mainly a mental cognitive process, based on hypotheses generation and its real-time verification. The verification is performed by comparing permanently the selected features of an image with expectations taken from earlier visual experience. The presented approach to image understanding is based on the same processes with a difference that it is performed by computers.

3 Active Contour as a High-Level Recognition Mechanism

From the previous section it is clear that there is a difference between classic *recognition* (last step in Fig. 3 resulting in diagnosis) and image *understanding.* Understanding means validation of correctness of some *hypothesis* from the set of available *concepts*, cf. Fig. 6. In simple situation, only some *generic concepts* are considered. However in the reality, also relations between those generic concepts needs to be identified and evaluated. In other words, general concepts are determined by generic concepts and their mutual relations [12]. Consequently, we define *hypothesis* as the (generic) concept under actual verification or in the more general case as the set of actually verified concepts together with their mutual relations. Note that frequently such general *concepts* can only be created if relations between simpler *concepts* composing the general one have correctly been identified. The task of hypothesis verification can only be performed if some external knowledge is available and can be implemented for machine use.

Originally, *active contour* methods were developed as tools for a low-level image segmentation (Fig. 5) but with ability for use of high-level information [7,8]. The main idea is to find an optimal contour in the space of considered contours representing certain region on the image. The search is performed in an evolution process (optimization) in which the given objective function, called energy, evaluates the quality of contour. As shown in [10], contours are contextual classifiers of pixels (one part of pixels belongs to the interior and another one to the exterior of given contour, cf. Fig. 5) and active contours are methods of optimal construction of classifiers.

Moreover, active contours can be used as methods of identification (selection) of groups of pixels composing certain objects which may be interpreted as *machine concepts* (or *concepts* for short). The possible way for hypothesis verification can be summarized as follows.

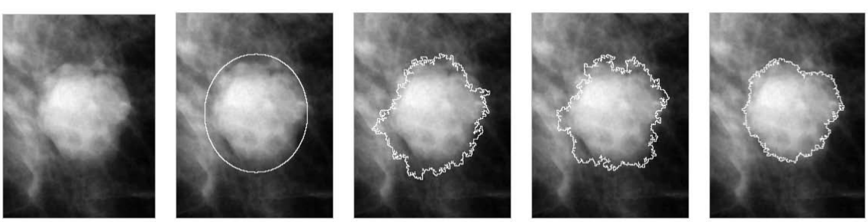


Fig. 5. Segmentation using active contour method

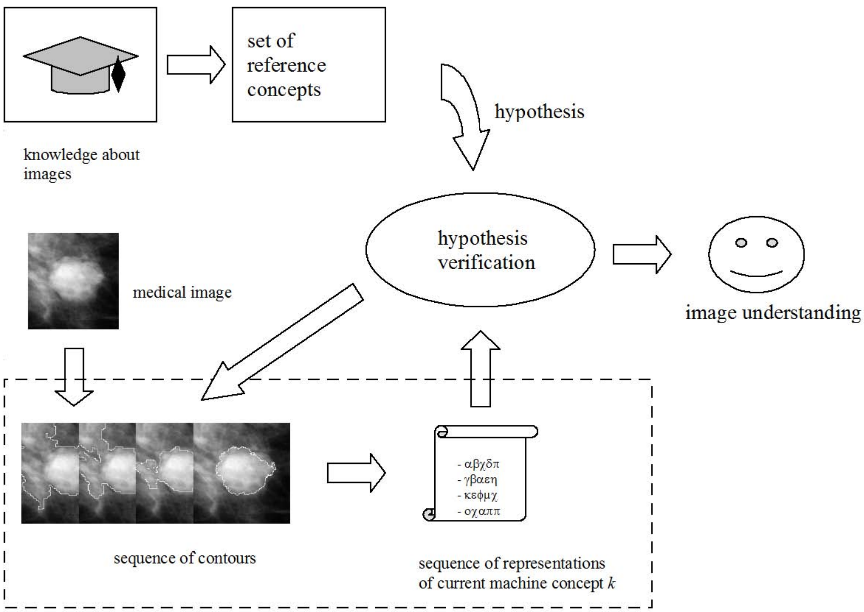


Fig. 6. Active contours as example of hypothesis verification

For each considered k-th hypothesis, the active contour method with relevant form of energy E^k can be applied.

– In the i-th step of the active contour method, C_i^k is machine concept (set of pixels) identified by current contour.
– Energy E^k evaluates quality of given set C_i^k as a potential concept of the type k. In the energy, some external high-level knowledge can be represented. It can model expectations with respect to machine concepts, i.e. sets of pixels or relations between those sets.
– Evolution (optimization of the k-th energy function) is the way of search for the optimal contour.

Hypothesis verification means search for the best C^k for all considered k.

Let us explain the lower part of Fig. 6 more exactly. Using certain form of energy function we decide about the actual form of the contour which we are looking for (hypothesis k). Due to its activity, contour method generates sequence of contours C_i^k. Each i-th proposed contour is evaluated by the energy function of the form k. If the quality of current contour is not good enough (value of the energy is too large) then the contour is modified in such a way that it should better fulfill the expectation closed in the form of energy in other words next, possible better contour is created. If the final result is not satisfactory (hypothesis θ^k has not been confirmed) then we can start the process with another form of energy, i.e. for another possible hypothesis θ^l , $l \neq k$.

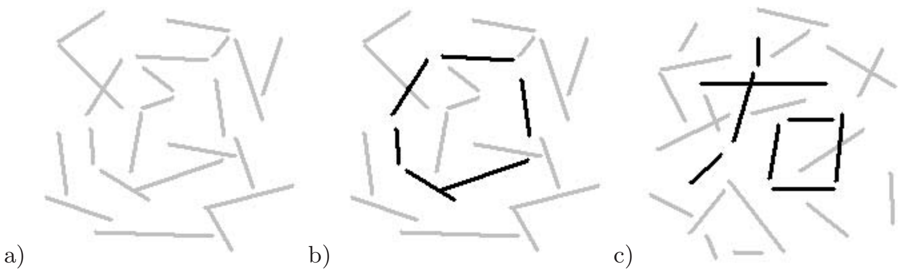


Fig. 7. Concept segmentation: (a) - segments of lines, (b) - approximate composition of circle, (c) - Chinese letter hidden in the set of segments, cf. Fig. 2

Let us consider an image where some segments of lines have already been detected by the use of some low-level method Fig. 7a). The goal is to find a circle composed by those segments. The form of energy should assure that its minima give compositions in a form which is similar to the *concept* of circle. The possible implementation is: consideration of coordinates of the endpoints of segments of lines and definition of the concept of circle as minimum of the mean difference between the expected radius and the distance from the endpoints of segments to the expected center of circle. Additionally, difference between expected perimeter and sum of lengths of segments can be minimized. This approach reflects some heuristic knowledge about *concept of circle* and it shows that active contour approach allows for use of external *expert knowledge* or user expectation. The computational result of search performed by evolution driven by simulated annealing is shown on Fig. 7b).

Similarly, we can try to implement search for any other hypothesis, e.g. for Chinese letter hidden in the picture Fig. 7c), assuming that we have an idea in construction of relevant energy function. Another question not discussed here is, if the active contour approach is effective in recognition of letters in comparison to other techniques.

With regard to the above consideration we conclude that active contours can be numbered as *image understanding* technique. The condition is that we are able to construct energy function in the way that desired concepts can be implemented. So far, the method was usually used to find concepts (confirm hypotheses) when operating on pixels which are the simplest image elements. As shown (Fig. 7), it can be generalized and used for more complex structures and on different levels of abstraction. Here, we can also speak about *concept segmentation.*

4 Adaptive Potential Active Hypercontours

Realization of automatic image understanding needs fulfilling at least two requirements: formalized *description (representation) of human expectation with*

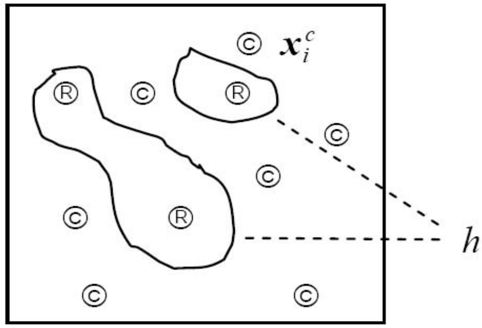


Fig. 8. Hypercontour h consisting of two contours determined by control points {©,®}; binary classification case

respect to image content and implementation of some *active mechanism* which can be understood as a kind of *cognitive resonance* cf. Fig. 4. In the current stage of technology, the mechanism may be rather simple but it must be some active process of recognition performed by machine and based on *human knowledge* reflected in demands about image merit content. Of course, we can simply imagine that the expectation in the form of mentioned demands about image content may arise from another machine. The notion image can also contain a more general meaning. However, both such assumptions does not change the main idea of image understanding.

Let us consider *potential active hypercontours* (PAH) an implementation of *active hypercontours* introduced in [9]; see also [10,11].

Let X - called feature space - be a metric space with metric $\rho : X \times X$.

The potential hypercontour h is determined by set D^c of labeled *control points* $\mathbf{x}_i^c$ and assigned to them potential functions P_i^c, $i = 1, 2, \ldots, N^c$ (Fig. 8).

Each point $\mathbf{x}_i^c$ is source of potential (similar to the electrical one) the value of which decreases the larger the distance from that point. There are many ways for definition of P_i^c; example are:

– exponential potential function

$$P_i^c(\mathbf{x}) = \Psi_i \exp\left(-\mu_i \rho^2\left(\mathbf{x}_i^c, \mathbf{x}\right)\right) \tag{1}$$

– inverse potential function

$$P_i^c(\mathbf{x}) = \frac{\Psi_i}{1 + -\mu_i \rho^2\left(\mathbf{x}_i^c, \mathbf{x}\right)} \tag{2}$$

where real numbers Ψ_i and μ_i are parameters characterizing the potential field of each control point.

Note that number of control points, their distribution and field parameters fully determine the shape of hypercontour.

Since each control point is labeled, i.e. one dispose of pairs $(\mathbf{x}_i^c, l_i^c)$, then h is in fact a classifier and it can be formally defined as follows:

$$\kappa(\mathbf{x}) = \arg\max_{l \in L} \sum_{i=1}^{N^c} P_i^c\left(\mathbf{x}_i^c, \mathbf{x}\right) \delta\left(l_i^c, l\right) \text{ for each } \mathbf{x} \in X \tag{3}$$

where L is set of labels, l_i^c, l denote labels of $\mathbf{x}_i^c$ and the examined one while $\delta\left(l_i^c, l\right) = \begin{cases} 0 & l \neq l_i^c \\ 1 & l = l_i^c \end{cases}$

Potential active contour (PAH) possesses ability to *evolution* with change of the location or number of control points, and with modification of parameters of potential functions. The search of optimal hypercontour is performed by optimization of some performance index E called energy in the theory of active contours; $E : H \to R^+ \cup \{0\}$ with H being space of all available hypercontours. It has been proved in [9] that each hypercontour generates the corresponding classification function. This statement is true if the space X is metric. In E almost any type of information can be used assuming that we are able to implement this information in the computer oriented form. We can also decide if the classification is *supervised* or *unsupervised*. In the following we deal with the supervised one because of its more intuitive realization.

As mentioned, there is certain freedom in definition of energy function. In terms of image understanding paradigm, we aim at incorporation of expert knowledge into E and there are many ways to do it. Let for example

$$E(h) = \sum_{j=1}^{J} \chi_j \left[1 - \delta\left(l_j^{cor}, \kappa(\mathbf{x}_j^{cor})\right)\right] \tag{4}$$

for each $h \in H$ where κ is a corresponding to hypercontour h classifier. Here $\left\{\left(\mathbf{x}_j^{cor}, l_j^{cor}\right)\right\}_{j=1,\ldots,J}$ denotes set of correctly labeled objects (*reference points*), χ_j is a real number introduced by expert and interpreted as significance of the j-th reference point for correct classification, and δ is the Kronecker symbol.

A similar concept of using expert knowledge was proposed in the unsupervised classification problem [13] where expert estimated proximity between pairs of objects from the training set.

The search of the optimal hypercontour may be driven in many ways e.g. by use of simulated annealing or genetic algorithm which perform global search and do not use gradient. In our work we apply the first mentioned method.

Adaptation is another interesting and powerful mechanism. Discrimination ability of given hypercontour is limited and it depends on the number of control points (assuming that other parameters are fixed). Flexibility of the potential active hypercontous (PAH) can be improved if we incorporate the change of the number of control points into the optimization procedure. For example, we can start with small number of that points and add new ones if necessary. The rate of misclassification in some area of the space X can be the reason for introducing a few new control points.

Let us summarize. On the basis of the above theory of potential contours we conclude that crucial role in *representation of human expectation with respect to image content* play: form of the energy function, labelled control points and definition of the contour. In this paper, the ways for contour representation have not been discussed. *Adaptation* and *evolution* are two processes that optimise the contour within the verified concept. Change of the hypothesis in a simple way imitates the cognitive resonance.

5 Concluding Remarks

Realization of *automatic understanding of images* needs preprocessing, application of both low-level and high-level methods of image analysis the contextual one including, as well as the cognitive resonance where machine concept propositions are verified by confrontation with formalized human knowledge (hypothesis). A lot of methods of low-level image analysis is known; some of them can operate on different levels of granulation, e.g. two- or multi-layer Kohonen neural network [14,15]. There may also exist diverse ways for technical imitation of human perception. Here, we show that certain potential stands behind the active contour approach. Anyway, embedding of mechanism known from investigation of human visual perception seems to be necessary if we aspire at consistence of human and machine perception. Of course, in many computer tasks of practical importance such conformability is not needed for effective use of machines.

Acknowledgment

This work has been supported by the Ministry of Science and Higher Education, Republic of Poland, under project no. N 519 007 32/0978, decision no. 0978/T02/2007/32.

References

1. Tadeusiewicz, R., Ogiela, M.R.: Medical Image Understanding Technology. Studies in Fuzziness and Soft Computing, vol. 156. Springer, Berlin (2004)
2. Ogiela, L., Tadeusiewicz, R., Ogiela, M.R.: Cognitive Approach to Visual Data Interpretation in Medical Information and Recognition Systems. In: Zheng, N., Jiang, X., Lan, X. (eds.) IWICPAS 2006. LNCS, vol. 4153, pp. 244–250. Springer, Heidelberg (2006)
3. Ogiela, L., Tadeusiewicz, R., Ogiela, M.R.: Cognitive Computing in Intelligent Medical Pattern Recognition Systems. In: Huang, D.-S., Li, K., Irwin, G.W. (eds.) Intelligent Control and Automation. Lecture Notes in Control and Information Sciences, vol. 344, pp. 851–856. Springer, Berlin, Heidelberg, New York (2006)
4. Ogiela, M.R., Tadeusiewicz, R., Ogiela, L.: Graph Image Language Techniques Supporting Radiological, Hand Image Interpretations. In: Computer Vision and Image Understanding, vol. 103/2, pp. 112–120. Elsevier, Amsterdam (2006)
5. Ogiela, M.R., Tadeusiewicz, R., Ogiela, L.: Image Languages in Intelligent Radiological Palm Diagnostics. Pattern Recognition 39/11, 2157–2165 (2006)

6. Tadeusiewicz, R., Ogiela, M.R.: Why Automatic Understanding? In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4432, pp. 477–491. Springer, Heidelberg (2007)
7. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. Int. Journal of Computer Vision 22(1), 61–79 (2000)
8. Kass, M., Witkin, W., Terzopoulos, S.: Snakes: Active Contour Models. Int. Journal of Computer Vision 1(4), 321–333 (1988)
9. Tomczyk, A.: Active Hypercontours and Contextual Classification. In: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications – ISDA 2005, Wroclaw, Poland, pp. 256–261. IEEE Computer Society Press, Los Alamitos (2005)
10. Tomczyk, A., Szczepaniak, P.S.: On the Relationship between Active Contours and Contextual Classification. In: Kurzyński, M., et al. (eds.) Computer Recognition Systems. Proceedings of the 4th Int. Conference on Computer Recognition Systems – CORES 2005, pp. 303–310. Springer, Heidelberg (2005)
11. Tomczyk, A., Szczepaniak, P.S.: Adaptive Potential Active Hypercontours. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 692–701. Springer, Heidelberg (2006)
12. Tomczyk, A., Szczepaniak, P.S.: Contribution of Active Contour Approach to Image Understanding. In: Proceedings of the IEEE International Workshop on Imaging Systems and Techniques – IST 2007, Krakow, Poland (2007), ISBN 1-4244-0965-9
13. Pedrycz, W., Loia, V.: P-FCM: A proximity-based fuzzy clustering. Fuzzy Sets and Systems 128, 21–41 (2004)
14. Tomczyk, A., Szczepaniak, P.S., Lis, B.: Generalized Multi-layer Kohonen Network and Its Application to Texture Recognition. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 760–767. Springer, Heidelberg (2004)
15. Lis, B., Szczepaniak, P.S., Tomczyk, A.: Multi-layer Kohonen Network and Texture Recognition. In: Grzegorzewski, P., Krawczak, M., Zadrony, S. (eds.) Soft Computing Tools, Techniques and Applications, Akademicka Oficyna Wydawnicza EXIT, Warszawa (2004)

KOSPI Time Series Analysis Using Neural Network with Weighted Fuzzy Membership Functions

Sang-Hong Lee[1] and Joon S. Lim[2]

[1,2] Division of Software, Kyungwon University, Korea
{shleedosa,jslim}@kyungwon.ac.kr

Abstract. Fuzzy neural networks have been successfully applied to generate predictive rules for stock forecasting. This paper presents a methodology to forecast the daily Korea composite stock price index (KOSPI) by extracting fuzzy rules based on the neural network with weighted fuzzy membership functions (NEWFM) and the minimized number of input features using the distributed non-overlap area measurement method. NEWFM supports the KOSPI time series analysis based on the defuzzyfication of weighted average method which is the fuzzy model suggested by Takagi and Sugeno. NEWFM classifies upward and downward cases of next day's KOSPI using the recent 32 days of $CPP_{n,m}$ (Current Price Position of day n : a percentage of the difference between the price of day n and the moving average of the past m days from day n-1) of KOSPI. In this paper, the Haar wavelet function is used as a mother wavelet. The most important five input features among $CPP_{n,m}$ and 38 numbers of wavelet transformed coefficients produced by the recent 32 days of $CPP_{n,m}$ are selected by the non-overlap area distribution measurement method. For the data sets, from 1991 to 1998, the proposed method shows that the average of accuracy rate is 67.62%.

Keywords: fuzzy neural networks, weighted average defuzzification, wavelet transform, KOSPI, nonlinear time series.

1 Introduction

With the help of recent development of information technology, the new research attempt using fuzzy theory, neural network, and wavelet transform is made in the wide range of economic field, such as money and banking, exchange rate, stock price, credit standing of companies, and etc. Fuzzy neural network (FNN) is the combination of neural network and fuzzy set theory, and provides the interpretation capability of hidden layers using knowledge based the fuzzy set theory [14-17]. Various FNN models with different algorithms such as learning, adaptation, and rule extraction were proposed as an adaptive decision support tool in the field of pattern recognition, classification, and forecasting [4-6][18]. Chai proposed the economic turning point forecasting using fuzzy neural network [19] and Gestel proposed the financial time series prediction using least squares support vector machines within the evidence framework [11].

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 53–62, 2008.

Stock forecasting has been studied using AI (artificial intelligence) approach such as artificial neural networks and rule-based systems. Artificial neural networks are to support for training stock data and rule-based systems are to support for making a decision in the higher and lower of daily change. Bergerson and Wunsch [10] combined of neural network and rule-based system in the S&P 500 index futures market. Xiaohua Wang [1] proposed the time delay neural network (TDNN). TDNN explored the usefulness of volume information in the explanation of the predictability of stock index returns. Kim proposed support vector machines (SVMs) to predict a financial time series and compared SVMs with back-propagation neural networks [2] and a genetic algorithm approach to instance selection (GAIS) in artificial neural networks (ANN) [12]. A new forecasting model based on neural network with weighted fuzzy membership functions (NEWFM) [3] concerning forecasting of stock by KOSPI using the Haar wavlet transforms (WT) is implemented in this paper.

In this paper, the five extracted input features are presented to forecast the daily Korea composite stock price index (KOSPI) using the Haar WT and the neural network with weighted membership functions (NEWFM), and the non-overlap area distribution measurement method [3]. The method extracts minimum number of input features each of which constructs an interpretable fuzzy membership function. The five features are interpretably formed in weighted fuzzy membership functions preserving the disjunctive fuzzy information and characteristics, locally related to the time signal, the patterns of KOSPI. The five features are extracted by the non-overlap area measurement method validated by the wine benchmarking data in University of California, Irvine (UCI) Machine Learning repository [7]. This study is to forecast the higher and lower of daily changes of KOSPI. They are classified as “1” or “2” in the data of KOSPI. “1” means that the next day’s index is lower than today’s index. “2” means that the next day’s index is higher than today’s index.

In this paper, the total numbers of samples are 2348 trading days used in Kim [12], from January 1991 to December 1998. Kim divided the samples into eight data sets according to the trading year. About 80% of annual trading data are used for training and 20% for testing, from January 1991 to December 1998. Kim used a genetic algorithm to search for optimal or near-optimal connection weights and relevant instances for ANN, and then the average of accuracy rate was 65.47% [12]. The result showed that the accuracy rate of a genetic algorithm approach to instance selection was higher than that of the conventional ANN with the genetic algorithm.

In this paper, the most important five input features are selected by non-overlap area measurement method [7]. The 5 generalized features are used to generate the fuzzy rules to forecast the next day’s directions of the daily changes of KOSPI. NEWFM shows that the average of accuracy rate is 67.62%. The fuzzy model suggested by Tagaki and Sugeno in 1995 can represent nonlinear system such as stock time series [13] and business cycle [19].

2 Wavelet Transforms

The wavelet transform (WT) is a transformation to basis functions that are localized in scale and in time as well. WT decomposes the original signal into a set of coefficients that describe frequency content at given times. The continuous wavelet transform (CWT) of a continuous time signal $x(t)$ is defined as:

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \psi\left(\frac{t-b}{a}\right) x(t)dt$$

where $\psi((t\text{-}b)/a)$ is the analyzing wavelet function. The transform coefficients $T(a,b)$ are found for both specific locations on the signal, $t=b$, and for specific wavelet periods (which are scale function of a). The CWT is defined as the dyadic wavelet transform (DWT), if a is discretized along the dyadic sequence 2^i where i = 1, 2, … . The DWT can be defined as [8]:

$$S_{2^i} x(n) = \sum_{k \in Z} h_k S_{2^{i-1}} x(n - 2^{i-1} k)$$

$$W_{2^i} x(n) = \sum_{k \in Z} g_k S_{2^{i-1}} x(n - 2^{i-1} k)$$

where S_{2^i} is a smoothing operator, W_{2^i} is the digital signal $x(n)$, $i \in Z$ (Z is the integral set), and h_k and g_k are coefficients of the corresponding low pass and high pass filters. A filtered signal at level i is down-sampled reducing the length of a signal at level i-1 by a factor of two, and generating approximation (a_i) and detail coefficients (d_i) at level i.

This paper proposes $\text{CPP}_{n,m}$ (Current Price Position) as a new technical indicator to forecast the next day's directions of the daily changes of KOSPI. $\text{CPP}_{n,m}$ is a current price position of day n on a percentage of the difference between the price of day n and the moving average of the past m days from day n-1. $\text{CPP}_{n,m}$ is calculated by

$$CPP_{n,m} = ((C_n - MA_{n-1,n-m}) / MA_{n-1,n-m}) \times 100$$

where C_n is the closing price of day n and $MA_{n\text{-}1,n\text{-}m}$ is the moving average of the past m days from day n-1.

In this paper, the Haar wavelet function is used as a mother wavelet. The Haar wavelet function makes 38 numbers of approximations and detail coefficients from $\text{CPP}_{n,5}$ to $\text{CPP}_{n\text{-}31,5}$ to extract input features. 38 numbers of approximations and detail coefficients consist of 16 detail coefficients at level 1, 8 detail coefficients at level 2, 4 detail coefficients and 4 approximations at level 3, 2 detail coefficients and 2 approximations at level 4, and 1 detail coefficient and 1 approximation at level 5.

The neural network with weighted membership functions (NEWFM) and the non-overlap area distribution measurement method [3] is used to extract minimum number of input features among 39 numbers of features. The following five minimum input features are extracted:

1) d_1 among 16 detail coefficients at level 1
2) d_1 among 8 detail coefficients at level 2
3) a_1 among 4 approximations at level 3
4) a_1 among 1 approximation at level 5
5) $\text{CPP}_{n,5}$

3 Neural Network with Weighted Fuzzy Membership Function (NEWFM)

3.1 The Structure of NEWFM

Neural network with weighted fuzzy membership function (NEWFM) is a supervised classification neuro-fuzzy system using bounded sum of weighted fuzzy membership functions (BSWFM in Fig. 2) [3][9]. The structure of NEWFM, illustrated in Fig. 1, comprises three layers namely input, hyperbox, and class layer. The input layer contains n input nodes for an n featured input pattern. The hyperbox layer consists of m hyperbox nodes. Each hyperbox node B_l to be connected to a class node contains n BSWFMs for n input nodes. The output layer is composed of p class nodes. Each class node is connected to one or more hyperbox nodes. An hth input pattern can be recorded as I_h={A_h=(a_1, a_2, … , a_n), *class*}, where *class* is the result of classification and A_h is n features of an input pattern.

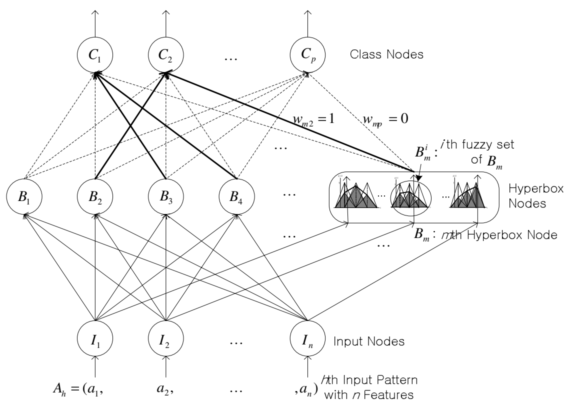


Fig. 1. Structure of NEWFM

The connection weight between a hyperbox node B_l and a class node C_i is represented by w_{li}, which is initially set to 0. From the first input pattern I_h, the w_{li} is set to 1 by the winner hyperbox node B_l and class i in I_h. C_i should have one or more than one connections to hyperbox nodes, whereas B_l is restricted to have one connection to a corresponding class node. The B_l can be learned only when B_l is a winner for an input I_h with class i and w_{li} = 1.

3.2 Learning Scheme

A hyperbox node B_l consists of n fuzzy sets. The ith fuzzy set of B_l, represented by B_l^i, has three *weighted fuzzy membership functions* (WFM, grey triangles ω_{l1}^i, ω_{l2}^i, and ω_{l3}^i in Fig. 2) which randomly constructed before learning. The *bounded sum of three weighted fuzzy membership functions* (BSWFM, bold line in Fig. 2) of B_l^i

combines the fuzzy characteristics of the three WFMs. The BSWFM value of B_l^i, denoted as $BS_l^i(.)$, is calculated by

$$BS_l^i(a_i) = \sum_{j=1}^{3} \omega_{l\,j}^i(a_i),$$

where a_i is an ith feature value of an input pattern A_h for B_l^i.

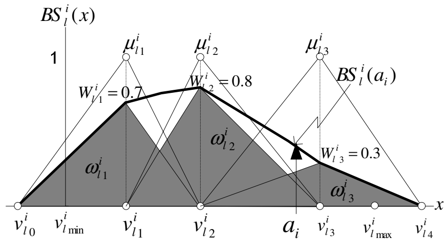


Fig. 2. An Example of Bounded Sum of Weighted Fuzzy Membership Functions (BSWFM, Bold Line) of B_l^i and $BS_l^i(a_i)$

The winner hyperbox node B_l is selected by the *Output*(B_l) operator. Only the B_l, that has the maximum value of *Output*(B_l) for an input I_h wtih class i and w_{li} = 1, among the hyperbox nodes can be learned. For the hth input A_h=(a_1, a_2, … , a_n) with n features to the hyperbox B_l, output of the B_l is obtained by

$$Output\ (B_l) = \frac{1}{n}\sum_{i=1}^{n} BS_l^i(a_i).$$

Then, the selected winner hyperbox node B_l is learned by the *Adjust*(B_l) operation. This operation adjusts all B_l^is according to the input a_i, where i=1, 2, … , n. The membership function weight $W_{l\,j}^i$ (where 0≤ $W_{l\,j}^i$ ≤1 and j=1,2,3) represents the strength of $\omega_{l\,j}^i$. Then a WFM $\omega_{l\,j}^i$ can be formed by ($v_{l\,j-1}^i, W_{l\,j}^i, v_{l\,j+1}^i$). As a result of *Adjust*(B_l) operation, the vertices $v_{l\,j}^i$ and weights $W_{l\,j}^i$ in Fig. 3 are adjusted by the following expressions:

$$\begin{aligned} v_{l\,j}^i &= v_{l\,j}^i + s\times\alpha\times E_{l\,j}^i \times \omega_{l\,j}^i(a_i) \\ &= v_{l\,j}^i + s\times\alpha\times E_{l\,j}^i \times \mu_{l\,j}^i(a_i)\times W_{l\,j}^i, where \end{aligned}$$

$$\begin{cases} s=-1, E_{l\,j}^i = \min(\left|v_{l\,j}^i - a_i\right|, \left|v_{l\,j-1}^i - a_i\right|),\ \ if\ v_{l\,j-1}^i \le a_i < v_{l\,j}^i \\ s=1, E_{l\,j}^i = \min(\left|v_{l\,j}^i - a_i\right|, \left|v_{l\,j+1}^i - a_i\right|),\ \ if\ v_{l\,j}^i \le a_i < v_{l\,j+1}^i \\ E_{l\,j}^i = 0,\ \ otherwise \end{cases}$$

$$W_{l\,j}^i = W_{l\,j}^i + \beta \times (\mu_{l\,j}^i(a_i) - W_{l\,j}^i),\ where$$

the α and β are the learning rates for $v_{l\,j}^{i}$ and $W_{l\,j}^{i}$ respectively in the range from 0 to 1 and j=1,2,3.

Fig. 3 shows BSWFMs before and after *Adjust*(B_l) operation for B_l^i with an input a_i. The weights and the centers of membership functions are adjusted by the *Adjust*(B_l) operation, e.g., $W_{l\,1}^{i}$, $W_{l\,2}^{i}$, and $W_{l\,3}^{i}$ are moved down, v_{l1}^{i} and v_{l2}^{i} are moved toward a_i, and $v_{l\,3}^{i}$ remains in the same location.

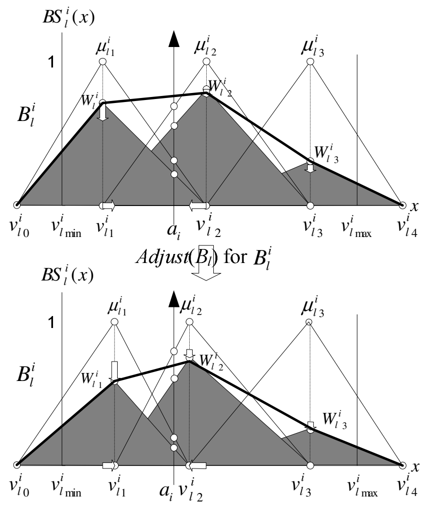


Fig. 3. An Example of Before And After *Adjust*(B_l) Operation for B_l^i

The *Adjust*(B_l) operations are executed by a set of training data. If the classification rate for a set of test data is not reached to a goal rate, the learning scheme with *Adjust*(B_l) operation is repeated from the beginning by randomly reconstructing all WFMs in B_l s and making all connection weights to 0 (w_{li} = 0) until the goal rate is reached.

4 Experimental Results

In this section, the total numbers of samples are 2348 trading days used in Kim [12], from January 1991 to December 1998. Kim divided the samples into eight data sets according to the trading year. Table 1 displays that about 80% of annual trading data are used for training and 20% for testing, from January 1991 to December 1998.

Kim used GANN (the conventional Artificial Neural Network with the Genetic Algorithm) and GAIS (the Genetic Algorithm approach to Instance Selection for Artificial Neural Network) to forecast the next day's directions of the daily changes of KOSPI. Table 2 displays the comparison of the numbers of features used in Kim and

Table 1. Number of instances used in Kim

Set	Year								Total
	1991	1992	1993	1994	1995	1996	1997	1998	
Training instances for GANN	234	236	237	237	235	235	234	234	1882
Selected instances for GAIS	74	71	87	66	93	86	93	85	655
Holdout instanced for GANN & GAIS	58	58	59	59	58	58	58	58	466

Table 2. Comparisons of Features of KIM With NEWFM

Kim	NEWFM
12 features such as CCI, RSI, Stochastic, and etc	5 features such as $CPP_{n,5}$ and 4 numbers of approximations and detail coefficients from $CPP_{n,5}$ to $CPP_{n-31,5}$

NEWFM. Kim used 12 features such as CCI, RSI, Stochastic, and etc. NEWFM uses 5 features, which consist of 4 numbers of approximations and detail coefficients made by the Haar wavelet function and $CPP_{n,5}$.

The accuracy of NEWFM is evaluated by the same data sets which were used in GANN. Table 3 displays the accuracy rates for about 20% of each eight data sets, from January 1991 to December 1998.

Table 3. Comparisons of Performance Results for KIM with NEWFM

Year	GANN	GAIS	NEWFM
	Accuracy (%)	Accuracy (%)	Accuracy (%)
1991	53.45	72.41	72.41
1992	56.90	58.62	62.07
1993	59.32	59.32	62.72
1994	57.63	61.02	61.02
1995	65.52	67.24	65.52
1996	65.52	77.59	79.31
1997	58.62	58.62	67.24
1998	56.90	68.97	70.69
Average	59.23	65.47	67.62

The five generalized features extracted from 39 numbers of input features are selected by non-overlap area measurement method [3]. The five generalized features are used to generate the fuzzy rules (BSWFM) to forecast the time series of KOSPI.

In this experiment, two hypoboxes are created for classification. While a hyperbox which contains a set of line (BSWFM) in Fig. 4 is a rule for class 1 (lower phase), the other hyperbox which contains a set of lines (BSWFM) is another rule for class 2 (upper phase). The graph in Fig. 4 is obtained from the training process of the NEWFM program and shows the difference between lower phase and upper phase for each input feature graphically. Lower phase means that the next day's index is lower than today's index. Upper phase means that the next day's index is higher than today's index.

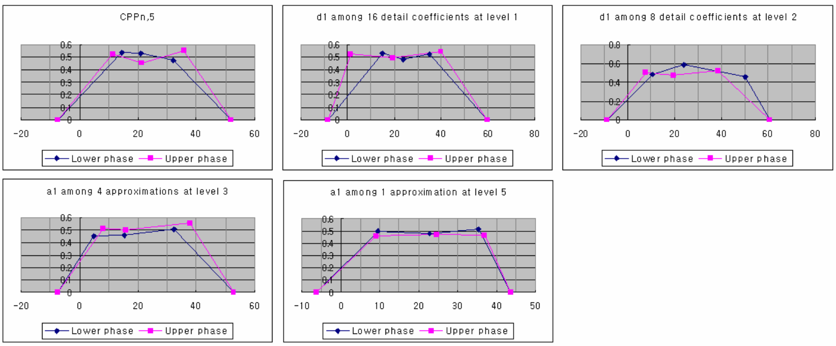


Fig. 4. Trained BSWFM of the Generalized Five Features for Lower Phase and Upper Phase Classification of KOSPI in 1998 among 8 years

The forecasting result of NEWFM can be represented by the trend line using the defuzzyfication of weighted average method which is the fuzzy model suggested by Takagi and Sugeno [13]. The Fig.5 shows the trend line of KOSPI in 1998. This result generally demonstrates the similar fluctuations with KOSPI.

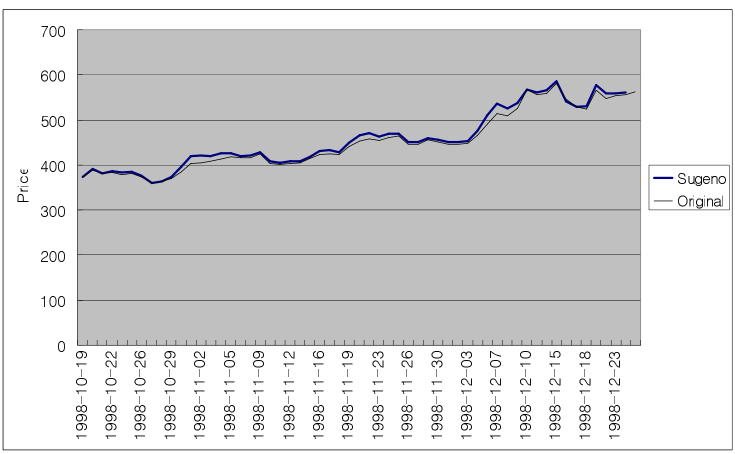


Fig. 5. Comparison of the Original KOSPI and the Fuzzy Model Suggested by Tagaki and Sugeno in 1998 among 8 years

5 Concluding Remarks

This paper proposes a new forecasting model based on neural network with weighted fuzzy membership function (NEWFM) and the time series of KOSPI based on the defuzzyfication of weighted average method which is the fuzzy model suggested by Takagi and Sugeno [13]. NEWFM is a new model of neural networks to improve forecasting accuracy rates by using self adaptive weighted fuzzy membership functions. The degree of classification intensity is obtained by bounded sum of weighted fuzzy membership functions extracted by NEWFM, and then weighted average defuzzification is used for forecasting the time series of KOSPI.

In this paper, the Haar wavelet function is used as a mother wavelet to extract input features. The five input features extracted by the non-overlap area distribution measurement method [3] are presented to forecast KOSPI using the Haar WT. The average result of accuracy rate from January 1991 to December 1998 is 67.62%. In Table 3, NEWFM outperforms GANN by 8.39% and GAIS by 2.15% for the holdout data.

Although further study will be necessary to improve the accuracy of the stock forecasting capability, the buy-and-hold investment strategy can be planned by the trend line of KOSPI. To improve the accuracy of the stock forecasting capability, some kinds of statistics such as CCI, normal distribution, and etc will be needed to study.

References

1. Wang, X., Phua, P.K.H., Lin, W.: Stock market prediction using neural networks: Does trading volume help in short-term prediction? In: Neural Networks, 2003. Proceedings of the International Joint Conference on Volume 4, July 20-24, 2003, vol. 4, pp. 2438–2442 (2003)
2. Kim, K.-j.: Financial time series forecasting using support vector machines. Neurocomputing 55, 307–309 (2003)
3. Lim, J.S., Ryu, T.-W., Kim, H.-J., Gupta, S.: Feature Selection for Specific Antibody Deficiency Syndrome by Neural Network with Weighted Fuzzy Membership Functions. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 811–820. Springer, Heidelberg (2005)
4. Ishibuchi, H., Nakashima, T.: Voting in Fuzzy Rule-Based Systems for Pattern Classification Problems. Fuzzy Sets and Systems 103, 223–238 (1999)
5. Nauk, D., Kruse, R.: A Neuro-Fuzzy Method to Learn Fuzzy Classification Rules from Data. Fuzzy Sets and Systems 89, 277–288 (1997)
6. Setnes, M., Roubos, H.: GA-Fuzzy Modeling and Classification: Complexity and Performance. IEEE Trans. Fuzzy Systems 8(5), 509–522 (2000)
7. Lim, J.S., Gupta, S.: Feature Selection Using Weighted Neuro-Fuzzy Membership Functions. In: The 2004 International Conference on Artificial Intelligence(IC-AI 2004), Las Vegas, Nevada, USA, June 21-24, 2004, vol. 1, pp. 261–266 (2004)
8. Mallat, S.: Zero Crossings of a Wavelet Transform. IEEE Trans. on Information Theory 37, 1019–1033 (1991)
9. Lim, J.S., Wang, D., Kim, Y.-S., Gupta, S.: A neuro-fuzzy approach for diagnosis of antibody deficiency syndrome. Neurocomputing 69(7-9), 969–974 (2006)

10. Bergerson, K., Wunsch, D.C.: A commodity trading model based on a neural network-Expert system hybrid. In: Proceedings of the IEEE International Conference on Neural Networks, pp. I289–I293 (1991)
11. Gestel, T.V., et al.: Financial Time Series Prediction Using Least Squares Support Vector Machines Within the Evidence Framework. IEEE Trans. Neural Networks 12(4), 809–821 (2001)
12. Kim, K.-j.: Artificial neural networks with evolutionary instance selection for financial forecasting. Expert System with Applications 30, 519–526 (2006)
13. Tagaki, T., Sugeno, M.: Fuzzy Identification of System and Its Applications to Modeling and Control. IEEE Trans. SMC(15), 116–132 (1985)
14. Carpenter, G.A., Grossberg, S., Reynolds, J.: ARTMAP : Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks 4, 565–588 (1991)
15. Jang, R.: ANFIS : Adaptive network-based fuzzy inference system. IEEE Trans. Syst., Man, Cybern. 23, 665–685 (1993)
16. Wang, J.S., Lee, C.S.G.: Self-Adaptive Neuro-Fuzzy Inference System for Classification Applications. IEEE Trans., Fuzzy Systems 10(6), 790–802 (2002)
17. Simpson, P.: Fuzzy min-max neural networks-Part 1 : Classification. IEEE Trans., Neural Networks 3, 776–786 (1992)
18. Lim, J.S.: Finding Fuzzy Rules by Neural Network with Weighted Fuzzy Membership Function. International Journal of Fuzzy Logic and Intelligent Systems 4(2), 211–216 (2004)
19. Chai, S.H., Lim, J.S.: Economic Turning Point Forecasting Using Fuzzy Neural Network and Non-Overlap Area Distribution Measurement Method. The Korean Economic Association 23(1), 111–130 (2007)

Facilitating MAS Complete Life Cycle through the Protégé-Prometheus Approach

Marina V. Sokolova[1,2] and Antonio Fernández-Caballero[1]

[1] Universidad de Castilla-La Mancha, Escuela Politécnica Superior de Albacete & Instituto de Investigación en Informática de Albacete, 02071-Albacete, Spain
[2] Kursk State Technical University, Kursk, ul.50 Let Oktyabrya, 305040, Russia
smv1999@yandex.ru,caballer@dsi.uclm.es

Abstract. The approach of this paper aims to support the complete multi-agent systems life cycle, integrated by two existing and widely accepted tools, Protégé Ontology Editor and Knowledge-Base Framework, and Prometheus Development Kit. A general sequence of steps facilitating application creation is proposed in this paper. We propose that it seems reasonable to integrate all traditional software development stages into one single methodology. This view provides a general approach for MAS creation, starting with problem definition and resulting in program coding, deployment and maintenance. The proposal is successfully being applied to situation assessment issues, which has concluded in an agent-based decision-support system for environmental impact evaluation.

Keywords: Multi-agent systems, Software life cycle, Methodologies.

1 Introduction

Nowadays there are many works and approaches dedicated to multi-agent systems (MAS) development, which pay attention to internal MAS functionality, reasoning and its coding. Creation, deployment and post-implementation of MAS as software products is a complex process, which passes through a sequence of stages forming its life cycle [13][21]. Every step of the life cycle process has to be supported and provided by means of program tools and methodologies. In case of MAS development, in our opinion there is still no solution to a unified approach to cover all the stages. However, there are some previous works dedicated to this issue [2][12]. For instance, de Wolf and Holvoet [2] have introduced a methodology in the context of standard life cycle model, with accent to decentralization and macroscopic view of the process. The authors offer their approach on the assumption that the research task has already been defined, omitting the problem definition and domain analysis stages of MAS development process. But, a complete software development in case of MAS should be based on the following steps: (1) domain and system requirements analysis, (2) design, (3) implementation, (4) verification, and, (5) maintenance [9][12].

Some well known alternative agent-oriented software engineering methodologies, including MaSE [1] , Gaia [23], MASDK [6], Prometheus [15], Tropos [4],

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 63–72, 2008.

INGENIAS [5], among others, support some of the cited stages of MAS life cycle process. Nonetheless, these methodologies often work under the condition that the developer has already defined the problem and determined the goals and the tasks of the system. However, domain analysis is a crucial stage and has to be scrutinizingly examined and planned. Indeed, the whole deployed system functionality and efficiency depends on how precisely the problem was defined and the domain ontology was elaborated. In the most general case, when a MAS is distributed and has to deal with heterogeneous information, the domain analysis becomes even more important.

Therefore, it seems reasonable to integrate all the software development stages into one single methodology, which should provide a general approach to MAS creation, starting with the problem definition and resulting in program coding, deployment and maintenance. As a tool for the system and domain requirements, we suggest using an OWL-language based toolkit, as OWL has become a standard for ontologies description [14]. The Protégé Ontology Editor and Knowledge-Base Framework [16] complies a set of procedures for ontology creation and analysis, offering a set of plug-ins covering viewers, problem-solving methods, knowledge trees, converters, etc. According to our proposal, ontologies can be represented by means of Protégé and later may be incorporated into MAS. In order to provide the following stages with tools we have tested different methodologies. We came to the conclusion to use the Prometheus Development Tool (PDT) [17], which provides a wide range of possibilities for MAS planning and implementation: the system architecture, the system entities, their internals and communications within the system and with outer entities. The most important advantages of PDT are an easy understandable visual interface and the possibility to generate code for JACK^{TM} Intelligent Agents [11]. The proposal is summed up in Fig. 1.

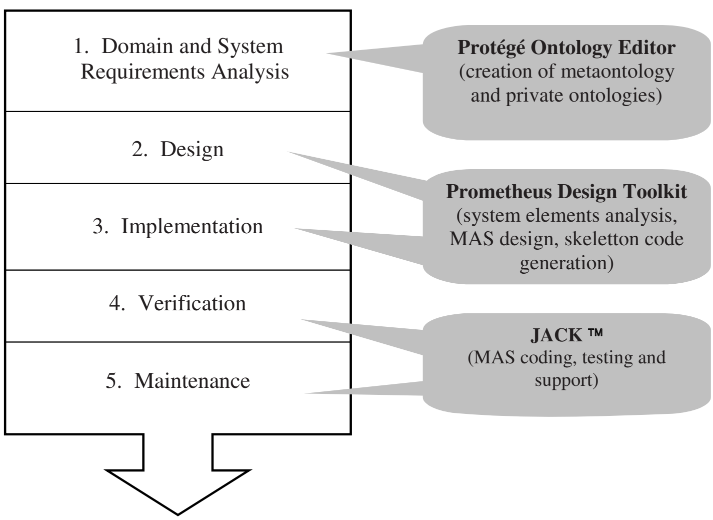


Fig. 1. The Protégé-Prometheus approach applied to the MAS life cycle

The paper is organized as follows. In section 2 the metaontology creation realized in Protégé is described and in section 3 the MAS designed in PDT is introduced. In section 4 our intention to implement the ideas for further usage of the integrated methodology are briefly explained.

2 Domain and MAS Requirements in Protégé

Ontology creation may be viewed as a crucial step in MAS design as it determines the system knowledge area and potential capabilities [7]. In the first part of this article a model of distributed metaontology that serves as a framework for MAS design is proposed. Its components - private ontologies - are described in extensive with respect to an application area and in terms of the used semantics.

When defining an ontology O in terms of an algebraic system, we have the following three attributes:

$$O = (C, R, \Omega) \tag{1}$$

where C is a set of concepts, R is a set of relations among the concepts, and Ω a a set of rules. The principal point of MAS is to determine the rules Ω and to evaluate them. Formula (2) proposes that the ontology for the domain of interest (or the problem ontology) may be described by offering proper meanings to C, R and Ω.

The model of the metaontology that we have created consists of five components, or private ontologies: the “Domain Ontology”, the “Task Ontology”, the “Ontology of MAS”, the “Interaction Ontology” and the “Agent Ontology”.

In first place, the “Domain Ontology”, includes the objects of the problem area, the relations between them and their properties. It determines the components C and R of expression (2), which is detailed as:

$$OD =< I, C, P, V, Rs, Rl > \tag{2}$$

where the set C (see formula (2)) is represented by two components: Individuals (I) and Classes (C), which reflect the hierarchical structure of the objects of the problem area; P - are class properties; V - are the properties values; Rs - are values restrictions; Rl embodies the set R, and includes rules which state how to receive new individuals for the concrete class.

The “Task Ontology” contains information about tasks and respective methods, about the pre-task and post-task conditions, and informational flows for every task. The “Task Ontology” has the following model:

$$OT =< T, M, In, Ot, R > \tag{3}$$

where T is a set of tasks to be solved in the MAS, and M is a set of methods or activities related to the concrete task, In and Ot are input and output data flows, R is a set of roles that use the task. Component R is inherited from the “Ontology of MAS” through the property `belong to role`. The tasks are shared and accomplished in accordance with an order.

The “Ontology for MAS” architecture is stated as:

$$OA =< L, R, IF, Or > \tag{4}$$

where L corresponds to the logical levels of the MAS (if required), R is a set of determined roles, IF is a set of the corresponding input and output information represented by protocols. Lastly, the set Or determines the sequence of execution for every role (orders).

The interactions between the agents include an initiator and a receiver, a scenario and the roles taken by the interacting agents, the input and output information and a common communication language. They are stated in the “Interaction Ontology” as:

$$OI =< In, Rc, Sc, R, In, Ot, L > \tag{5}$$

Actually, as In and Rc Initiator and Receiver, respectively, of the interaction we use agents. The component Sc corresponds to protocols. R is a set of roles that the agents play during the interaction. In and Ot are represented by informational resources, required as input and output, respectively. Language L determines the agent communication language (ACL).

In our approach BDI agents [3], which are represented by the “Agent Ontology”, are modeled. Hence, every agent is described as a composition of the following components:

$$Agent =< B, D, I > \tag{6}$$

Every agent has a detailed description in accordance with the given ontology, which is offered in a form of BDI cards, in which the pre-conditions and post-conditions of agent execution and the necessary conditions and resources for the agent successful execution are stated. Evidently, B, D and I stand for Believes, Desires and Intentions, respectively.

Metaontology is a specification for further MAS coding; it provides the necessary details about the domain, and describes system entities and functionality. It includes five components:

$$MO =< OD, OT, OA, OI, Agent > \tag{7}$$

where OD stands for the ”‘Domain Ontology”’, OT for the “Task Ontology”, OA “Ontology for MAS” architecture, OI is the “Interaction Ontology”, and, $Agent$ is the “Agent Ontology”.

Private ontologies mapping is made through slots of their components. So, the “Agent Ontology” has four properties:

1. `has intentions` - which contains individuals of the methods “M” class from the “Task Ontology”.
2. `has believes` - which contains individuals from the “Domain Ontology”.
3. `has desires` - which contains individuals from the “Task Ontology”.
4. `has type` - which contains variables of *String* type.

There is a real connection between the "Task Ontology" and the "Domain Ontology". The *OT*, in turn, refers to the "Ontology of MAS" (*OA*), which is formally described by four components. The first two

- `level value`
- `order`

contain values of *Integer* type, which determine the logical level number and the order of execution for every role. Roles (*R*) are the individuals of the named ontology. The next two properties

- `has input`
- `has output`

refer to individuals of the "Interaction Ontology"; in particular, to protocols, which manage communications. Their properties are of type *String*:

- `has scenario`,
- `language`,
- `roles at scenario`.

The "Interaction Ontology" slots named `has initiator` and `has receiver` are the individuals of the "Agent Ontology" (*Agent*). Thus, agents are linked

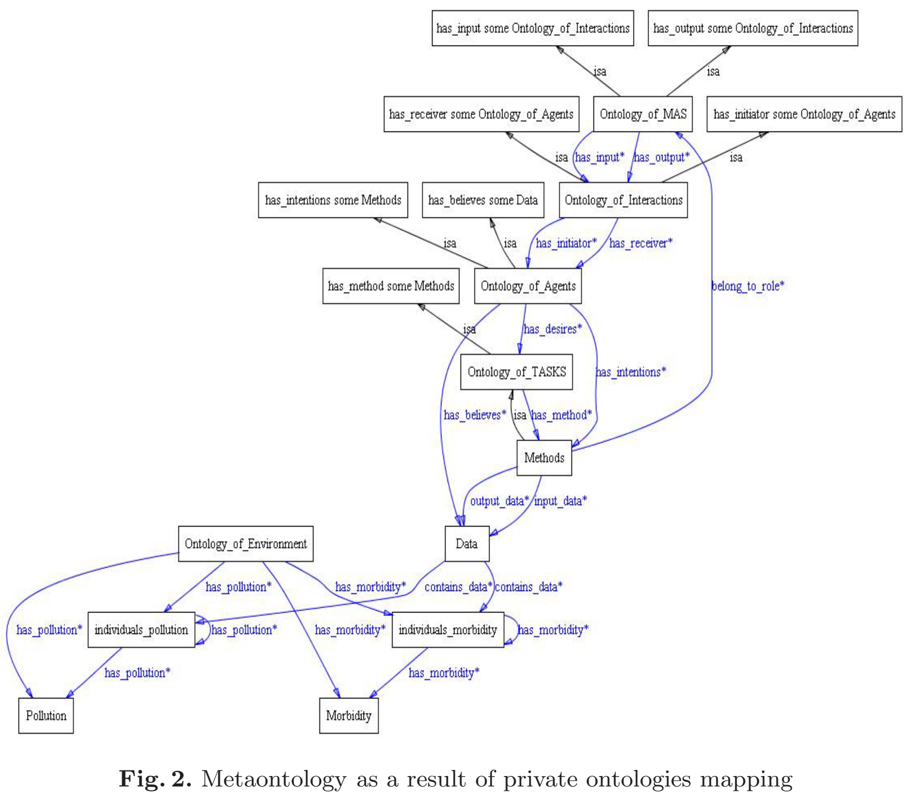


Fig. 2. Metaontology as a result of private ontologies mapping

to the proper protocols within the MAS. The *OD*, by means of its individuals - which contain data records - is connected to *Agent*, which uses the knowledge on the domain area as its believes. This way, the proposed metaontological model realized in Protégé covers the first four steps of the software development life cycle. The "System Elements Analysis" phase is covered by establishing the terminological basis for the further design.

3 System Design with Prometheus Development Tool

In order to validate the second step of our approach, we introduce a running example, consisting in an agent-based decision support system (ADSS) dedicated to monitoring environmental pollution information, analyzing this data, simulating with respect to health consequences, and making decisions for responsible system users [8][10].

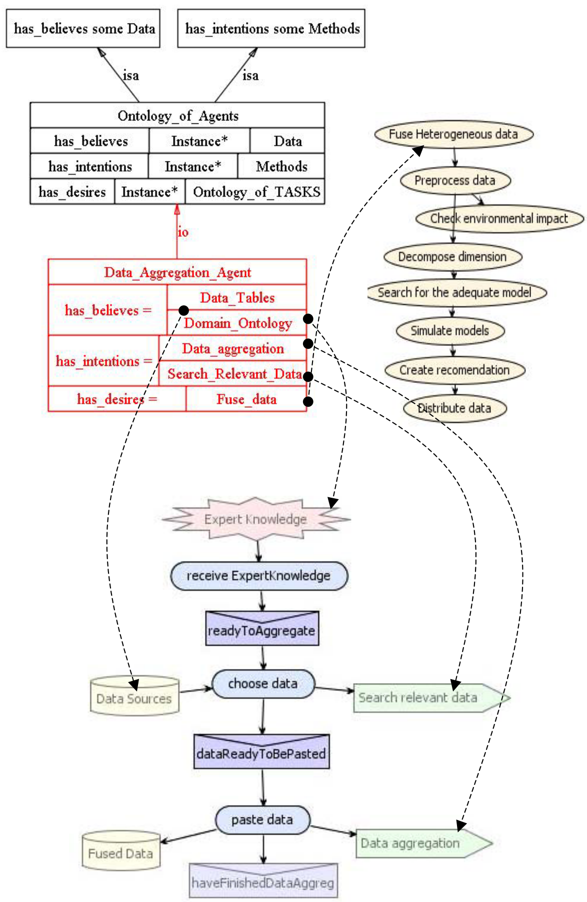


Fig. 3. An example of Protégé and Prometheus agent internals integration

Following the orientations described in section 2 for metaontology and private ontologies creation, the ADSS has been modeled as a logical three layer MAS architecture. The first layer is named **Information Fusion** and it acquires data from diverse sources, and preprocesses the initial information to be ready for further analysis. The second layer is named **Data Mining** and there are three roles at this level, dedicated to knowledge recovering through modeling, and calculation impact of various pollutants upon human health. The third layer, **Decision Making**, carries out a set of procedures including model evaluation, computer simulation, decision making and forecasting, based on the models created in the previous level. At every level of the system certain goals and tasks have to accomplished [18][19].

The system resembles a typical organizational structure. The agents are strictly dedicated to working with the stated sets of data sources. They solve the particular tasks and are triggered when all the necessary conditions are ful-

Table 1. Mapping between Protégé and Prometheus entities

Protégé Entity	**Prometheus Entity**	**Prometheus View**
"Domain Ontology"	Data	Data Coupling
"Task Ontology"		
Tasks Methods Input Output Roles	Goals Actions Data Data Roles	Goal Overview System Roles Data Coupling Data Coupling System Roles /Agent-Role Grouping
"Ontology of MAS" structure		
Levels Roles Information Flows Order	- Roles Protocols -	System Overview System Overview /Agent-Role Grouping System Overview System Overview
"Interaction Ontology"		
Initiators Receivers Scenarios Roles at Scenario Input Output Language	Agents Agents Scenarios - Data Data -	Agent Acquaintance Agent Acquaintance Scenarios - Data Coupling Data Coupling -
"Agent Ontology"		
Believes Desires Intentions	Data Perceptions Goals Actions	Data Coupling Analysis Overview Goal Overview System Overview /System Roles

filled, and there are positive messages from previously executed agents [22]. The system includes a set of roles, correlated with the main system functions and a set of agents related to each role. Actually, mostly every agent is associated to one role; only in case of "Function Approximation" role, there are two agents, one for data mining, and the other one for validation.

In Fig. 3 there is an illustration of the integration between metaontological concepts (and their properties) in Protégé and the Prometheus entities in the context of the agent internals. The "Data Aggregation agent" uses "Domain Ontology" and "Task Ontology", which are parts of the metaontology previously realized in Protégé. In order to closely analyze the integration of these methodologies, the mapping of the Protégé entities into Prometheus ones is shown in Table 1. For instance, Table 1 states similarities between entities of OD ("Domain Ontology") in Protégé and "Data" in Prometheus, which can be observed in the "Data Coupling diagram". The components of OT (see equation (3)) are converted into Prometheus entities and can be displayed as well. Some components of OA, such as *Levels* and *Order* do not have equivalents, as well as *Roles at Scenario* and *Language* components of OI, which serves more for a researcher during the MAS planning stage.

4 Conclusions

The integration of two existing and widely accepted tools, Protégé Ontology Editor and Knowledge-Base Framework, and Prometheus Development Kit, into a common methodology has been introduced in this paper. The Protégé Ontology Editor complies a set of procedures for ontology creation and analysis, offering a set of plug-ins such as viewers, problem-solving methods, knowledge trees, converters, and so on. To provide the following stages with tools, we have tested different methodologies, and finally decided to use the Prometheus Development Tool, which offers a wide range of possibilities for MAS planning and implementation, namely the system architecture, the system entities, their internals and communications within the system and with outer entities.

As the Prometheus methodology has been developed in collaboration with Agent Oriented Software, and a modified version of the Prometheus modeling language has been partially implemented in their JACK Intelligent Agents^{TM} development environment as a tool for visual modeling of the architectural design and plans, the next logical step of our approach is to implement under this environment. The JACK Design Tool is a software package for agent-based applications development in Java-based environment JACK.

Thus, the integrated approach covers all the stages of MAS planning and implementation, supporting them with tools and frameworks. The proposed fusion of methodologies, Protégé and Prometheus, was chosen because of the wide range of functions offered and their conformance to international standards. We believe that the common use of Protégé and Prometheus in complex developments would prevent researchers and developers from numerous misunderstandings. It should greatly help in domain requirements description, facilitating the complete

MAS development life cycle. However, other combinations of agent-oriented tools could be used, whenever it helps getting the same result and support during MAS development, deployment and maintenance.

Acknowledgements

Marina V. Sokolova is the recipient of Postdoctoral Scholarship 0000253836, Program II.E (Becas MAE) awarded by the Agencia Española de Cooperación Internacional of the Spanish Ministerio de Asuntos Exteriores y de Cooperación.

This work is supported in part by the Spanish Ministerio de Educación y Ciencia TIN2004-07661-C02-02 and TIN2007-67586-C02-02 grants, and the Junta de Comunida-des de Castilla-La Mancha PBI06-0099 grant.

References

1. DeLoach, S.A., Wood, M.F., Sparkman, C.H.: Multiagent systems engineering. International Journal of Software Engineering and Knowledge Engineering 11, 231–258 (2001)
2. de Wolf, T., Holvoet, T.: Towards a full life-cycle methodology for engineering decentralised multi-agent systems. In: Proceedings of the Fourth International Workshop on Agent-Oriented Methodologies, pp. 1–12 (2005)
3. Georgeff, M., Pell, B., Pollack, M., Tambe, M., Wooldridge, M.: The Belief-Desire-Intention model of agency. In: Rao, A.S., Singh, M.P., Müller, J.P. (eds.) ATAL 1998. LNCS (LNAI), vol. 1555, pp. 1–10. Springer, Heidelberg (1999)
4. Giunchiglia, F., Mylopoulos, J., Perini, A.: The Tropos software development methodology: Processes, models and diagrams. In: Third International Workshop on Agent-Oriented Software Engineering, pp. 162–173 (2002)
5. Gómez-Sanz, J., Pavon, J.: Agent oriented software engineering with INGENIAS. In: Mařík, V., Müller, J.P., Pěchouček, M. (eds.) CEEMAS 2003. LNCS (LNAI), vol. 2691, pp. 394–403. Springer, Heidelberg (2003)
6. Gorodetsky, V., Karsaev, O., Konushy, V., Mirgaliev, A., Rodionov, I., Yustchenko, S.: MASDK software tool and technology supported. In: International Conference on Integration of Knowledge Intensive Multi-Agent Systems, pp. 528–533 (2005)
7. Guarino, N., Giaretta, P.: Ontologies and knowledge bases: Towards a terminological clarification. In: Towards Very Large Knowledge Bases, pp. 25–32. IOS Press, Amsterdam (1995)
8. International Classification of Diseases (ICD) home page, http://www.who.int/classifications/icd/en/
9. ISO/IEC 12207 home page, http://www.iso.org/iso/
10. ISO 14031:1999 home page. Environmental management - Environmental performance evaluation - Guidelines, http://www.iso.org/iso/
11. Jack^{TM} Intelligent Agents home page, http://www.agent-software.com/shared/home/
12. Konichenko, A.V.: Distribution Information Systems Design Management. Rostov-Don Press, Russia (2005)
13. Marík, V., McFarlane, D.: Industrial adoption of agent-based technologies. Intelligent Systems 20, 27–35 (2005)

14. OWL Web Ontology Language home page. (2004), http://www.w3.org/TR/owl-features/
15. Padgham, L., Winikoff, M.: Prometheus: A pragmatic methodology for engineering intelligent agents. In: Proceedings of the Workshop on Agent Oriented Methodologies (Object-Oriented Programming, Systems, Languages, and Applications), pp. 97–108 (2002)
16. Prometheus Design Tool home page, http://www.cs.rmit.edu.au/agents/pdt/
17. Protégé home page, http://protege.stanford.edu/
18. Sokolova, M.V., Fernández-Caballero, A.: A multi-agent architecture for environmental impact assessment: Information fusion, data mining and decision making. In: 9th International Conference on Enterprise Information Systems, ICEIS 2007, vol. AIDSS, pp. 219–224 (2007)
19. Sokolova, M.V., Fernández-Caballero, A.: An agent-based decision support system for ecological-medical situation analysis. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2007. LNCS, vol. 4528, pp. 511–520. Springer, Heidelberg (2007)
20. Schelfthout, K., Holvoet, T.: ObjectPlaces: an environment for situated multi-agent systems. In: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 1500–1501 (2004)
21. Vasconcelos, W.W., Robertson, D.S., Agusti, J., Sierra, C., Wooldridge, M., Parsons, S., Walton, C., Sabater, J.: A lifecycle for models of large multi-agent systems. In: Wooldridge, M.J., Weiß, G., Ciancarini, P. (eds.) AOSE 2001. LNCS, vol. 2222, pp. 297–318. Springer, Heidelberg (2002)
22. Weiss, G.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. The MIT Press, Cambridge (2000)
23. Wooldridge, M., Jennings, N.R., Kinny, D.: The Gaia methodology for agent-oriented analysis and design. Journal of Autonomous Agents and Multi-Agent Systems 3, 285–312 (2000)

Holonic Multi-agent System Model for Fuzzy Automatic Speech / Speaker Recognition

J.J. Valencia-Jiménez and Antonio Fernández-Caballero

Departmento de Sistemas Informáticos, Universidad de Castilla-La Mancha
Escuela Politécnica Superior de Albacete, 02071-Albacete, Spain
caballer@dsi.uclm.es

Abstract. An automatic speech / speaker recognition (ASSR) system has to adapt to possible changes of speaker and environment conditions, and act as close as possible to the way a human recognizes speeches / speakers. This kind of very complex system has to deal with speech signals, looking for the integration of different information sources; and this is precisely the reason to use fuzzy logic. The main objective of this paper is the description of a robust, intelligent and adaptive system, modeled as a multi-agent system (MAS), forming a recursive hierarchy of MAS denominated holonic MAS.

Keywords: Holony, Multi-agent systems, Fuzziness, Speech recognition, Speaker recognition.

1 Introduction

Automatic speaker and/or speech recognition (ASSR) is a challenge per excellence for the Turing Test [17]. No computational efforts, throw as fast and exact results as a person does. The essential question is the integration of information that comes from diverse sources of knowledge (acoustic, phonetic, phonologic, lexical, syntactic, semantic and pragmatic). These sources incorporate great doses of uncertainty and errors due to the noise corruption of the input data. Also it is negatively influenced by the inherently ambiguous nature of natural language in the different types of knowledge that are integrated.

Therefore, the option to use fuzzy logic is very attractive [19][20]. The proposal is to quantify with words, more diffuse in its meaning, instead of with numbers. The process of symbolic inference proper of fuzzy logic rests on the denominated "decision trees" structures. Their nodes are linguistic variables that represent intermediate concepts; the connections between nodes represent the set of rules that relate the connected concepts. The membership functions express the probability that a value belongs to one or several linguistic variables. Recent is also the paradigm that possibly more affects the relations between vision and auditory decoding. We are introducing the Fuzzy Logical Model of Perception, developed throughout the last years by Massaro [10]. The fact that this paradigm is applied to decoding of the sensorial information presented to a person, leads to the application of its principles to the processing of the speech

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 73–82, 2008.

signal with the knowledge sources previously mentioned. Another paradigm with ample development in the field of Artificial Intelligence is the use of neuronal networks. In fact, neuronal networks are so intimately bound to fuzzy logic. For a long time architectures and hybrid systems denominated *neurofuzzy* have been developed, taking advantage of the properties and flexibility of both paradigms.

The main objective of our proposal is the design of a speech recognition system independent of the speaker, and the identification of the speaker independently of the speech. The system has to be robust and intelligent, and able to adapt to the environmental circumstances of noise and the characteristics of the speaker. The system also has to deal with the difficult analysis of the pronunciation of the speaker. For it, due to its versatility, the option of using the holonic multi-agent systems (MAS) [18] paradigm is considered. The remainder of this paper introduces the holonic multi-agent architecture, its bases and what are the more interesting reasons to use it, as well as a preliminary design of the fuzzy recognition system, using Prometheus [12] MAS design methodology.

2 Holonic Multi-agent Architecture

A holonic architecture is based on the model of distributed systems architectures. It is a solution based on the theory of complex adaptable systems, and the name comes from the combination of the Greek word " holos" (everything) with the suffix "on", that conceives the idea of particle or part of something. What impelled Koestler [8] Koestler to propose the concept of holon were mainly two observations: (1) First, it is easier to construct a complex system when it is composed of intermediate elements - or subsystems - that are stable. Also, complex systems, like biological organisms or societies, always are structured as a stable hierarchy of subsystems in multiple levels. These, as well, are recursively divided in subsystems of an inferior order. Structurally it is not a simple aggregation of elementary parts, and functionally it is not a global behavior like a superposition of behaviors of elementary units. Therefore a real "synergy" exists. (2) Second, when hierarchies are analyzed, it is discovered that, although it is easy to identify "sub-all" and "parts" of the "all", the "all" and the "parts" do not exist in an absolute sense, because there is a double nature of "sub-all/part".

Holonic systems are modeled in terms of components (holons) that possess their own identity and at the same time belong of a greater set. This superior set is known as holarchy [13]. Holons are self-contained with respect to their subordinated parts and are simultaneously dependent parts when they are observed from superior hierarchic levels. Therefore, the word holarchy denotes hierarchic organizations of holons with a recursive structure. Holarchy guarantees stability, predictability and global optimization of the hierarchic control. Simultaneously it provides flexibility and adaptability [3][4][5][7], since, in an independent way, each holon is able to adapt to events and to cooperate with others holons. This could directly be compared to a distributed system with cooperative nodes. Nevertheless, the key characteristic introduced is that holons cooperate in dynamic hierarchies. They reorganize every certain time or when there have been significant changes to reach a global objective. This description of adaptable systems agrees with the one of the MAS. Indeed, a holon could be an agent, a group of them or a complete independent multi-agent subsystem. Therefore, it may be stated that the holon introduces the concept of "recursive agent". And holarchy could consider the hierarchy of agents in that system for a given moment. Therefore, given the interchangeable characteristics of holarchy and multi-agent architecture, it seems reasonable to consider the design of a complex speech/speaker recognition system as the design of a holonic MAS, with all the advantages in performance, efficiency and scalability that it entails.

3 Fuzzy Speech / Speaker Recognition

Fuzzy logic applied to speech recognition or to speaker identification offers a great advantage. It may be included in any of the main approaches to the problem - phonetic-acoustic models and pattern matching - in almost any level of the different processing steps [1]. Fuzzy logic offers the possibilities of assimilating or of surpassing uncertainty, and in a similar way humans do, of offering different results, according to the knowledge level (phonetic, lexical, syntactic, semantic), without blocking the recognition process. A key question is the representation of the spectrogram of sounds, alophones, produced in speech. The fuzzification process consists of the transformation of the analogical signal spectrum into a fuzzy description.

The multidimensional representation influences directly in the amount of fuzzy variables that characterize the different properties of the sounds to recognize - the phonemes. Basing on the International Phonetic Alphabet (IPA) chart [6], the different features or dimensions for a sound can be inferred. A sound in this alphabet is fitted and its phoneme is determined. Also, by means of a combination of phonemes, the diphons (the union of two phonemes including the transition information from one to another), or the syllables (the set of two or the more phonemes around a phoneme representing a vowel, which is the common structure most used by the humans when processing spoken language) can be obtained. In these phonetic combinations it is necessary to consider the borders and overlapping of the phonetic units and the prosody. The previous and posterior phonemes give rise to co-articulation phenomena (energy and frequencies are transferred to the adjacent units). This causes an enormous computational explosion.

Of course, it is necessary to consider that there are numerous cases, in very homophonous languages, in which several different phonemes share sound features. In fuzzy set theory this is translated in that the intersection of the sets that identify their features is not the empty set. For almost any given phoneme the values in its different features are not exclusive, but for some central values of that phoneme's dimension some peripheral values exist. Also the opposite case can occur. Several phonemes of the international alphabet really correspond with the same one in a certain language. Therefore, a word could have several phonetic transcriptions in IPA, and, detecting any of them indifferently would be correct in that language. This can also be established by means of clustering algorithms to determine optimal patterns for each phonetic unit of the vocabulary and several values as optimal segmentation thresholds.

The process of speech and speaker recognition is made up of a series of not totally sequential stages [14]. They are rather functional phases, corresponding to phonetic, lexical, syntactic and semantic processing. In an initial stage of preprocessing or digitalization a pre-emphasis and filtering of the input sound signal is performed by means of a set of filters. Its fundamental task is to heighten the signal in a non-homogenous way, weighing and better discriminating in those frequency bands of the auditory spectrum that contribute more information to the recognition. Also the effects of atmospheric noise are diminished, increasing this way the signal to noise ratio (SNR). The result of this stage is a non-homogenous spectrogram, with the different values of intensity or sound amplitude for the different frequencies, throughout the time the pronunciation lasts. Next, the phases of frame blocking are carried out, where the signal is divided into frames of N samples, and windowing, where the discontinuities of the signal in the beginning and at the end of each frame are minimized.

Also end point detection (EDP) of the word, the silence, is necessary [16]. More specifically, the separation between the phonetic units has to be detected for a later segmentation and classification. Generally, the beginning and the end of a word is detected in speech, processing the samples of the already filtered input wave and compressing the useful information of the used phonetic unit. Usually this is based on the analysis of the linear prediction of the mean square error. The output is a vector of the samples of the phonetic unit. It is a difficult process, since in natural language pronunciation usually there are no pauses between these phonetic units. Much more if the units are smaller than words. This characteristic of the fluid speech, where the sound of a phoneme is influenced by the adjacent phonemes (mainly by the previous one), is called the "co-articulation" effect.

It is faster to perform the extraction and comparison of patterns in a fuzzy way when the numerical values have previously been transformed into fuzzy values. In some representations a progressive scale of colors is used - from black to white - that correspond to the linguistic variables that express the different intensities. Along with their corresponding membership functions, they are applied to each division of the frequencies range where the signal is sampled, providing a fuzzy value of the quantification of the signal intensity in that frequencies. In other cases quantifying linguistic variables and its corresponding membership functions are used. For instance, there is *nothing, very little, little, enough*, etc. We have also to consider that usually the duration of the phonetic unit, or phoneme, syllable or word, is identified with quantifiers such as *very short, short, medium, long* and *very long*. Therefore, the application of the "Dynamic Time Warping" algorithm [11] (also in a fuzzy way) is made in a rather economic way in terms of consumed resources and time, since much numerical precision is not necessary. A certain freedom of the values between the recognized pattern and the phonetic unit tests is allowed, fundamentally when a superior phonetic unit is constructed from more elementary ones. This occurs, for example, in the words composed of syllables, where the different lengths of the elementary units and their viable lengths have to be combined.

The comparison among the vectors is realized by means of inference rules, where the linguistic number and value of the antecedents are based on the fuzzy measures available in the own structure of the pattern. For that reason there also partial patterns or fuzzy patterns, since according to the human mechanism of analysis and speaker/speech recognition, the different data are combined or integrated in different ways. Sometimes, they are even omitted to deal with uncertainty and the result is a series of fuzzy decisions that are weighed to give the final decision [21]. The use of fuzzy logic in the comparison phase eases to separate the classes represented by several phonetic units. This segmentation eliminates ambiguity and more efficient decisions are generated to compare pattern templates with more dimensions and features in a flexible way. In addition, to determine a sequence of phonetic units (for example, syllables) with the intention to construct a superior unit (for example, a word) dynamic programming techniques are used. The computational load is considerably increased when trying to find the optimal paths by means of backtracking. By means of fuzzification and defuzzification this can be simplified enormously, since the different calculations of the costs of dynamic programming are more complicated than the different methods of calculation of centroids of defuzzification surfaces. Thus the individual values of the simpler phonetic units can be integrated directly to form the superior unit.

A very important part of an automatic patterns recognition and identification system is the initial training algorithm. In later phases the learning algorithm, automatic or supervised, must optimize the recognition of the patterns of the speech phonetic units as well as the phonetic patterns that identify the speaker. Without fuzzy logic, the test patterns are usually constructed using a clustering algorithm. It will be adaptive in case the independence of the speaker is looked for. Test patterns or templates are obtained from a superior number of training pronunciations for each individual class of phonetic unit. Nevertheless, genetic algorithms are also used to optimize the representation of the patterns. In the case of using fuzzy logic, the templates with partial and fuzzy representation are more adaptable than a representation with numeric parameters. In addition, learning using logical fuzzy is also based on the construction and modification of the inference rules from the observed data. A very effective method is ANFIS (Adaptive Neuro-Fuzzy Inference System) [2] where different genetic algorithms and different parameters to implement different learning models are applied [15].

4 Design of the Holonic Fuzzy Recognition System

In the preliminary design of the holonic multi-agent system the Prometheus methodology has been used [12]. The main advantage in the analysis and design of holonic systems with respect to multi-agent systems is that a holon allows differentiating the roles in a more compact way and obtaining affine behaviors from the very beginning. Thus, for example, the analysis of the characteristics of a wave can be made in the temporal or frequency domains. The temporal analysis is flexible, fast and simple, but less precise than in the frequency domain. Therefore, holons could in parallel be dedicated in temporal and frequency analysis.

As soon as the sound signal is captured by a microphone, the following main goals have to be performed:

1. To extract the main features or dimensions those identify the speaker. The aim is to construct a partial or fuzzy pattern with a minimum number of features able to identify the speaker as rapidly as possible.
2. To identify the speaker by means of a fuzzy pattern-matching process of the previous partial patterns. If the identification is not possible, it is necessary to extract all possible features of the unknown speaker with the purpose of registering him as a new one.
3. To extract the main features or dimensions of the speech. This process is optimized by the previous identification of the speaker.
4. To recognize what is spoken by means of a fuzzy pattern-matching process of the previous partial patterns. If it is not possible to determine it, then all possible features or dimensions of the unknown speech are extracted to repeat fuzzy pattern-matching with more pattern features.
5. Learning. The positive results are used to obtain the different combinations of fuzzy dimensions of the patterns by means of inverse analysis and genetic algorithms. This way speech recognition and speaker identification is improved. These results produce the modification of the inference rules base optimized to each speaker. We are looking for a most self-learning system by selecting and adapting the learning algorithms.

The main advantage of the holonic system is without a doubt the structure of holarchy. The system can dedicate holons to obtain each dimension of the phonemes. Thus, we are in front of a multi-dimensional or hyper-dimensional phonetic analysis. Later, in a successive manner, when navigating through the holarchy, in superior levels the possible combinations of the results of inferior levels are obtained. Thus the effects of co-articulation are obtained with the diphons and later with the syllables. There is really a process of construction of superior phonetic and lexical units. In the holarchy the different phonological language rules can be expressed, besides expressing the integration of the different knowledge sources of the spoken signal, phonetic, lexical, syntactic, semantic and pragmatic. In addition, in the particular case of phonetic and lexical analysis, fuzzy patterns allow the flexibility of the recognition when rising in the holarchy level. This is possible by exploring the most coherent options, or even by treating the word or lexical unit like a whole in which the humans interpolate phonetic information if the initial and final syllables are correctly pronounced. A degree of finer analysis is also allowed to distinguish between phonetically similar words, constructing directed phonetic graphs in a fuzzy way. These connect the different phonemes in a single direction. Thus the temporary sequential nature of the phonemes is represented, reflecting the uncertainty due to the different alternatives for a pronunciation. Fuzzy values of certainty (the conditional fuzzy probability) can be introduced according to the available dictionaries or corpus.

That distributed parallelism of holarchy allows the definition and resolution of a goal of the system in a holonic way as an integration of subgoals. This way, the most complex processes, such as learning and training, can be made in a parallel and independent way. And this even at the time when recognizing words of different languages and different interlocutors, where the training and learning processes for a given language and interlocutor are optimized. This is also the case when using the IPA language between the different languages. Genetic learning and training algorithms allow "mutations" by means of flexible fuzzy rules of the representations or patterns, which portray co-articulation phenomena between phonetic neighbors - the subtle differences of pronunciation between two related phonemes that share phonetic features.

A preliminary analysis of the system with Prometheus methodology throws the hierarchy of initial goals that must be fulfilled (as shown in Fig. 1). From a first moment the hierarchy of goals could be used as bases of the holarchy, assigning a goal to each holon. When developing each goal in subgoals, when responding to the questions "how?" and "why?", holons are added recursively, forming themselves a holarchy to fulfil the subgoals. The main goals at this initial hierarchy are speaker and speech recognized (*Recognized Speaker/Speech*). In order to arrive to this state they have to fulfill each one of the following goals:

– To digitize the input signal (*Digitize Signal*). This is the pre-processing phase where the signal passes through different filters to heighten certain frequencies ranges, to eliminate noise and to decompose an analogical signal in digital values.
– To detect the main parameters of speaker and speech (*Detect Speech/Speaker Main Parameters*). Here the most decisive parameters for the recognition are calculated.
– To construct the fuzzy or partial pattern for speaker and speech (*Build Speaker/Speech Main Fuzzy Pattern*). The results of the previous goal are fuzzified and the fuzzy feature vector is constructed.

The previous goals form the initial phase of the recognition, the digitalization and the construction of the pattern. The next primary objectives are *Recognize Speaker Pattern)* and *Recognize Speech Pattern.* Both recognition goals must fulfill a sub-hierarchy of goals: *Recognize Word Pattern, Recognize Syllable Pattern, Recognize Diphon Pattern* and *Recognize Phoneme Pattern.*

In this hierarchy a superior goal is made up of several subgoals of the inferior level. There is the composition from an inferior phonetic unit to a superior one - or to a lexical unit, a word with meaning. All these goals, each one at its level and with its patterns, are related to the goal in which the comparison of the fuzzy patterns is made (*Fuzzy Pattern Matching*). For the sake of legibility of the diagram it has been related to both main goals (*Recognized Speaker/Speech*) and to the superior goal of the sub-hierarchy of word recognition (*Recognize Word Pattern*). The comparison of speech and speaker patterns is assumed to be similar. The difference is in the inference rules base that determines the

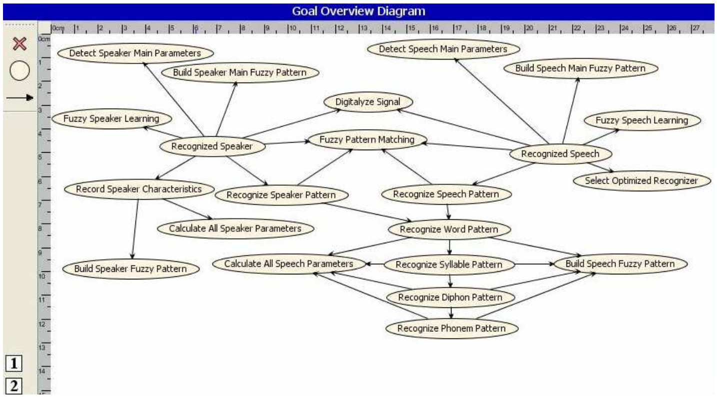


Fig. 1. Initial goals hierarchy of the holonic fuzzy recognition system

values of the features, the combination of the antecedents and the data base of the reference patterns.

In case the speaker has still not been identified, the goal to register the features of the speaker (Record Speaker Characteristics) has to be fulfilled. For it, all that speaker's possible parameters have to be calculated (*Calculate All Speaker Parameters*) and their fuzzy patterns (*Build Speaker Fuzzy Pattern*) must be constructed. In a similar form, when some part of the speech has not been recognized, then all the possible features (*Calculate All Speech Parameters*) have to be calculated and the fuzzy patterns (*Build Speech Fuzzy Pattern*) have to be construct at any phonetic unit level. Finally, the goals of fuzzy learning for speech and speaker (*Fuzzy Speaker/Speech Learning*) optimize the inference rules bases to fulfill the goal to select an optimal recognizer of the speech based on the speaker (*Select Optimized Recognizer*).

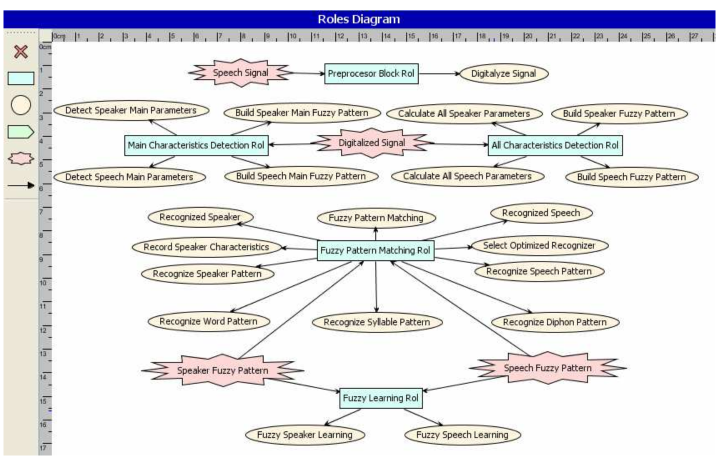


Fig. 2. Roles or functionalities diagram of the holonic fuzzy recognition system

Another interesting diagram is the roles or functionalities diagram (Fig. 2). The roles receive input information or percepts and give rise to actions, fulfilling the system goals. There are four basic percepts: (1) the input signal (*Speech Signal*), (2) the (*Digitized Signal*), (3) the fuzzy pattern of the speaker (*Speaker Fuzzy Pattern*), and, (4) the fuzzy pattern of the speech (*Speech Fuzzy Pattern*). The percepts relate the different basic roles to each other. Role pre-processor (*Preprocessor Block Role*) provides as result the digitized signal that relates the two functionalities in charge of computing the parameters or features (*Main Characteristics Detection Role, All Characteristics Detection Role*) of the fuzzy patterns of speaker and speech. The percepts are used by the role of comparing fuzzy patterns (*Fuzzy Pattern Matching Role*) and fuzzy learning (*Fuzzy Learning Role*).

5 Conclusions and Future Work

The interest to use holonic multi-agent systems comes from the desire to integrate, in an intelligent manner, systems already implemented that provide a very useful function in a given field. The paradigm of the holonic multi-agent systems can extend the use of smaller or specific holonic systems so that they take part of a constellation of stable systems that fulfill a hierarchy of goals. We considered that the paradigm of holonic multi-agent architecture is a robust, independent and scalable way to approach the construction of hierarchies of stable systems. This is the case of speech recognition and speaker identification systems, which must adapt and even learn to identify the speaker and recognize its speech, forming a dynamic hierarchic structure, oriented to the fulfilment of the goals.

The proposed system is part of an upper system of integral surveillance for facilities [9][18][13], able to integrate information from many heterogeneous platforms with multiple sensor types. In this ongoing research project, the microphones are used to extract sound information, which may be noise or human speech. The noise can be analyzed and useful information can be extracted about what occurs in the environment. However, human speech supplies very useful information in a surveillance system.

Acknowledgements

This work is supported in part by the Spanish Ministerio de Educación y Ciencia TIN2004-07661-C02-02 and TIN2007-67586-C02-02 grants, and the Junta de Comunidades de Castilla-La Mancha PBI06-0099 grant.

References

1. Beritelli, F., Casale, S., Cavallaro, A.: A robust voice activity detector for wireless communications using soft computing. IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing for Wireless Communications 16(9), 1818–1829 (1998)

2. Jang, J.S.R.: ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man & Cybernetics 23, 665–685 (1993)
3. Holland, J.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975)
4. Honma, N., Abe, K., Sato, M., Takeda, H.: Adaptive evolution of holon networks by an autonomous decentralized method. Applied Mathematics and Computation 91(1), 43–61 (1998)
5. Huang, B., Gou, H., Liu, W., Xie, M.: A framework for virtual enterprise control with the holonic manufacturing paradigm. Computers in Industry 49(3), 299–310 (2002)
6. International Phonetic Association home page (2005), http://www.arts.gla.ac.uk/IPA/index.html
7. Jarvis, J., Rönnquist, R., McFarlane, D., Jain, L.: A team-based holonic approach to robotic assembly cell control. Journal of Network and Computer Applications 29(2–3), 160–176 (2005)
8. Koestler, A.: The Ghost in the Machine. Arkana Books (1971)
9. López, M.T., Fernández-Caballero, A., Fernández, M.A., Mira, J., Delgado, A.E.: Visual surveillance by dynamic visual attention method. Pattern Recognition 39(11), 2194–2211 (2006)
10. Massaro, D.: Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. The MIT Press, Cambridge (1998)
11. Myers, C.S., Rabiner, L.R.: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal 60(7), 1389–1409 (1981)
12. Padgham, L., Winikoff, M.: Developing Intelligent Agent Systems: A Practical Guide. Wiley, Chichester (2004)
13. Pavón, J., Gómez-Sanz, J., Fernández-Caballero, A., Valencia-Jiménez, J.J.: Development of intelligent multisensor surveillance systems with agents. Robotics and Autonomous Systems 55(12), 892–903 (2007)
14. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs (1993)
15. Russo, M.: FuGeNeSys: A genetic neural system for fuzzy modeling. IEEE Transactions on Fuzzy Systems 6(3), 373–388 (1998)
16. Tsao, C., Gray, R.M.: An endpoint detector for LPC speech using residual error look-ahead for vector quantization applications. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 97–100 (1984)
17. Turing, A.M.: Computing Machinery and Intelligence. Mind 49, 433–460 (1950)
18. Valencia-Jiménez, J.J., Fernández-Caballero, A.: Holonic multi-agent systems to integrate independent multi-sensor platforms in complex surveillance. In: Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance, p. 49 (2006)
19. Zadeh, L.: From computing with numbers, to computing with words: A new paradigm. International Journal on Applied Mathematics 12(3), 307–324 (2002)
20. Zadeh, L.: Fuzzy logic, neural networks and soft computing. Communications of the ACM 37(3), 77–84 (1994)
21. Zadeh, L.: Outline of a new approach to the analysis of complex systems and decisión processes. IEEE Transactions on Systems, Man and Cybernetics, 28–44 (1973)

Social Semantic Cloud of Tag: Semantic Model for Social Tagging

Hak-Lae Kim[1], John G. Breslin[1], Sung-Kwon Yang[2], and Hong-Gee Kim[2]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway, IDA Business Park,Lower Dangan, Galway, Ireland
haklae.kim@deri.org, john.breslin@deri.org
[2] Biomedical Knowledge Engineering Lab, Seoul National University, 28-22 Yeonkun-Dong, Chongno-Ku, Seoul 110-749, Korea
sungkwon.yang, hgkim@snu.ac.kr

Abstract. Tagging has proven to be a successful and efficient way for creating metadata through a human collective intelligence. It can be considered not only an application of individuals for expressing one's interests, but also as a starting point for leveraging social connections through collaborative user participations. A number of users have contributed to tag resources in web sites such as Del.icio.us, Flickr etc.

However, there is no uniform structure to describe tags and user's activities. This makes difficult to share and represent tag data among people. The SCOT (Social Semantic Cloud of Tags) ontology is aimed to represent the structure and semantics of a set of tags and promotes their global sharing. The paper introduce the SCOT ontology and methods of its representation.

1 Introduction

A number of social bookmarking and tagging sites have become popular, and tagging in traditional web sites is getting adopted at a good pace. Many people already know how to bookmark and tag online resources such as Web sites, bookmarks, photos, and blog posts. Tagging is a way for representing concepts by cognitive association techniques, but does not force us to categorize. Each tag tells us about what we are interested in and improves social reinforcement through enabling social connections and search. There is an advantage that social bookmarking and tagging is a simple way that allows a user to save and share anything in online communities.

But the critical problem is that the social bookmarking and tagging systems do not provide a uniform way to share and reuse tag data among users or communities. Although some systems support an export functionality using open APIs, there are no uniform structure and semantics to represent tag data. Therefore, it is not easy to meaningfully search, compare or merge "*similar collective tagging data*" [2] on different sources.

This paper will explore the idea of defining the structure and the semantics of social tagging through a new approach we have called SCOT (Social Semantic

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 83–92, 2008.

Cloud of Tags)[3]. This approach can help to represent the context required in the use of social tagging and to provide methods for sharing and reusing tagged data.

To remedy the shortcoming of social tagging, we advocate using Semantic Web technologies semantically to represent tagging data. In particular, the contributions of the paper are as follows:

- We suggest the semantic model to represent the structure and semantics of tagging data on social tagging spaces.
- The SCOT can enhance sharing and reuse of social tagging data.

The next section will discuss the motivation for presented ideas and address a number of issues for social tagging. In Section 3 the design principles of the SCOT ontology will be introduced. Section 4 will be devoted to the core concepts of the SCOT ontology to describe the structure and the semantics of tag data. Section 5 explains how a SCOT space expands. Section 6 describes some works related to the topic. Finally, future directions for further development of the SCOT ontology will be discussed.

2 Motivations

We propose the SCOT ontology to address a number of issues of note in the domain of tagged social content.

Firstly, people have a certain amount of inherent laziness when it comes to tagging content. Over time, they may build up a rich collection of tags on a certain site or application, but the size and scope of this collection will depend on how regularly the site or application is being used. When beginning to tag content on a new system, the user may have a certain amount of reluctance to create new tags if they do not expect to be using the system that much, and also they may not wish to recreate their existing tag information on other sites.

Related to this issue is a problem with the lack of reuse of tags between an individual's various applications. At the moment, there is no consistent method for reusing one's personal set of tags between either web-based systems, desktop applications, or for transferring tags between the desktop and the Web. For example, in Windows Vista users can now tag files and use those keywords to organise and search all documents on their desktop. However, they cannot reuse this collection of tags from their desktop when they signup and start annotating content on a new social bookmarking website.

The third issue is in relation to a consistent use of tagging. Even a single user may forget the tags that they have used previously on a single site (not to mention all the sites or applications that they may use) when they are prompted with an empty tag field for annotating a new content item. Some systems may provide tag recommendations based on previously used tags (or from a user's social network), but this could be augmented by suggesting tags from any system that a user has previously tagged content on. In order to provide a consistent and wider view of all the content that a person has tagged with a particular keyword through multiple applications, such cross-platform exchanges of tag cloud metadata is necessary.

Another motivation is to allow tag reuse between people in existing social networks, and to connect people who may have a common interest of set of interests. On signing up for a site, one may quickly realise that a particular tagging scheme used by a network contact is of interest when tagging one's own content. Also, social networks of people may be tagging content across various sites, and therefore a tag cloud for not only one's own content but also for one's friends' can be useful for maintaining a cohesive social network across different systems. Even if there is no existing link between people with similar interests, these can be formed through a serendipitous use of similar tags or by browsing from one's community tag cloud to others who use tags in that cloud.

Finally, another problem that has been noticed with the exporting of tag information from social applications is the lack of explicit information on tag structure, how often tags occur together, and who has used a particular set of tags. We aim to address this lack of semantics of tag structure with a number of interesting properties in the SCOT ontology.

3 Design Principles of the SCOT

In this section, the principles of the development of the SCOT ontology are described. The SCOT ontology[1] is an ontology for sharing and reusing tag data and representing social relations among individuals. It provides the structure and semantics for describing resources, tags, users and the extended tag information such as tag frequency, tag co-occurrence frequency, and tag equivalence. As the SCOT ontology is established to support semantically social tagging, the framework is designed to represent tag data at both personal and group-level. There is an opportunity to build group-level SCOT through integrating multiple personal SCOT ontologies. This idea has emerged on the following major principles:

- ***lightweight.*** Our approach follows the principle "a little semantic goes a long way[6]." The ontology model must be designed both with minimal structure and minimal semantics. Accordingly, the SCOT is expressed in a simple RDF format.
- ***share and reuse.*** In order to share and reuse the data to other applications, the ontology model must provide a consistent method for sharing existing sets of tags among users.
- ***compatibility.*** Except for the core concepts of our ontology that represent tagging activities, there is no intention to define redundant classes and properties that already exist in other RDF vocabularies.

4 Overview of the SCOT

The SCOT ontology generically models tagging activities for typical online communities and relations between each component of the activity. Figure 1 shows simplified model of SCOT ontology with its main top-level concepts and relations.

[1] http://scot-project.org/scot/ns/

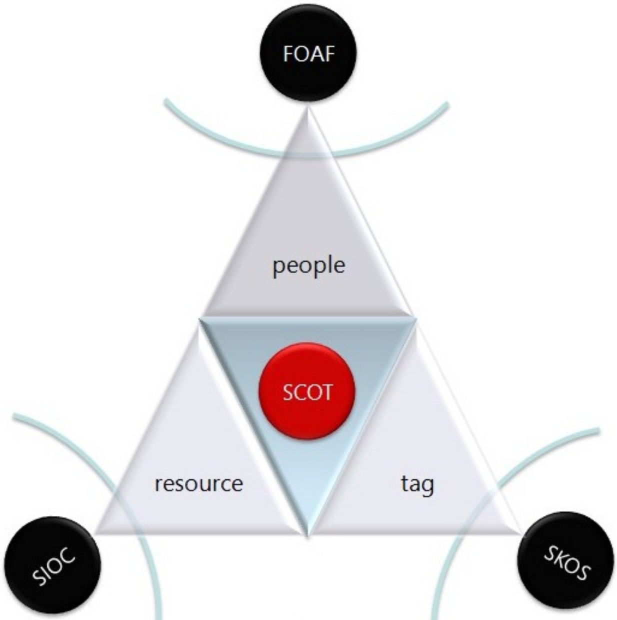


Fig. 1. SCOT Ontology Model

The SCOT ontology is linked to the three dimensional relationships that are represented in SIOC[7,10], FOAF[9], and SKOS[8]. We use the SIOC concepts to describe site information and relationships among container-item as well as site-site, and use the FOAF concepts to represent either a human or machine agent because a tag can be generated manually by a human user or automatically by a machine. Also we try to represent the relationships among users. This relationship has the two aspects: agent-agent and agent-group. When we are tagging and are using them, we assume these relationships. Finally, we use the SKOS to allow semantically relate a tag with another tag using properties such as `skos:broader` and `skos:narrower`. The SCOT consists of the Tagcloud, the Tag and the Cooccurrence class as the core concepts. In the following section, we describe in detail.

4.1 Tagcloud Class

The `Tagcloud` class is the central element of the ontology. Basically the term *tagcloud* is used for navigation and visualization of content tags used on a website. Often, more frequently used tags are displayed in a larger font or otherwise emphasized, while the displayed order is generally by alphabet or by popularity. In the sense, "*text cloud*" can be considered a more accurate term than tagcloud from the visualization point of view.

Tagging itself, however, includes users, tags, and the relation among them. The tagcloud, the result of tagging, represents *user-tag* relation and *tag-tag* relations for particpants of tagging activities. The term tagcloud in our model has comprehensive perspectives to represent the entire tagging activity rather than just a simple visualization.

The `scot:Tagcloud` is a class that can be used to represent groups of users and a set of tags for users in certain sites or forums. For instance, there are the properties such as `scot:hasUsergroup`, `scot:hasMember`, `scot:composedOf`, `scot:totalTagFrequency`, `scot:totalCooccurFrequency`, `scot:totalItems`, `scot:totalTags`, `dcterms:created`, and `dcterms:modified` and so on.

The range of `scot:hasUsergroup` property is an instance of any of the classes such as the `Usergroup, User, Site,` and `Forum` from SIOC and `Agent, Group` from FOAF. Thus, a user group of the `scot:Tagcloud` class can be considered a common interest group according to set of tags. The scot:composedOf property is to describe that if a Tagcloud has more than two Tagcloud (i.e. group SCOT), each Tagcloud is part of the Tagcloud.

4.2 Tag Class

The `Tag` class is used to represent the concept of a tag and its statistical and linguistical properties. This class, a member of Tagcloud class, has a name through URIs. There are some well-known flaws in using free-tagging classification such as "*tags variation*", "*tags ambiguity*", and "*flat organization of the tags*"[11]. Those flaws are critical barriers to more precise categorization and to methods for better navigation.

statistical properties. A single tag has a frequency which is the number of occurrences. In SCOT, a tag's frequency is represented by the `scot:Frequency` property. This property has a value from the XML schema datatype `xsd:integer`.

A certain tag would often appear tags, and the meaning of the tag become more specific when the tag is combined with a set of tags. Generally it is so called '*tag co-occurrence*' with those other tags. It can play an important role to reduce '*tag ambiguity*'. In SCOT, a tag co-occurrence is represented by the `scot:cooccurTag` property. When we use this property, we must consider the frequency of the individual tag itself and also the frequency among co-occurring tags. Thus we define two properties for this purposes. The First is `scot:ownFrequency` property that is used for representing a frequency of a single tag. The next is `scot:cooccurFrequency` to which describes co-occurrence frequency amongst a set of tags.

Both properties are subproperty of `scot:Frequency` property which is an instance of `rdf:Property`. These three properties (`scot:Frequency, scot:ownFrequency, scot:cooccur Frequency`) are called "*statistical properties*" of tag.

When we want to describe cooccurrence and its frequency among tags, we need to represent it by n-ary relations. For instance, let us assume that the tag *blog* has a cooccurrence with *web2.0* with a frequency of 5 (see Figure 2). There is a binary relation between the Tag *blog* and the tag *web2.0* and there is a coocccurrence value describing this relation. We try to represent this information using an

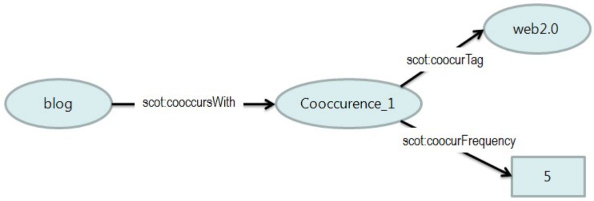


Fig. 2. Cooccurrence and its frequency between blog and web2.0

instance of the class `scot:Cooccurrence`. The individual Cooccurrence_1 here represents a single object encapsulating both the tag (web2.0, a specific instance of `scot:Tag`) and the cooccurrence value among the individuals (5).

linguistical properties. These properties are used to enhance the semantics of individual tags and to create connections between different variations of tags. In SCOT all tags have a concept and can be represented by a hierarchy among tags in SKOS. The `skos:broader` property is used to describe a more general term inverse of `skos:narrower`. In fact, it is not easy to build a tag hierarchy from a set of individual tag. However, we can build tag frequencies using the frequency of the tag and co-occurrence among tags, and can then represent their concept hierarchy based on this information. It can be provided a different structure to visualize a tagcloud beyond the flat organization of the tags.

A tag would often appear in different conventions across a systems or web sites. Firstly, let us think of the term '*Weblog*'. We can find some different conventions such as '*weblogs*' as the plural, 'weblog' in the lowercase, or '*blog*' as the short term. All terms can be considered as having the same meaning, if not the intended purpose. In addition, we can consider different examples like '*Semantic Web*'. In this case, the issue is placed on how to combine a compound word rather a number or case-sensitive version of the term. We can intuitively think '*semantic_web*', '*SemanticWeb*', or '*Semantic-Web*' as having same meaning with '*Semantic Web*'.

The above examples are typical real-world cases encounter when we use a tag. Even if some popular Web 2.0 sites provide methods to handle different conventions, they do not allow one to represent this information explicitly. For solving these flaws, we define several properties such as `scot:spellingVariant`, `scot:acronym`, `scot:plural`, `scot:singular`,`scot:synonym`, `scot:delimited`, `scot:hypenated`, `scot:underscored`, `scot:slashed`, and `scot:spaced` and so on. It can reduce tag ambiguity from acronyms and even recommend more common patterns of tag name.

5 How SCOT Space Expands

A huge amount of tagging data has been created from different users, applications, and domains. But it is difficult to share and reuse the tagging data across the different environments. SCOT provides a comprehensive way for sharing and reusing the tagging data. Basically SCOT can be created by the efforts of a single user or a machine agent. In this case, the user owns the tagcloud. We are called it as the *Personal SCOT*. However, if a tagcloud is made by a group of users or reflects tagging efforts from different sites, its ownership is not an individual, but rather includes all those who participate in this activity. Such group of users would share the tagcloud under a boundary. We call it as the *Group SCOT*.

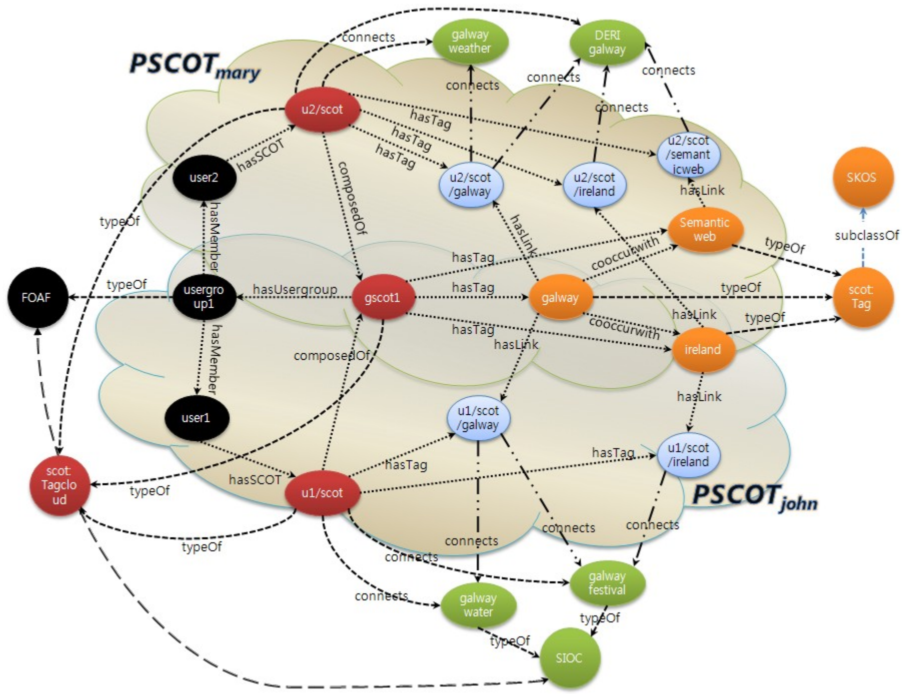


Fig. 3. Relationship between RDF vocabularies and SCOT: $gscot_1$ is composed of both $PSCOT_{mary}$ and $PSCOT_{john}$. FOAF has information about Mary and John and also has URIs for their SCOTs. Each SCOT can be connected with sioc:Post as the URIs.

A user can create a personal SCOT from his/her data source and then publish and share it. The personal SCOT can be used for searching in the user's site or blog, but also can be contained in other's SCOT. This means that SCOT can be composed of combinations of either multiple personal SCOTs or multiple group SCOTs. Moreover, SCOT can be created across different types of applications such as weblogs, social bookmarking sites etc.

We make it possible to exchange the tagging data and to navigate resources using SCOT across various and varied applications. It make it possible to build the SCOT spaces which is the network based on user-driven tagging data. In addition, It can be considered as a kind of folksonomy for those in the group. Figure 3 illustrates the scenario to interconnect across users or communities and shows how the SCOT is connected with other RDF vocabularies.

6 Related Work

In [1], the author describes purposes of "semantics for tagging" with two aspects: to help in people's understanding and to allow computers to process the tag for supporting people's understanding.

The semantics of a tag is primarily about the meaning among people or a community group in the social space. There are several efforts that try to represent the concept of tagging, the operation of tagging, and the tag themselves.

Newman[14] describes the relationship between an agent, an arbitrary resource, and one or more tags. In his ontology, there are three core concepts such as Taggers, Tagging, and Tags to represent tagging activity. Taggers are represented by `foaf:Agents` and Taggings "reify the n-ary relationship between a tagger, a tag, a resource, and a date." Tags are represented by the Tag class with URIs and are linked to `skos:Concept` and `skos:subject`.

Gruber[15] describes the core idea of tagging that consists of object, tag, tagger, and source. One notable thing is that he defines the source as the "scope of namespaces or universe of quantification" for objects. This allows one to differentiate between tagging data from different systems and is the basis for "collaborative tagging across multiple applications."

Knerr[17] describes the concept of tagging in the *Tagging Ontology*. Since his approach is based on the ideas from [14] and [15], the core element of the ontology is Tagging. The ontology consists of time, user, domain, visibility, tag, resource, and type.

The approaches in the related work are focused on tagging activities or events that people used to tag in resources using terms. Therefore the core concept of the ontologies is Tagging, and there are Tagger and Resource class to represent user and resource respectively. However, there are no way to describe frequency of tags in the ontologies. The SCOT ontology is easy to represent this information using three properties of frequency. In addition, we provide a number of properties to represent social tagging activity and relationships among elements occurring on online community.

7 Conclusion

We have introduced the SCOT ontology, which focuses primarily on representing the uniform structure and the semantics for tag data. And we also have proposed how to share tag data among different sources or different people.

A tag can be used with many different meanings depending on user's contexts. Thus, when tags are represented by kinds of ontology, we should take into account relationships among a tagger, a resource, and a tag. The SCOT ontology can describe three concepts and the relationships to define context for a tagging activity. Furthermore, the SCOT has some properties for handling tag frequencies and tagging activities such as total posts, number of users etc. In other words, the ontology has numerical and linguistical properties. The classes and the properties of SCOT is essential to represent tag data and share them.

Our approach is a starting point to share tag data beyond sites or people. To realize SCOT space, we have provided the SCOT Exporter that exposes a SCOT instance from legacy databases and the int.ere.st web site to search, bookmark, integrate, and share SCOT instances. We plan to extend and improve the Exporter and the int.ere.st for applying to various online communities. We provide more detailed information about the SCOT ontology at http://scot-project.org.

Acknowledgments

This study was partly supported by a grant of the interoperable EHR Research and Development (A050909), Ministry of Health & Welfare, Republic of Korea and the Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

References

1. Beckett, D.: Semantics Through the Tag. In: Proceeding of XTech 2006: Building Web 2.0 (2006), Available at: http://xtech06.usefulinc.com/schedule/paper/135
2. TagCommons project website, http://tagcommons.org
3. SCOT project website, http://scot-project.org
4. Kim, H.L., Yang, S.K., Breslin, G.J., Kim, H.G.: Simple Algorithms for the SCOT Exporters using Tag Frequency and Cooccurrence. In: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2007), Silicon Valley, USA (2007)
5. Kim, H.L., Yang, S.K., Song, S.J., Breslin, G.J., Kim, H.G.: Tag Mediated Society with SCOT Ontology. In: Proceedings of the Semantic Web Challenge 2007 in conjunction with the Sixth International Semantic Web Conference, Busan, Korea, November 11-15 (2007)
6. Hendler, J.: The Dark Side of the Semantic Web. IEEE Intelligent Systems and Their Applications 22(1), 2–4 (2007)
7. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards Semantically-Interlinked Online Communities. The UPGRADE European Journal for the Informatics Professional 6(6) (2005)
8. Brickley, D., Miles, A.: SKOS Core Guide-W3C Working Draft 2 (2005), Available at: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
9. Brickley, D., Miller, L.: FOAF Vocabulary Specification (2005), Available at: http://xmlns.com/foaf/0.1
10. Brickley, D., Stefan, S., Miles, A., Miller, L., Caoimh, D.O., Neville, C.M.: SIOC Ontology Specification (2007), Available at: http://rdfs.org/sioc/spec/
11. Passant, A.: Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs: Theoretical background and corporate use-case. In: Proceeding of International Conference on Weblogs and Social Media, Boulder, Colorado, U.S.A. (2007)
12. Finin, T., Ding, L., Zhou, L., Joshi, A.: Social networking on the semantic web. The Learning Organization: An International Journal 12(18), 418–435, May 1 (2005)
13. Lux, M., Granitzer, M., Kern, R.: Aspects of Broad Folksonomies. In: 4th International Workshop on Text-Based Information Retrieval (2007)

14. Newman, R.: Tag Ontology design, (viewed 03/04/2007) (2005), Available at: http://www.holygoat.co.uk/projects/tags/
15. Gruber, T.: Ontology of Folksonomy: A Mash-up of Apples and Oranges, First on-Line conference on Metadata and Semantics Research (MTSR 2005), http://www.metadata-semantics.org/, (to appear in International Journal on Semantic Web and Information Systems vol 3(1) (2007)
16. Gruber, T.: Collective Knowledge Systems: Where the Social Web meets the Semantic Web. To appear in Journal of Web Semantics (2007), Avaiable at: http://tomgruber.org/writing/Collective-KnowledgeSystems.htm
17. Knerr, T.: Tagging Ontology- Towards a Common Ontology for Folksonomies (2006), Available at: http://code.google.com/p/tagont/

Realization of an Intelligent Frog Call Identification Agent

Chenn-Jung Huang[1,*], Yi-Ju Yang[2], Dian-Xiu Yang[1], You-Jia Chen[1], and Hsiang-Yu Wei[2]

[1] Department of Computer & Information Science
National Hualien University of Education, Hualien, Taiwan
cjhuang@mail.nhlue.edu.tw
[2] Institute of Ecology and Environmental Education
National Hualien University of Education, Hualien, Taiwan

Abstract. An intelligent frog call identification agent is developed in this work to provide the public to easily consult online. The raw frog call samples are first filtered by noise removal, high frequency compensation and discrete wavelet transform techniques in order. An adaptive end-point detection segmentation algorithm is proposed to effectively separate the individual syllables from the noise. Four features are extracted and serve as the input parameters of the classifier. Three well-known classifiers, the *k*-th nearest neighboring, Support Vector Machines and Gaussian Mixture Model, are employed in this work for comparison. A series of experiments were conducted to measure the outcome performance of the proposed agent. Experimental results exhibit that the recognition rate for Gaussian Mixture Model algorithm can achieve up to the best performance. The effectiveness of the proposed frog call identification agent is thus verified.

Keywords: intelligent agent, gaussian mixture model, support vector machines, k-th nearest neighboring, Mahalanobis distance.

1 Introduction

Pattern recognition forms a fundamental solution to different problems in real world applications [1]. The function of pattern recognition is to categorize an unknown pattern into a distinct class based on a suitable similarity measure. Thus similar patterns are assigned to the same classes while dissimilar patterns are classified into different classes.

In speech recognition, a source model is assumed and the signal is expected to obey the laws of a specific spoken language with a vocabulary and a grammar. Frog vocalization is a representative instance of a category of natural sounds where a vocabulary and other structural elements are expected. In comparison with the human speech recognition problem, animal sounds are usually simpler to recognize. Speech recognition often proceeds in a quiet and similar environment, while frog's sounds are

* Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 93–102, 2008.

usually recorded in a much noisy environment, under which we must recognize simpler vocalizations.

In general, features include time domain and frequency domain features. Time domain features are calculated directly from the sound waveform such as zero crossing rate and signal energy. Time domain features signal is first transformed to the frequency domain using Fourier transform and new features are thereby derived from transformed frequency signals.

An agent is an intelligent and autonomic assistant which is able to take the place of humans. We thus propose a frog call identification agent in this work to automatically recognize the frog species based on the recorded audio signals that were sampled from recordings of frog sounds in an outdoor environment. The sampled signals were first converted into frequency signals. Then syllable segmentation and feature extraction methods are employed to separate the original frog calls into syllables and to derive the input features for the classifiers. Experimental results and analysis are given to verify the effectiveness of the proposed agent.

The remainder of this article is organized as follows. The architecture of the intelligent frog call identification agent is presented in Section 2. Experimental results and analysis is given in Section 3. Finally, Section 4 concludes the work.

2 Architecture of the Intelligent Frog Call Identification Agent

The architecture of the proposed frog call identification agent can be divided into four main modules, including signal preprocessing, syllable segmentation, feature extraction, and classification modules, as illustrated in Fig. 1. Undesirable information is first removed from the raw input signals in order to preserve the desired characteristics of frog call during signal preprocessing stage. The resulting signal is then segmented by the syllable segmentation method and the segmented syllables are further processed during feature extraction stage to produce meaningful parameters for the classifier.

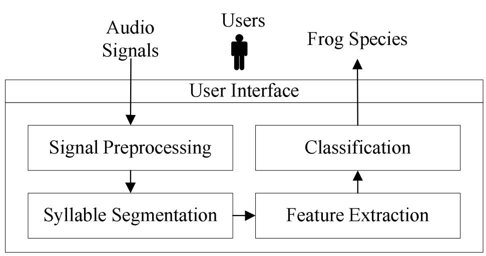


Fig. 1. Architecture of the intelligent frog call identification agent

2.1 Signal Preprocessing

The recorded sound signal is resampled at 16k Hz frequency and saved as 16-bit mono format. The amplitude of each sound signal is normalized within the range [−1, 1] for the ease of further processing. Three techniques, including pre-emphasis, de-noise and discrete wavelet transform (DWT) are applied in order to "purify" the data in the noisy environment.

The motivation of using pre-emphasis technique is to compensate the high-frequency part that was suppressed during the recording of audio signal by using sound production mechanism. De-noise filter is employed to remove the noise during the signal analysis; while the application of DWT is to resolve high frequency and low frequency components within a small and large time windows, respectively.

2.1.1 Pre-emphasis Filter

Pre-emphasis filter is used to amplify the importance of high-frequency formants in this work. Base on the input sequence, the signal derived by pre-emphasis technique can be expressed by,

$$S'_n = S_n - \alpha \times S_{n-1}, \tag{1}$$

where S_n and $S_{n\text{-}1}$ S_{n-1} denote the raw signal sequence, and α is obtained from $H(z) = 1 - \alpha \times z^{-1}$. Here the *z*-transform $H(z)$ represents a high-pass filter.

2.1.2 De-noise Filter

De-noise filter is a well-known technique to remove the noise during the signal analysis. De-noise filter kernel function makes use of wavelet threshold function in 1-dimention signal. The output signal of De-noise filter is expressed by,

$$D'_n = f(g(h(D_n, c_1, c_2))), \tag{2}$$

where D_n denotes the raw signal, c_1 and c_2 represent the coarsest level and quadrature mirror filter, respectively, and h, g and f stand for the forward wavelet transform, soft threshold function, and the inverse wavelet transform, respectively.

2.1.3 Discrete Wavelet Transform

Discrete wavelet transform (DWT) has been applied to various speaker identification problems. Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks, and for accurately deconstructing and reconstructing finite, non-periodic and non-stationary signals. The DWT performs the recursive decomposition of the lower frequency band obtained by the previous decomposition in dyadic fashion.

2.2 Adaptive End-Point Detection Segmentation

A syllable is basically a sound that a frog produces with a single blow of air from the lungs. The rate of events in frog vocalization is so high that the separation of individual syllables is difficult in a natural environment due to reverberation. Once the syllables have been properly segmented, a set of features can be collected to represent each syllable. In this work, we propose an adaptive end-point detection segmentation method, in which a cleaner waveform can be obtained by accurately measuring the positions of two endpoints of the waveforms. It is observed that the peak amplitude of each individual syllable is located around the middle portion of each separate frog call. We thus extract the major portion of each syllable by trimming two sides of the peak to filter out the noise contained in the audio signal. Meanwhile, the parts of the filtered signal whose volumes fall below a predefined threshold are discarded and the longest continuous segment is extracted for the ease of the processing at the feature extraction stage.

The first part of the algorithm that trims two sides of the peak amplitude is summarized as follows:

I. Compute the amplitude of the input acoustic signal using iterative time-domain algorithm. We denote the amplitude a matrix $S(a,t)$, where a represents amplitude value and t is the time sequence. Initially set $n = 1$.

II. Find a_n and t_n, such that $S(a_n, t_n) = \max\{S(|a|, t)\}$. Set the position of the nth syllable to be $S(a_n, t_n)$.

III. If $|a_n| \le a_{threshold}$, stop the segmentation process. The $a_{threshold}$ is the empirical threshold. This means that the amplitude of the nth syllable is too small and hence no more syllables need to be extracted.

IV. Store the amplitude trajectories corresponding to the nth syllable in function $A_n(\tau)$, where $\tau = t_n - \varepsilon, \ldots, t_n, \ldots, t_n + \varepsilon$ and ε is the empirical threshold of the syllable. The step is to determine the starting time $t_n - \varepsilon$ and the ending time $t_n + \varepsilon$ of the nth syllable around t_n.

V. Set $S(a, (t_n - \varepsilon, \ldots, t_n + \varepsilon)) = 0$ to delete the area of nth syllable. Set $n = n + 1$ and go to (II) to find the next syllable.

The second part of the algorithm that extracts the longest continuous segment is summarized as follows:

I. The volume level of j-th frame for the i-th syllable can be expressed by,

$$v_{i,j} = 10 \log_{10}\left(\sum_{t=1}^{n} a_{i,t}^2\right), \tag{3}$$

where n is the number of empirical frame size, $a_{i,t}$ represents the amplitude value at time t.

II. Initially set $i = 1$.

III. A volume sequence S_i is obtained via (I) and (II). The volume sequence of i-th syllable can be expressed by,

$$S_i = \bigcup_{k=1}^{m} v_{i,k}, \tag{4}$$

where $v_{i,k}$ is the volume level of k-th frame for the i-th syllable, and $\{k\}$ is a continuous integer sequence. Initially set $k = 1$.

IV. Find a subset s in S_i, such that each frame of the subset s is greater than $v_{threshold}$ and s is the longest continuous segment. The value of $v_{threshold}$ is determined by the experiments.

V. The first and the last elements in the subset s are regarded as the start and end points of the filtered syllable.

VI. Set $i = i + 1$. If there are unprocessed syllables, go to (III).

2.3 Feature Extraction

There are three well-known features [1] used in the pattern recognition work in the literature, including spectral centroid, signal bandwidth and spectral roll-off, employed in this work. Besides them, a new developed feature, threshold-crossing rate, is proposed in this work to reduce the impact of the noises in the sound samples.

2.3.1 Spectral Centroid

Spectral centroid is center point of spectrum and in terms of human perception it is often associated with the brightness of the sound. Brighter sound is related to the higher centroid. Spectral centroid for signal syllable is calculated as:

$$S = \frac{\sum_{n=0}^{M} n|x_n|^2}{\sum_{n=0}^{M} |x_n|^2}, \tag{5}$$

where x_n is discrete Fourier transform (DFT) of signal syllable for the n-th sample and M is half of the size of DFT. The DFT of each frame is expressed by,

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}nk}, \tag{6}$$

where $k = 0,\dots,N-1$.

2.3.2 Signal Bandwidth

Signal bandwidth is defined as the width of the frequency band of signal syllable around the center point of spectrum. The bandwidth is calculated as,

$$B = \sqrt{\frac{\sum_{n=0}^{M} (n-S)^2 |x_n|}{\sum_{n=0}^{M} |x_n|^2}}. \tag{7}$$

Notably, the bandwidth of syllable is calculated as average of bandwidth of DFT frames of syllables.

2.3.3 Spectral Roll-Off

This feature measures frame-to-frame spectral difference. In short, it tells the changes in the spectral shape. It is defined as the squared difference between the normalized magnitudes of successive spectral distribution. Spectral roll-off frequency can be expressed by,

$$S = \max\left(M \cdot \sum_{n=1}^{M} \left|x_n\right|^2 \le C \cdot \sum_{n=1}^{M} \left|x_n\right|^2 \right), \tag{8}$$

where C is an empirical constant ranged between zero and one, x_n is the DFT of signal syllable for the n-th sample and M is half of the size of the DFT. The DFT of each frame is expressed from (6).

2.3.4 Threshold-Crossing Rate

Traditionally zero-crossing rate is the number of time domain zero-crossings in each individual syllable. A zero-crossing occurs when adjacent signals have different signs. Zero-crossing rate is closely related to spectral centroid as they both can be used to derive the spectral shape of syllable. The so-called threshold-crossing rate is adopted in this work to ignore the time domain zero-crossings in each individual syllable produced by the noises. The threshold-crossing rate is defined as:

$$T = \frac{1}{2}\sum_{n=0}^{M-1} \left| \mathrm{tsgn}(x_n) - \mathrm{tsgn}(x_{n+1}) \right|, \tag{9}$$

where the function tsgn(·), which represents the threshold, is defined as,

$$\mathrm{tsgn}(x_n) = \begin{cases} 1, & x_n > \eta \\ -1, & x_n < -\eta \end{cases}, \tag{10}$$

Here η is an empirical amplitude threshold set for the syllable.

2.4 Classification

Three well-known machine learning techniques, *k*-nearest neighbor classifier (*k*NN), support vector machines (SVM) and Gaussian Mixture Model (GMM), are used to classify the frog species in this work. The spectral centroid, signal bandwidth, threshold-crossing rate and spectral roll-off are input parameters to the classifier..

*k*NN and GMM have been widely applied to various music sound analysis and music information retrieval problems in the literature. Although SVM is rarely used in the application domain of audio analysis problems, we attempt to apply this advanced machine learning technique because it has shown excellent performance in binary classification problems. A trustworthy open source implementation for SVM, LibSVM [4], is adopted in this work. It supports multi-class classification based on one-against-all approach, in which the SVM is trained using all the instances in a specific class as positives and the rest of instances as negatives.

2.4.1 *k*-Nearest Neighbor Classifier

k-nearest neighbor classifier (*k*NN) simply stores the presentative training data. This algorithm assumes all the examples correspond to vectors in *n* dimensional vector space. The neighbors of an example are defined in terms of one of distance functions, such as Euclidean distance and Mahalanobis [2] distance.

In the *k*NN, the *k*-nearest samples in the training data set are found, for which the majority class is determined. The distance measure for features is of critical importance for *k*NN classifier. Among different distance measures in the literature, Euclidian distance function is the most widely used in *k*NN. It is defined as

$$d_E(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}, \tag{11}$$

where $\mathbf{a} = (a_1, a_2, ..., a_n)$ and $\mathbf{b} = (b_1, b_2, ..., b_n)$ are two points in the *n*-dimensional space.

The well known Euclidian distance shows good performance for clean and regular features, but loses performance in case there is some correlation between some of the features. We thus adopt Mahalanobis distance measure for *k*NN as well because Mahalanobis distance is useful in proximity measures when the attributes are relatively correlated. The Mahalanobis distance is defined as,

$$d_M(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{i=1}^{n} \frac{(a_i - b_i)^2}{\sigma_i^2}}, \tag{12}$$

where σ_i is the standard deviation of the a_i over the sample set.

2.4.2 Support Vector Machines Classifier

There has been a surge of interest in Support Vector Machines (SVM) in recent years. SVM is a new generation learning system based on recent advances in statistical learning theory and was first introduced in the early 1990s by Vapnik [3] as a binary classification tool. It is rapidly growing in popularity due to many attractive features, and promising empirical performance.

Support vector machine [3][4] is a discriminative model classification technique that mainly relies on two assumptions. First, transforming data into a high-dimensional space may convert complex classification problems with complex decision surfaces into simpler problems so that linear discriminant functions can be used. Second, SVMs use those training patterns that are near the decision surface.

2.4.3 Gaussian Mixture Model Classifier

Gaussian Mixture Model (GMM) [5] classifier is quite popular in speech and speaker recognition. The probability density functions that most likely resemble the training patterns of each of the classes are generated in GMM, assuming that they can be modeled by a mixture of a number of multidimensional Gaussian distributions. The iterative expectation-minimization algorithm is then used to estimate the parameters of each Gaussian component and the mixture weights.

3 Experimental Results

In this work, a database that consists of five frog species as listed in Table 1 was used in the experiments to verify the effectiveness of the proposed frog call identification agent. Five frog species are evenly distributed in 25 wave files, and each file contains only one species. The resampling frequency is 16 kHz and each sample is digitized in 16-bit mono. Each acoustic signal is first segmented into a set of syllables. The value of the threshold level used in the adaptive end-point detection segmentation method is set to 35% of the maximum volume of the input sequence. The frame size and overlapping size are 512 and 256 samples, respectively. The total number of samples is 727 syllables, spanned from 25 wave files. The number of the training syllables is 469 and that of holdout syllables is 258. That is, about one-third of the syllables are used for holdout and the rest are used for training.

The following classification accuracy rate is used to examine the performance of the proposed work,

$$A = \frac{N_C}{N_T} \times 100\%, \tag{13}$$

where N_C is the number of syllables which were recognized correctly and N_T is the total number of testing syllables.

Table 1. Frog species for Family Microhylidae

Family	Scientific Name	Common Name
Microhylidae	Microhyla heymonsi	Heymonsi Narrow-Mouthed Toad
	Kaloula pulchra	Malaysian Narrow-Mouthed Toad
	Microhyla butleri	Butler's Rice Frog
	Microhyla ornate	Ornate Rice Frog
	Microhyla steinegeri	Stejneger's Paddy Frog

Table 2. Accuracy rates of *k*NN-E, *k*NN-M, SVN and GMM classifiers

Classifier \ Frog Species	Butleri	Heymonsi	Ornate	Pulchra	Steinegeri
kNN-E	94.64%	93.49%	88.89%	100.00%	99.23%
kNN-M	94.25%	95.02%	89.66%	99.62%	99.62%
SVM	96.17%	93.49%	89.66%	100.00%	100.00%
GMM	95.02%	95.02%	91.19%	99.23%	99.62%

The accuracy rate comparison for the three calssifiers, including *k*NN, SVM and GMM, are listed in Table 2. Notably, the notations *k*NN-E and *k*NN-M stand for the *k*NNs that employs Euclidean distance measure and Mahalanobis distance measure, respectively. The number of neighbors, *k*, set for *k*NN-E and *k*NN-M is three. In SVM and GMM classifier, the adopted kernel function is linear and the selected haar wavelet is a detail coefficients vector. Generally speaking, GMM slightly outperforms the other three classifiers, *k*NN-E, *k*NN-M and SVM.

Tables 3 to 5 give the comparisons of the three well-known performance measures, including precision rate, recall rate and F-measure, in the literature. The three measures have been widely used in the multi-label multi-category classification. The expressions of the three performance measures are given by,

$$precision = \frac{\text{Number of frog retrieved calls that are relevant}}{\text{Total number of retrieved frog calls}} \times 100\% \tag{14}$$

$$recall = \frac{\text{Number of retrieved frog calls that are relevant}}{\text{Total number of frog calls that are relevant}} \times 100\% \tag{15}$$

$$F-measure = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \tag{16}$$

Table 3 shows the precision rate comparison for the five frog species. The precision rates for the five frog species range from 65.5% to 100%. The identification of Ornate by using SVM gives the worst performance, while the recognition of Pulchra and Steinegeri by using *k*NN-E, *k*NN-M and SVM, and the classification of the Ornate by using *k*NN-E all achieve 100% precision rate.

Table 3. Precision rate comparison for five frog species

Classifier \ Frog Species	Butleri	Heymonsi	Ornate	Pulchra	Steinegeri
kNN-E	84.6%	83.5%	100.0%	100.0%	100.0%
kNN-M	84.6%	92.9%	72.7%	100.0%	100.0%
SVM	88.5%	91.5%	65.5%	100.0%	100.0%
GMM	86.3%	97.4%	70.9%	81.8%	98.1%

Table 4. Recall rate comparison for five frog species

Classifier \ Frog Species	Butleri	Heymonsi	Ornate	Pulchra	Steinegeri
kNN-E	100.0%	100.0%	19.4%	100.0%	96.23%
kNN-M	100.0%	92.9%	44.4%	100.0%	100.0%
SVM	100.0%	88.3%	52.7%	100.0%	100.0%
GMM	98.7%	87.2%	61.1%	100.0%	100.0%

Table 4 shows the recall rate comparisons for five frog species by using the four classifiers. The recall rate for Ornate is apparently worse than the other four species. Parts of the frog calls for Ornate were misclassified as those for Butleri and Heymonsi. This leaves much room for improvement of the segmentation techniques to better the recognition rate for Ornate species.

The F-measure comparisons for five frog species by using the four classifiers are listed in Table 5. Notably, F-measure for the recognition of Ornate by using *k*NN-E is only 32.5%, whereas that by using *k*NN-M achieves up to 84.21. It can be inferred that the Euclidian distance measure used for *k*NN degrades the performance due to the correlated features extracted in Ornate samples.

Table 5. F-measure comparison for five frog species

Classifier \ Frog Species	Butleri	Heymonsi	Ornate	Pulchra	Steinegeri
kNN-E	91.66%	91.01%	32.50%	100.00%	98.08%
kNN-M	88.57%	81.60%	84.21%	100.00%	100.00%
SVM	90.01%	76.38%	79.17%	100.00%	100.00%
GMM	91.52%	82.11%	76.01%	89.24%	99.07%

4 Conclusion and the Future Work

An intelligent frog call identification agent is proposed in this work to effectively recognize the five frog species based on the recorded audio samples. An adaptive end-point detection segmentation method is employed to separate the syllables from raw sound samples. Four features, spectral centroid, signal bandwidth, threshold-crossing rate and spectral roll-off, are extracted from the syllables to serve as the parameters for three well-known classifiers, including *k*NN, SVM and GMM. Two kinds of distance measures are used in this work; they are Euclidean distance measure and Mahalanobis distance measure. The Mahalanobis distance measure is used to measure the correlation among the four extracted features. It can be seen from the experimental results that the proposed intelligent frog identification agent is adequate for the identification of frog sounds.

Based on the results of this study, the characteristics of the sounds of some species, such as Microhyla butleri and Microhyla ornate, clearly require further analysis in the future work to improve the syllable segmentation so that the features extracted for the above-mentioned species can be easily recognized by the classifier.

Acknowledgment. The authors would like to thank the National Science Council of Taiwan (Contract No. NSC 96-2628-E-026-001-MY3) and Forestry Bureau Council of Agriculture of Taiwan (Contract No. 96-FM-02.1-P-19(2)) for financially supporting this research.

References

1. Fagerlund, S.: Bird Species Recognition Using Support Vector Machines. EURASIP Journal on Applied Signal Processing (2007)
2. Younis, K., Karim, M., Hardie, R., Loomis, J., Rogers, S., DeSimio, M.: Cluster Merging based on Weighted Mahalanobis Distance with Application in Digital Mammography. In: IEEE Aerospace and Electronics Conference (1998)
3. Vapnik, V.: Statistical Learning Theory. Wiley, Chichester (1998)
4. Chang, C.-C., Lin, C.-J.: LIBSVM: A Library for Support Vector Machines (2003), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
5. Lung, S.-Y.: Feature Extracted from Wavelet Eigenfunction Estimation for Text Independent Speaker Recognition. Pattern Recognition 37, 1543–1544 (2004)

Towards a Distributed Fuzzy Decision Making System

Kevin K.F. Yuen and H.C.W. Lau

Department of Industrial and Systems Engineering,
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
kevinkf.yuen@gmail.com, {ise.kevinyuen,mfhenry}@polyu.edu.hk
Tel.: (852) 60112169, 27666628

Abstract. As many decisions need to be made in enterprises, various Fuzzy Decision Making Systems (FDMS) have been proposed to solve decision making problems. FDMS can be adapted for use with many different problems and they are extensible to cover new problems. The decision making process can also be streamlined by including the benefits of dynamic collaboration. To illustrate this capability this paper proposes a multi-agent framework for FDMS as a distributed fuzzy decision making system (D-FDMS). The multi-agent framework uses fuzzy algorithms as reasoning agents, and applies the architecture concept of Service Oriented Architecture (SOA) which is implemented by web-services. A case study illustrates how this proposed model functions.

Keywords: Agent-Oriented Software Engineering, Fuzzy Decision Making System, Decision Support System, Web Services, SOA.

1 Introduction

In the competitive business era, decision makers always make decisions in an environment of uncertainty, vagueness, and fuzziness or with insufficient information, in a limited time frame. An efficient and effective decision making system plays an essential role for the enterprise in such an environment. In the enterprise, there are various classes of decision making activities such as vendor selection, material selection, location selection, partner selection, promotion selection, product development selection, and in strategy making. Each class involves different projects with different features. Thus reusable design patterns are required for the decision system.

There have been many studies dealing with decision making problems. This research has typically chosen Fuzzy Decision Making Systems [e.g. 1-2, 4, 7, 9, 12-13, 15, 17-22]. In order to incorporate the distributed system with FDMS, the distributed system should be capable of being extended to include new algorithms and incorporate changes in the existing algorithms. As the application domain using FDMS is broad, most studies focus on the mathematical model for the purposes of development. Surprisingly little research can be found in the literature which discusses the aspects of software engineering and computer architecture, and the

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 103–112, 2008.

building of scalable, reusable, distributed, secure and extensible platforms for decision making systems. Also, there is a lack of studies discussing the design patterns and approaches for modeling the algorithms relating to the intelligent agents in the system. As the calculation of the fuzzy decision algorithm is rather complicated, a software agent is the best solution for making the algorithms user friendly. The agent can be shown in a web service which can allow different application agents to use its services.

This paper proposes a Distributed Fuzzy Decision Making System to bridge the gap mentioned above. D-FDMS is a multi-agent framework which applies the concept of a Service-Oriented Architecture (SOA) and is realized by web services. The paper is structured as follows: section 2 briefly reviews the related work on FDMS, and web service-based SOA and agent technology. Section 3 illustrates the inside services of the FDMS agents. Section 4 introduces an SOA environment for D-FDMS. Section 5 introduces the FDMS service agent architecture and its distributed mechanism with client agents. A case study on developing a service agent is discussed in section 6. Section 7 contains a conclusion and suggestions for future study.

2 Related Works

2.1 Fuzzy Decision Making System

This research uses modeling fuzzy theory related to the decision making process which is discussed in many articles [e.g. 1-2, 4,7,9,12-13, 15,17-22].. The main reason for the need for fuzzy theory is that, in the real world, the uncertainty, constraints, and even the vague knowledge of the experts imply that decision makers cannot provide exact numbers to express their opinions [2].

Regarding the decision making process, Yager's [21] is a two stage process: Assessment and Aggregation. The resolution scheme of Roubens [15] and Herrera and Martinez [7] is in two phases: Aggregation and Exploitation. By making a suitable combination, this paper uses the core activities, Assessment, Aggregation and Exploitation, as well as some supportive activities. The algorithms of the different fuzzy decision making approaches are categorized as these three modules. The details of the modeling are discussed in section 3.

2.1 Multi-Agent System and Web Services-Based SOA

An agent is a hardware or (more usually) a software entity with some of these characteristics [5,6,9,16]: ongoing execution, environment awareness, agent awareness, autonomy, adaptiveness, intelligence, mobility, anthropomorphism, and ability to reproduce. A Multi Agent System (MAS) is a group of the agents joined together to complete a task. A MAS has the characteristics [6,8]: (1) each agent has partial information or limited capabilities (knowledge, information, or resources), thus each agent has a limited viewpoint; (2) there is no global system control; (3) data in an MAS are decentralized; (4) computation is asynchronous; and (5) different agents could be heterogeneous, for example, with respect to knowledge representation, data format, reasoning model, solution evaluation criteria, goal, architecture, algorithm, language, or hardware platform.

The conceptual design of MAS can be effectively applied by Services-Oriented Architecture in views of software development. SOA advocates that developers create distributed software systems whose functionality is provided entirely by services [14]. SOA services can be invoked remotely, have well-defined interfaces described in an implementation-independent manner, and are self-contained (each service's task is specific and reusable in isolation from other services) [14]. Web services are becoming the technology of choice for realizing service-oriented architectures (SOAs).[11]. Web services simplify interoperability and, therefore, application integration [11]. They provide a means for wrapping existing applications so developers can access them through standard languages and protocols. [11].Although researchers have proposed various middleware technologies to achieve SOA, Web services standards better satisfy the universal interoperability needs [14]. Services will be invoked using a Simple Object Access Protocol (SOAP) typically over HTTP and will have interfaces described by the Web Services Description Language (WSDL) [3,14].

There is little in the literature discussing the development of the distributed agent system for fuzzy decision making. Regarding applying fuzzy computing and XML, [1] proposed a fuzzy ubiquitous computing framework using XML for Ambient Intelligence. As web services are XML-based, this structure can be further realized by web services. Yuen and Lau [22] introduced the concept of a distributed Fuzzy Qualitative Evaluation System using Intelligent Agent Technology implemented by web service-based SOA. This paper attempts to bridge the gap in development of the distributed agent system for fuzzy decision making.

3 Decision Making Process Service Agents Tier

FDMS needs many agents or a multi-agent system to complete a task. The design of a process service agent is process-oriented from a decision making point of view. The decision making process is complex. A typical decision making process includes three core activities, assessment, aggregation and exploitation, and supporting activities. As the process usually involves various people, collaboration in a single environment is necessary to improve the efficiency and knowledge sharing between them. Thus distributibility and connectivity of the decision system are required. Fig. 1 shows Decision Process Service Agents Tier:

Template Service Agents provide the format to establish the scope and objective of the services, and support the analysis planning of the whole decision support system. Such analysis templates include the services of template design or the knowledge including linguistic terms schema, linguistic defuzzification structure and contents, common evaluation criteria, aggregation and exploitation processes, and the corresponding users and related information.

Definition Service Agents provide the entry to establish the details of the evaluation criteria and statements in the Assessment process, parameters for aggregation and exploitation processes, and the hierarchy showing personal deployment.

Assessment Service agents generate and delivers the correct information to the right people and devices, and also collect the input from the decision makers. Aggregation Process Service agents aggregate the granules of decision inputs from different experts. Exploitation Service agent generate the final results. The algorithms in Aggregation or Exploitation Process Service agents represent how the agents can think and reason. Information service agents are responsible for the retrieval of the decision results from authorized users.

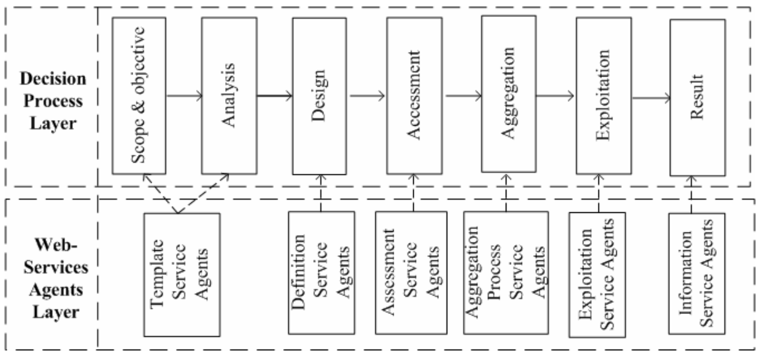


Fig. 1. Decision Process Service Agents Tier

4 D-FDMS SOA Framework

The distributed fuzzy decision making system framework (Fig. 2), which is based on SOA design, contains seven tiers described as follows: In the presentation tier, different roles of users can interact with the systems by using different devices such as web browsers, window client applications, office documents such as spread sheets and office Word documents, as well as mobile devices. Each device can be regarded as an agent and can communicate by web services. In the Security tier, the authentication agent identifies the user. The Authorization agent manages the user's rights, and the Transportation Security Agent responds to encrypt and decrypt data

The Project Service Agents Tier provides services for end users, such as auditors, domain experts and decision makers, so that they can interact with their dedicated projects. This agents tier is dynamically created or inherited from the decision template in the Application Service Agents tier, and further modified to meet the objectives by the domain experts. The Application Service Agents Tier provides a set of decision templates to meet different decision project requirements, thus this layer enables knowledge sharing of designing project criteria, aggregation rules and other attributes. Each template can be regarded as a template service agent such as: Vendor Selection Service Agent, and Partner Selection Service Agent. The Application Service Agent Tier follows the business process managed in the Decision Making Process service agents tier which was discussed in the previous section.

A Component Service Agents Tier consists of four component service agents. The Intelligence Service agent contains fuzzy algorithms for linguistic modeling, linguistic defuzzification, granular aggregation, fuzzy rule reasoning, and result exploitation. The Resources Access Service Agents provide the services for data access among their upper and low tiers. The External Access Agent provides services to access external agents that are out of this framework such as email server agents. The Common Library includes the general components for the application such as the Graphic Engine agent which is usually created from third party software obtained from vendors.

The Databases Service Agents Tier contains three database systems which perform different roles. The Intelligent database service agent provides the general patterns of parametric inputs from domain experts. The Knowledge Database Service Agent stores the general templates for each project and the templates are derived from the application services layer. The Information Database Service Agent organizes the data for each project queried and updated in the Project Service Agent Tier. Separation of these three databases services improves accessibility and manageability of the system.

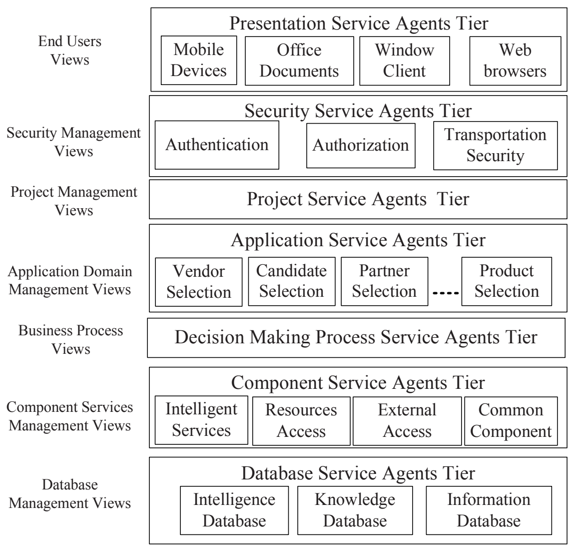


Fig. 2. D-FDMS SOA Framework

5 Agent Architecture and Communication Mechanism

5.1 The Agent Architecture

For realization of these intelligent components as distributed intelligent components, the distributed mechanisms should be developed in the components. Several mechanisms such as CORBRA, .Net Remoting, COM+ are the possible ways. However, they face challenges especially development workloads for cross platforms, cross applications, reusability and extensibility. Web services are the ideal distributed mechanism. The advantages of using web services are that web services provide flexible (loose) binding, and service descriptions can be easily extended and upgraded. Fig. 3 illustrates the distributed architecture of a FDMS Agent, as follows:

The Web Service interface layer includes a service contract (e.g. WSDL) and a service adapter (e.g. disco or UDDI). The Service contract defines the behaviors of a service and the messages required as the basis for interaction. A Service adapter implements the service contract and to expose this functionality at a certain endpoint.

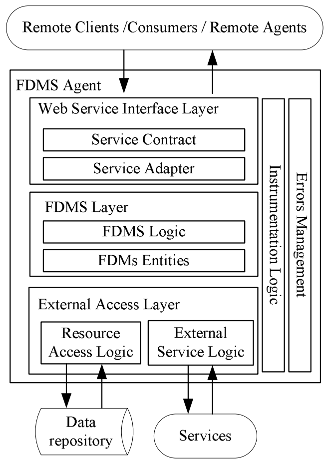


Fig. 3. FDMS service Agent Architecture

The FDMS layer incorporates FDMS Logic and FDMS Entities that are used by the FDMS logic components. FDMS logic contains the methods. FDMS Entities contains the data entity objects and data type mapping. The separation is typically beneficial for code management such as the extensibility of the application. For example, if the data type is changed in the remote agents, the developers do not need to expend much effort to modify the codes. Instead, the developers simply modify the entities files to meet the requirements of the remote agents.

The External Access Layer includes Resource Access Logic and External Service Logic. Resource Access Logic includes the methods to access the external data repository through database service agents. External service Logic includes the logic to contact the external service through external service agent. External services may be other intelligent service systems that are capable of integrating with the FDMS system.

Exception management and instrumentation are used for troubleshooting. Instrumentation logic includes the logic to control the transaction logs such as the number of requests and responses, system failures and the length of the queue for the messages. Error management handles any invalid SOAP requests and system failures.

5.2 Distribution Mechanism

Web service applications are the technologies which send and receive SOAP messages, expose WSDL contracts, and embed the functionality that is needed to fulfill the service's behavior. Fig. 4 shows the distributed communication mechanism between remote agents in the application service agents layer, using web services. The communication process has two main stages: Web Services Definition Discovery and Web Reference Creation, i.e. steps 1 to 4, and Object Creation and Task Completion, i.e. steps 5 to 14.

In phase one, remote agents explore the service contracts in the FDMS service agent, and creates Web Reference classes. The word "remote" means the SOAP client, e.g. the application service agents tier. The web methods of the FDMS service agent are described by the Web Services Definition Language (WSDL) files, which are the XML schemas describing the interfaces, data types, attributes, methods of the services and binding information. The XML schema files, or the WSDL contracts, are generated by the FDMS Service Agent. The web services interface is created to display the services of FDMS using the web methods. In short, services are described by a WSDL contract, not by type libraries. Remote clients can discover the services and make a web reference through the interfaces. This process in fact creates the remote web reference classes in the SOAP client. Thus a client application in this remote web reference class can communicate with the FDMS service agent using SOAP.

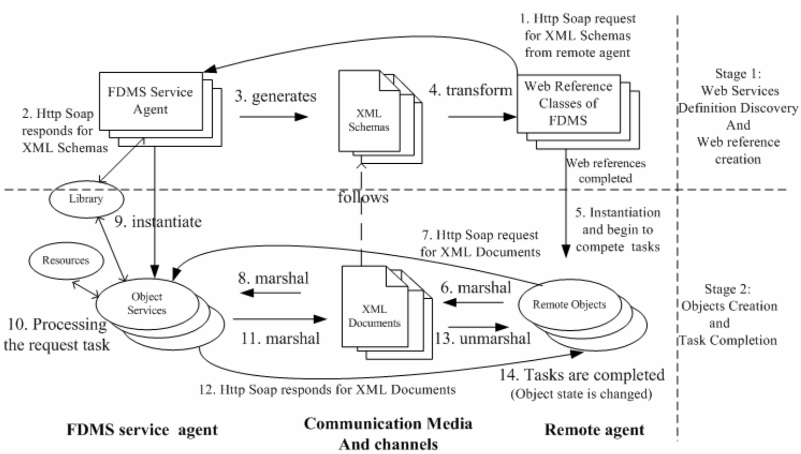


Fig. 4. Communication Mechanism of D-FDMS

In phase two, the remote objects that send the messages to the FDMS service agent invoke the agent's associated operation via the web services. The remote object is instantiated from the web reference classes. The object with request parameters is further marshaled to the XML documents which describe the structure of the object with request parameters. Next it will be sent as SOAP (simple object access protocol) message, which is an XML document containing the marshaled XML about the requested objects, to the FDMS service agent by Http Request. When the SOAP messages arrive in the FDMS service agent, the XML files are unmarshaled into objects on the basis of instantiation in the service agent. The FDMS agent processes the requests, and responds with objects. These objects are marshaled as another XML

Document again, which will be packed as a SOAP message and will be sent to the SOAP client via Http Responds. When the soap message is received in the remote side, the XML file containing the object's information is unmarshaled into new remote objects. The SOAP client will use these new objects for further processing. In addition, these two binding relationships are named as marshal and unmeshed in java. In .Net framework, they are called Serialization and Deserialization.

6 Case Study

In this case study, OWA Aggregation operator [20], one of the aggregation operators, is used as the reasoning algorithm for the Aggregation Process Service Agent. The

Fig. 5. Web Reference files in client agent

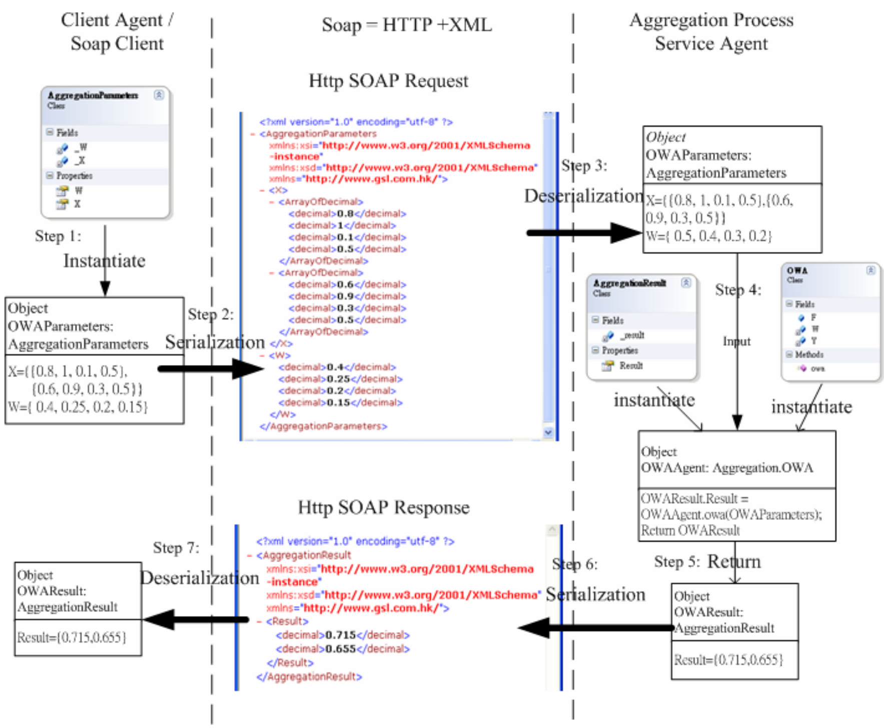


Fig. 6. Distribution Mechanism in Aggregation Service Agents

agent is implemented by Microsoft .net technology. In phase one, after the SOAP client agent has discovered Web Services Definition and added Web References from the remote agent, the agent now contains three sets of files named: disco, wsdl and map (Fig.5). WSDL describes the services from the provider. The Discovery Protocol (disco) specification enables clients to discover services at a known URL. (Disco is different from UDDI, which is outside the scope of this paper). Map is used for mapping web services with local files. The related documents can be found in [10].

In phase two, after the soap client agent is built, it will perform various operations and perform the services provided by the service provider- OWA APS agent. Fig.6 illustrates the mechanism of how the remote agent requests services from the Aggregation Process Service Agent in order to complete its tasks. The distribution mechanism is based on the description from section 5.2.

7 Conclusion and Future Work

Fuzzy decision making problems have attracted the interest of many researchers who have discussed the algorithms for this solution. However, in the literature, we did not find any other approaches to model and implement the FDMS using web service-based SOA. This paper proposes a framework using web-service based SOA agent technology for FDMS. The multi-agent system framework using the SOA as conceptual architecture, and web services as realization tools, is applied to FDMS. The contribution of this model is to illustrate the fundamental concept to enable system architects to form solutions to the problems of the multi-service agent who wishes to build the D-FDMS system. A case study illustrates how the proposed model functions. The advantages of our proposed design include role separations, code reusability, knowledge sharing, extensibility, and cross platforms and cross applications.

In this paper, the FDMS model addresses only a small portion of the field. As the discussion of the algorithms is a potentially fruitful field, ongoing research is needed into the ability of Unified Modeling Language to model some well-known FDMS algorithms as modules which will contribute to software design. As high degrees of flexibility, extensibility, and connectivity are required in order to implement the system, more discussion of the proposed framework is needed and more research needs to be carried out. A further study will be implemented the system in the company, and thus a case study will be conducted to test the validity of the system.

Acknowledgements

The authors wish to thank the Research Office of the Hong Kong Polytechnic University for supporting this research project. This research was supported in part through a grant from the Innovation Technology Fund (UIT/74) from the Innovation and Technology Commission HKSAR and Group Sense Limited. Thanks are also extended to the editors and anonymous referees for their constructive criticisms to improve the work.

References

1. Acampora, G., Loia, V.: A proposal of ubiquitous fuzzy computing for Ambient Intelligence. Information Sciences (to be published, 2007)
2. Ben-Arieh, D., Chen, Z.: Linguistic-Labels Aggregation and consensus measure for autocratic decision making using group recommendations. IEEE Transactions on Systems, Man, and Cybernetics-Part A: system and Humans 36(3) (May 2006)
3. Christensen, E., Curbera, F., Meredith, G., Weerawarana, S.: Web Services Description Language (WSDL) 1.1. W3C note (March 2001), `http://www.w3.org/TR/wsdl.html`
4. Ding, J.-F., Liang, G.S.: Using fuzzy MCDM to select partners of strategic alliances for linear Shipping. Information Sciences 173, 197–255 (2005)
5. Ferber, J.: Multi-agent systems: an introduction to distributed artificial intelligence. Addison- Wesley, Reading (1999)
6. He, M.H., Leung, H.-f.: Agents in E-commerce: State of the Art. Knowledge and Information System 4, 257–282 (2002)
7. Herrera, F., Martinez, L.: A Model based on Linguistic 2-tuples for dealing with multigranular Hierarchical Linguistic Contexts in Multi-Expert Decision-Making. IEEE Trans on SMC, Part B 31(2), 227–234 (2001)
8. Jennings, N., Sycara, K., Wooldridge, M.: A roadmap of agent research and development. Autonomous Agent and Multi-Agent Systems 1, 7–38 (1998)
9. Jiang, Y.-P., Fan, Z.-P., Ma, J.: A method for group decision making with multi-granularity linguistic assessment information. Information Science (to be published, 2007)
10. Microsoft developer network (2007), `http://msdn.microsoft.com`
11. Nezhad, H.R.M., Benatallah, B., Casati, F., Toumani, F.: Web Services Interoperability specifications. Computer, 24–32 (May 2006)
12. Olcer, A.I., Odabasi, A.Y.: A new fuzzy multiple attributed group decision making methodology and its application to production/manoeuvring system selection problem. European Journal of Operational Research 166, 93–114 (2005)
13. Pasi, G., Yager, R.R.: Modeling the concept of majority opinion in group decision making. Information Science 176, 390–414 (2006)
14. Pasley, J.: How BPEL and SOA Are Changing Web Services Development. IEEE Internet Computing, 60–67 (May-June 2005)
15. Roubens, M.: Fuzzy Sets and Decision Analysis. Fuzzy Sets Syst. 90, 199–206 (1997)
16. Shoham, Y.: What we talk about when talk about software agents. IEEE Intelligent Systems 14(2), 28–31 (1999)
17. Sirbiladze, G., Gachechiladze, T.: Restored fuzzy measures in expert decision-making. Information Sciences 169, 71–95 (2005)
18. Tzeng, G.-H., Yang, Y.-P.O., Lin, C.-T., Chen, C.-B.: Hierarchical MADM with fuzzy integral for evaluating enterprise intranet web sites. Information Sciences 169, 409–426 (2005)
19. Viedma, E.H., Martinez, L., Mata, F., Chiclana, F.: A Consensus Support System Model for Group Decision-Making Problems With Multigranular Linguistic Preference Relations. IEEE Transactions On Fuzzy Systems 13(5) (October 2005)
20. Yager, R.R.: Ordered weighted averaging aggregation operators in multicriteria decision making. IEEE trans on Systems, Man and Cybernetics 18, 183–190 (1988)
21. Yager, R.R.: Non-numeric multi-criteria multi-person decision making. Group Decision and Negotiation 2, 81–93 (1993)
22. Yuen, K.K.F., Lau, H.C.W.: A Distributed Fuzzy Qualitative Evaluation System. In: Proceedings of 2006 IEEE / WIC / ACM International Conference on Intelligent Agent Technology, Hong Kong, December 18-22, 2006, pp. 560–563 (2006)

A Color Set Web-Based Agent System for 2-Dimension Emotion Image Space

Pei-Lin Chen[1] and Jen Yen[2]

[1] Graduate School of Computational Design, National University Yunlin Science and Technology, 123 Section 3, University Road, Douliou, Yunlin, Taiwan, R.O.C.
g9534702@yuntech.edu.tw
[2] Graduate school of Visual Communication Design, National University Yunlin Science and Technology, 123 Section 3, University Road, Douliou, Yunlin, Taiwan, R.O.C.
yenj@yuntech.edu.tw

Abstract. In this paper we present a conceptual framework for various emotion image spaces based on web design. In order to assist novices with color set as senior designers, we build a web-based agent system for novices. We use questionnaire to get sample data and use multiple regression analysis to predict the population of various emotion image spaces. The system can tell designers that whether the image-words match various colors or not. It provides designer suggestions of color set and analyzes. To conclude, the concept of this research can apply to all design domains. We build the emotion image database fits any questionnaire and the color set of this web-based agent system can stand for local color and culture.

Keywords: Web-based agent, color science, color set, regression analysis.

1 Introduction

Even just hundreds of colors, coloring process of 2D graphic design are already a time consuming process. Therefore, providing a traditional color picker is not effective enough for designers (especial novice designers), neither intelligent nor efficient. Color image words are the most used approach (up to day) for understanding and interpreting emotion images pairs. Novice designers may lack of experience or knowledge of color sets as compared with experts. This is to say that if the expertise of coloring can be computed and represented computationally. The effective usages of such knowledge can be significant. Furthermore, two main issues raised in this context: how color-image-words can be adapted in mapping the expert knowledge? And how this knowledge can be retrieved?

Regarding of the image-words, most products use two image-words in Combination while related researches often just use a single image-word. Consumers frequently use color words when describing their perceptions of a product. However, only having one image word allows consumers to choose the preferred image in currently studies.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 113–121, 2008.

For implementing issues, we use Web-based technology since most of our works for coloring is done in web-based media. In addition, with agent-based design strategy, the little helper of our coloring system is designed as a agent-based supporting system, called Novice Coloring Supporting System (NCSS). (See in Figure6)

Consequently, the hypothesis unleashed in this research is as follow:

Table 1. Study hypothesis

uestion: Can various emotion images be computational?
Proof. ………….
Answer: Color Set Web-based Agent System of Various Emotion Image

2 Literature Reviews

Then in reviews, what we are going to find out for building up a novice coloring supporting system (NCSS)? What should we know? We review cases for similar approaches (a supporting system / system cases / user analysis for designers) and coloring theories (modern) and image-words application. And we also study the case of Web-based.

2.1 Similar Approaches

Tsai (2007) provides extensive discussions of the applications. Tsai's paper offered a sounder theoretical basis for analyzing two evaluation method. The study develops an automatic design support system to search for the near-optimal two colors set, and combines a gray theory-based color–association evaluation method and a color-harmony-based aesthetic evaluation method. We use multiple regressions to replace the gray theory. It cans analysis to predict the population of emotion image spaces. (See in Figure1)

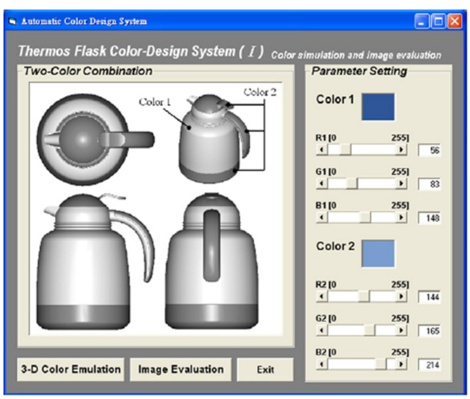


Fig. 1. Interface for constructing 3-D model and performing color-image prediction

Min-Yuan (2007) presents a decision-making support model that assists consumers in choosing among options and can effectively design products. He discusses two issues: "(1) factor analysis is utilized to differentiate multifarious products into several styles, which will allow consumers to choose one style according to their preferences. (2)The fuzzy analytic hierarchy process (FAHP) combined with the image compositing technique is applied to construct the design decision-making support model provided for choosing the optimum product." (See in Figure2)

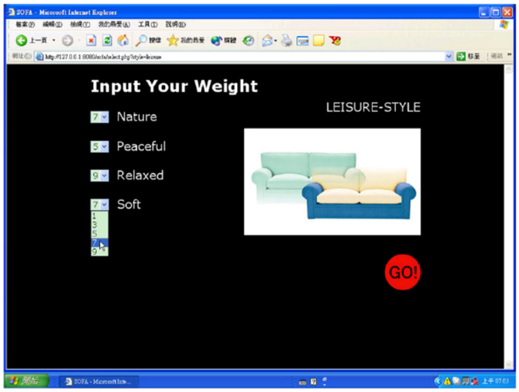


Fig. 2. The customer inputs the required image weights

Ming Ronnier (2005) describes that "some computer systems have been recently developed to assist designers for designing color palettes through color selection by means of a number of widely used color order systems, for creating harmonized color schemes via a categorical color system, for generating emotion colors using various color emotional scales and for facilitating color naming via a color-name library." (See in Figure3)

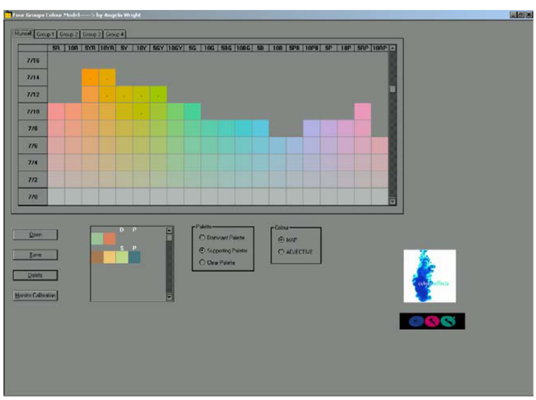


Fig. 3. Color design system

2.2 Image-Words Application

Giani (2006) describes the emotions collection as short sentences were first codified online by means of a set of labels negotiated at distance by two members of the research staff. The codes were then iteratively revised by the members of the research staff until a shared codification of each emotional statement was obtained. (See in Table2).

Table 2. Codes of the emotions

Appreciation	Panic
Disappointment	Bewilderment
Dissent	Professional detachment
Empathy	Regret
Esteem	Sadness
Helplessness	Surprise
Identification	Tenderness
Impulse to be active	Understanding
Leniency	Worry

2.3 Factor Analysis

Factor analysis was originally developed by psychologists. The subject was first put on a respectable statistical footing in the early 1940s by restricting attention to one particular form of factor analysis that based on maximum likelihood estimation. [Mardia, 1979]

Factor analysis is a mathematical model which attempts to explain the correlation between a large set of variables in terms of a small number of underlying *factors*. A major assumption of factor analysis is that it is not possible to observe these factors directly; the variables depend upon the factors but are also subject to random errors. Such an assumption is particularly well-suited to subjects like psychology where it is not possible to measure exactly the concepts one is interested in (e.g. ”intelligence”), and in fact it is often ambiguous just how to define these concept. [Mardia, 1979]

The definition of the factor model:

Let $x(p\times 1)$ be a random vector with mean μ and covariance matrix $\sum$.

Then we say that the k-factor model holds for x if x can be written in the form

$$x = \Lambda f + u + \mu$$

Where $\Lambda(p\times k)$ is a matrix of constants and $f(k\times 1)$ and $u(p\times 1)$ are random vectors. (1)

The elements of f are called common factors and the elements of u specific or unique factors.

2.4 Multiple Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical sciences, economics, management, and life and biological sciences, and social sciences. In fact, regression analysis may be the most widely used statistical technique. [NetLibrary, 1998]

A regression model that involves more than one regressor variable is called a multiple regression model. If two or more regressors are highly correlated, we say that multicollinearity is present in the data. Regression models fit to data by the method by the method of least squares when strong multicollinearity is present are notoriously poor prediction equations, and the values of the regression coefficients are often very sensitive to the data in the particular sample collected. [NetLibrary, 1998]

I use multiple regression analysis to predict the sample of the population space.

Definition of multiple regressions

The **general additive multiple regression model equation** is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dotplus \ldots + \beta_k x_k + \varepsilon \tag{2}$$

where $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. In addition, for purposes of testing hypotheses and calculating CIs or PIs, it is assumed that ε is normally distributed.

2.5 Web-Based Agent

A case of web-based agent is UDDI. Shuying (2006) describes "the service-orientated computing paradigm is transforming traditional workflow management from a close and centralized control system into a worldwide dynamic business process. A complete workflow serving inter-enterprise collaboration should include both internal processes and ad hoc external processes. This paper presents an agent-based workflow model to address this challenge. In the proposed model, agent-based technology provides the workflow coordination at both inter- and intra-enterprise levels while Web service-based technology provides infrastructures for messaging, service description and workflow enactment. A proof-of-concept prototype system simulating the order entry, partner search and selection, and contracting in a virtual enterprise creation scenario are implemented to demonstrate the dynamic workflow definition and execution for inter-enterprise collaboration."

3 Methodology and Implementation

There are two parts in the study methods, one is to build emotion image words questionnaire, and the other is to build the database of two emotion images. The first part of building various emotion images database is the method of questionnaire survey. The second part is method of statistics analysis includes factor analysis, and multiple linear regression analysis. (See in Figure4)

3.1 Build Single-Color Sample and Prototypical Eighteen Image Words

Step1. The first part of questionnaire survey

Giani's study(2006) provides the eighteen code of emotions that are "Appreciation", "Disappointment", "Dissent, "Empathy", "Esteem"," Helplessness", "Identification", "Impulse to be active", "Leniency", "Panic", "Bewilderment", "Professional detachment", "Regret", "Sadness", "Surprise", "Tenderness", "Understanding", "Worry".

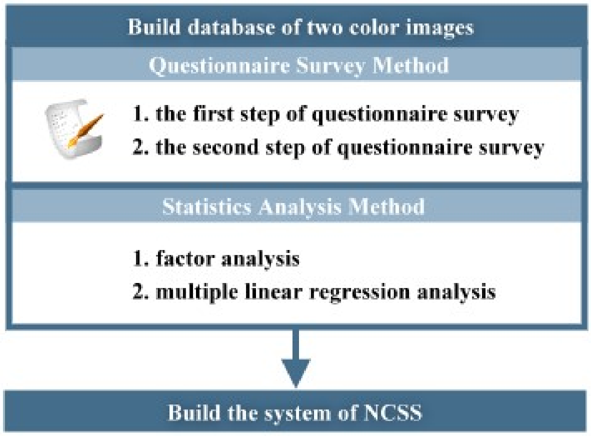


Fig. 4. Flow hart

And we choose 30 color set experts of having more than ten years design experience of including 20 graduates of visual communication department and 10 teachers to follow the single emotion image evaluation experiments.

Step2. The second part of questionnaire survey
30 color set experts of having more than ten years design experience of including 20 graduates of visual communication department and 10 teachers follow the single emotion image evaluation experiments. The method is to use the eighteen image words of the fittest emotion image from judgment sampling to match color of three or five color samples. Then we cancel repeated color sets and choose pre-ten color set to for the user test questionnaire survey to verify. Then we use those color set to be the web-based agent system of basis color set samples.

3.2 Build the Emotion Image Database

To find the mapping color of the image words and let the different senior designers select three or five corresponding color Combination of eighteen words of previous emotion image words. Then I use this sample data to generate the emotion image space between 3 to -3 of x-coordinates vale. To find the same attributes of those sample data and to analyze the factors, then we use the factors to conjecture the population space. Use the mathematics method of multiple regression analysis to predict the 2-dimension emotion images space.

Step1. Factor analysis experiments
Then we use semantic differential to design the first part of questionnaire and choose factor analysis to analyze. This basic single-color sample use CIE RGB to create 120 basic single-color samples, including ten kinds of Hun, eleventh kinds of Tone, and ten neutral. (See in Table5) We also choose method of factor analysis to verify attributes of color set samples and to find experimenter, explanatory variables, of a multiple linear regression model. The variables are common attributes of color set samples. We use explanatory variables to calculate the observed value, response variable.

Step2. Multiple linear regression analysis experiments

We use statistics method of multiple linear regression analysis to predict color set population of having common attributes (x_i). The method is to use less color set samples to predict the multiple linear regression model, the two emotion image space, and to use normalization method to acquire the closest population space. (See in Figure5)

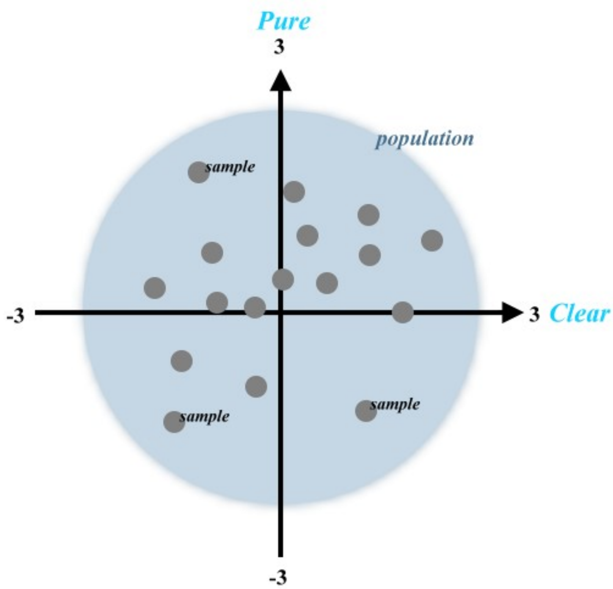


Fig. 5. Using sample to predict population

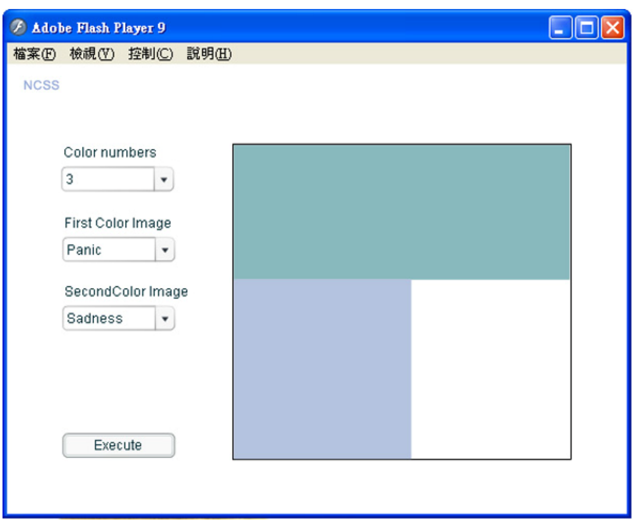


Fig. 6. Novice Coloring Supporting System

4 Result and Discussion

The system combines the several results of objective determinations and provides various virtual images for styles among which designers can compare, and each designer obtains topic combinations from their satisfied selections.

The system can tell designers that whether the color set match the topic emotion images or not. It provides designer suggestions of color set and analyzes. Furthermore, the system can tell the immediate design digresses or not, which is shown by the degrees of percentage numbers.

Additionally, the image compositing approach allows designers to view the several combinations and avoid wasting time evaluating too many options and efficiently choose color combinations according to their needs.

I use the statistics method of multiple regression analysis to predict the 2-dimension emotion images space and find the statistics formulas. The systems also can build 2-dimension emotion images spaces and the emotion image database may fits other questionnaires and the color combination from the system can stand for local color and culture.

5 Conclusion and Further Work

In the past, the emotion images for color set have less research papers, the thing points the emotion image is huger and more complex. While most of researchers use color harmony to evaluate the aesthetics of design, they are not suitable for this research with three arguments: (a) the size of the graphic area affects the human feeling, (b) the relative location influences the human vision; (c) the definition of emotion image words makes different result. Therefore, we only use the methodology from knowledge-base system to achieve the knowledge of coloring. There are enormous sample data that I should analyze and then predict the 2-dimension emotion images space. So I choose simple and useful eighteen emotion image-words, and analyze factor precisely to reduce sample errors.

In the future, this study can continuously on various emotions using color theories. Then evaluating those complicated emotion represents what kind of color combination. I would like to find mathematics formula of color combination rules upon mathematics theorem. After that, I will use additive mixture and subtractive mixture of colors, and color combination algorithm to show color mixture's color which is produced by various colors.

References

1. Hayzelden, A.L.G., Bourne, R.A.: Agent Technology for Communication Infrastructures, England (2001)
2. Craig, S.A.: Emotion, Cognitive, and Decision Making. Psychology Press, UK: East Sussex (2000)
3. Devore, J.L.: Probability and Statistics. Thomson, USA (2004)
4. Richard, D.N., Harry, S.: NetLibrary Inc. Applied Regression Analysis. Wiley, UK (1998)
5. Eiseman, L.: Pantone® Guide to Communication with Color. Grafix Press, USA (2000)

6. Ferber, J.: Multi-Agent Systems: An Introduction to Distributed Artifical Intelligence, New York (1995)
7. Agneta, H.F., Antony, S.R.M.: Culture and Emotion: A Special Issue of Cognition and Emotion. Psychology Press, UK: East Sussex (2002)
8. Kuehni Rolf, G.: Color: An Introduction to Practice and Principles. Wiley, UK (2005)
9. Kim, H.S., Cho, S.B.: Application of Interactive Genetic Algorithm to Fashion Design. Engineering Applications of Artificial Intelligence 13, 9 (2000)
10. Shevell Steven, K.: Optical Society of America, the Science of Color. Elsevier, UK: Elsevier (2003)
11. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis, USA (1979)
12. Kalay, Y.E.: Computability of Design, USA (1987)
13. Liu, c.-j.: An Adaptive Design Support System for Interactive Image Searching with Data Mining Methodology (2006)
14. Liu, H.-w.: A Collaborative Knowledge Management and Representing Different Industrial Design Stakeholders by Multi-Agents -the Case of Company Asus (2007)
15. Liu, M.C.: A Morphing Method for Shape Generation and Image Prediction in Product Design 23, 23 (2002)
16. Luo, M.R.: Applying Color Science in Color Design. Optics & Laser Technology 38, 7 (2005)
17. d'Inverno, M., Luck, M.: Understanding Agent Systems, New York (2004)
18. Wei, C.: Color Coordination Support System Based on Color Image. National Yunlin University of Science & Technology (2005)
19. Ma, M.-Y., Chen, C.-Y., Wu, F.-G.: A Design Decision-Making Support Model for Customized Product Color Combination. Computers in Industry 58, 15 (2007)
20. Montgomery Douglas, C., Peck Elizabeth, A.: Introduction to Linear Regression Analysis. Wiley, UK (1992)
21. Picard, R.W.: Affective Computing. The MIT Press, England (1997)
22. Karsten, R., Kapatthi, S.: Using Dynamic Explanations to Enhance Novice Programmer Instruction Via the Www. Computers Edvc. 30, 7 (1998)
23. Wang, S., Shen*, W., Hao, Q.: An Agent-Based Web Service Workflow Model for Inter-Enterprise Collaboration. Expert Systems with Applications 31, 13 (2006)
24. Stiles, G.W.W.S.: Color Science: Concepts and Methods, Quantitative Data and Formulae, USA (2000)
25. Tsai, H.-C., Choub, J.-R.: Automatic Design Support and Image Evaluation of Two-Coloured Products Using Colour Association and Colour Harmony Scales and Genetic Algorithm. Computer-Aided Design, 11 (2007)
26. Giani, U., Brascio, G., Bruzzese, D., Garzillo, C., Vigilante, S.: Emotional and Cognitive Information Processing in Web-Based
27. Medical Education. Biomedical Informatics 40, 11 (2007)
28. Walpole, Myers, Myers, Y.: Probability and Statistics for Engineers and Scientists, USA (2002)
29. Yang, Y.-H., Su, Y.-F., Lin, Y.-C., Chen, H.H.: Music Emotion Recognition: The Role of Individuality, Taiwan

Approach to Solving Security Problems Using Meta-Agents in Multi Agent System

Esmiralda Moradian and Anne Håkansson

Department of Information Science, Computer Science,
Uppsala University,
Box 513, SE-751 20, Uppsala, Sweden
Esmiralda.Moradian@dis.uu.se, Anne.Hakansson@dis.uu.se

Abstract. In this paper, we present an approach of integrating meta-agents in Web Services. Web Services requires an appropriate level of security to support business systems to be used by anyone, anywhere, at anytime and any platform. However, the increased use of distributed systems with message passing produces a growing set of security problems. This especially concerns threats and attacks on Web Services consisting of transactions with valuable resources. To prevent possible attacks on Web Services, we propose using meta-agents over software agents in a multi-agent system. The multi-agent system consists of a network of software agents that secure the message passing at the transport and application levels. To avoid attacks on these agents, we use meta-agents to monitor the software agents' actions and then to direct the software agents' work. These meta-agents are also used to handle unexpected events in the actions of the software agents.

Keywords: Software agents, meta-agents, intelligent agents, multi-agent systems, networks, Web Services, SOAP, security, reliability, auditing.

1 Introduction

Due to a growing use of the Web for commercial and business applications, security has become more important. Web Services promise business systems to be used by anyone, anywhere, at anytime and any kind of platform [18]. However, this requires that the Web Services have an appropriate level of security including confidentiality, integrity, reliability, and availability. With the increased number of users, the number of attacks and threats has increased. These especially concern Web Services consisting of transactions with valuable resources, which are communicated by following a request/response model of interaction [5].

A primary purpose of Web Services architecture is to integrate different Web Services and to open up the corporate network to the external world. Applications based on Web Services provide broad connectivity across a multitude of back-end systems [14]. This facility has increased the number of enterprises and organizations that discover the strength of Web Services. Unfortunately, the Web Services can challenge privacy and security of the corporation since Web Services open pipelines to its most valuable information assets, which can be used by an attacker [14].

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 122–131, 2008.

One of the main challenges of Web Services security is to maintain security while routing between multiple Web Services [5]. The routing is the message carrying facility where WS-Routing defines “how to insert routing information into the header of a SOAP message through a chain of intermediaries” [5]. Every SOAP message has a sender, an ultimate receiver and an arbitrary number of nodes, also called intermediaries, which process the message and route it to the receiver. The challenge is to prevent the attacks, which is not easy since one SOAP message may traverse multiple SOAP “hops” between the originator and the endpoint [13].

One of the benefits with SOAP, which is also a problem for security reasons, is that SOAP does not require a strong connection between the client and the server. There is no centralized middleware in Web Services and, therefore, when invoking the service it is necessary to specify the address (URI) of the service and the transport protocol (HTTP). Recently, the number of threats and attacks on Web Services has become a larger problem. The attacks on SOAP are now more sophisticated than earlier, making the web an unsafe media for passing delicate information.

To decrease the number of attacks, we present a solution of using meta-agents on top of software agents in a multi-agent system on the networks that can provide secure routing for SOAP messages. In the system, the software agents perform message passing at the transport level and the application level. These agents work as message carriers between routers, at which the agents collect information about the routers and store the information to be reused in the next request to the same router. Thus, the software agents carry sensitive messages through the Web by using the predefined routers, which can be reliable for avoiding security problems. Carrying secure messages requires total insight and control of the message, including the protocols and the servers involved in the message passing. Our solution is applying meta-agents and integrating them in Web Services architecture. The meta-agents are created from the software agents and, then, used to supervise the agents when they are carrying the sensitive messages. The meta-agents give better control of the nodes that the SOAP messages are transported through. Hence, by using meta-agents there will be possibility of providing intelligent information management.

2 Related Work

Multi-agent systems have been developed for network security. One solution is to build co-operative agents, responsible for communication between different network nodes. The co-operative agents determine if suspicious events are a part of a distributed attack and, if so, warn other agents about the possible threats [6]. Another solution is to adopt agents for e-commerce, to handle trust, security and legal issues. For example, the author, Fasli, (2007) discuss issues regarding security, and how trust can be addressed through the use of cryptography [1]. Other solutions are to simulate network attacks [2].

Adsett *et al.*, (2004) propose a method and an implementation to realize weak workflow, called DeFleX, [9]. We propose another method, namely using meta-agents to monitor and control the message path, starting from initial sender via intermediaries to ultimate receiver. The meta-agents communicate the message between the ground-level agents at the transport level and the application level. These ground-level agents report the possible attacks back to the meta-agent. Furthermore, using meta-agents makes it possible to monitor and control message on transport level as well, since each meta-agent can travel with the SOAP message across different layers in OSI model. Hence, instead of solving security problems with cryptography we apply agents in network by using Web Services to avoid inappropriate routes in the network.

3 Web Services

WS-security should protect data both in storage and transit. Business data transfers across untrusted networks and is manipulated by numerous distributed systems. If any part of this chain is compromised, the entire WS-security model breaks down [18].

As long as the packages are transported within the network of the organisation, the transfer can be considered safe. However, as Peterson (2006) says, "The Web Services approach involves interoperability with many services that are not in your organization's direct control"[15] and, moreover, there is a limited ability to provide assurance proof [15]. Thus, the control of the transfer is lost. Therefore, monitoring and auditing Web Services operations, for security violations by using meta-agents, is an essential aspect of defence in depth.

Web Services rely on message security that focuses on confidentiality and integrity. A shortcoming of Web Services lies in its use of HTTP for message transport. The HTTP transport can use Secure Socket Layer (SSL) as a tunnel for securing message content. However, this is not an adequate solution for protecting message during transport where the scenario is a SOAP request, routed via different intermediaries. In a Web Services context, the problem of several intermediaries is unavoidable. SOAP is designed to support more than one intermediary that can forward or reroute SOAP messages based upon information either in the SOAP header or in the HTTP header [17].

3.1 SOAP and WS- Routing

SOAP is a lightweight protocol recommended by the W3C. It is the essential component for exchanging structured information in a Web Services decentralized, distributed environment. SOAP is used to carry a payload from the initial SOAP message sender, either directly to ultimate SOAP receiver or via one or more SOAP nodes. Each SOAP node can act as sender, receiver and intermediary [7].

SOAP messages are in the form of SOAP envelope, which is divided in two parts: a header and a body. The header is an optional part, while message body is mandatory part of the SOAP message. When processing a message, SOAP header elements are targeted at different SOAP nodes, where they can play one or more roles, i.e., *next, ultimateReceiver or none* [5]. "Next" is especially important from a security point of view. "Each SOAP intermediary and the ultimate SOAP receiver MUST act upon next" [7] making SOAP messages pass through multiple hops compromising security.

SOAP does not define any security mechanism for Web Services. However, the SOAP security extensions can be defined as, a security element in the headers of the SOAP message. The security elements have specific roles, where each role can have at most one security element. Security elements contain all security information relevant to the role [5]. Message security information, targeted at different receivers, must appear in different security header elements.

According to W3C specification, SOAP intermediaries can be considered as "men-in-the-middle" [7]. Compromising one of intermediate stations opens possibilities for an attacker to perform a man-in-the-middle attack by inserting fake routing instructions. The message travels to a malicious location from which the attacker can send malicious instruction to original destination [16].

Security breaches in systems, that run SOAP intermediaries, can result in serious security problems. A compromised SOAP intermediary, an implemented or configured intermediary without security considerations, might be used in a wide range of potential attacks. There is no requirement in SOAP that all hops between participating SOAP nodes use the same underlying protocol. Even so, the use of SOAP intermediaries is likely to reach beyond the scope of transport-level security [7]. SOAP does not define any routing or forward semantics corresponding with a message path.

Until now it has been impossible to specify which intermediaries should be visited by a SOAP message while routing to its destination. Recently, SOAP Routing Protocol (WS-Routing) opens the possibility to describe the "entire" message path within the SOAP message structure using the SOAP extensibility model. However, WS-Routing is not intended to provide a "complete set of services often considered part of reliable messaging, security, or other services that may be required in a messaging environment" [8]. WS-Routing intermediaries may, dynamically, insert additional nodes in the forward message path and there is no requirement that the complete forward message is known at the time it leaves the initial WS-Routing sender [8].

4 The Software Agents

As the network grows in size, it becomes essential to control the flow of information across organisations. Using a multi-agent system, where each software agent acts as a ground-level information collecting agent, would be an efficient and ubiquitous way of monitoring and controlling the network.

The exact structure of the routing chain of SOAP message is unknown to both service requestor and service provider. By using meta-level agents, we have the possibility to identify the exact structure of the routing chain of a SOAP message since the meta-level agent monitors the whole routing path (or chain), and, if needed, generates a record listing the nodes that were passed. Moreover, meta-level agent can calculate the time it actually took for a specific node to process the message.

The software agents work with one task at the time in a straightforward environment. The environment is fully observable, deterministic, episodic, static and discrete. In a fully observable environment, the agents have access to the complete state of the environment at each point at the time [11]. The agents can detect all aspects that are relevant to the choice of action. In our system, the software agents have access to all information about IP addresses, protocols, servers, routers and the all the components involved in the message passing. The agents follow the messages from one computer, i.e., server or router, to another, where the route in between is determined by the protocol, at application-level and transport-level. For each request, the agents send information about the transport to a meta-level agent, which has to judge the success of the transport. A successful transport will pass without restrictions but an unsuccessful transport will be halted and cancelled.

The agent works in an environment that is deterministic. Deterministic means that next state of the environment is determined by the current state and the action that is being executed by an agent [11]. Our software agents work in a deterministic environment. All information about the steps is found in the servers and the protocols. When an agent has reached its goal, i.e. an IP address, next agent proceeds by taking the message to the next IP address according to the information in the protocol. The information about the IP address is sent to a meta-level agent, which keeps track of the all IP addresses that have been visited.

Episodic defines the task environment [11]. The task environment is divided into atomic episodes, where each episode has an agent that performs a single task. In our work, each software agent performs a single task. The agent is moving between a sender and a receiver delivering the request to the receiver. When the agent has reached the receiver, it becomes a launching point for the next agent. The agent "hands over" the message to the successor agent that can proceed to deliver the message. Handing over messages will continue until the message has reached its end destination.

Static environment refers to a surrounding that is unchanged during execution [11]. In these environments, the agents do not need to control the surrounding world while determining the actions. The software agents, in our work, perform actions within a static environment where they move between servers, routers and network components following the protocols. The protocol directs each package through the network and the agent performs the request. However, the agents collect all the information available during its travel. Except for time, it is information about, but not limited to, IP addresses and message digests. The information can easily be expanded with protocols, including other levels than transport and application.

A discrete environment describes the state representation of the environment, as well as, the way time is handled. Moreover, the environment describes perceptions and actions of the agents [11]. A discrete environment can have a finite number of states; it also can have a discrete set of perceptions and actions. The network has a finite number of states but the first request uses an unknown number of IP addresses. When the sensitive message is sent, the number of IP addresses must be known and finite. The actions of the agents follow the messages between the routers; meanwhile the agents record everything that happens. Thus, the agents perceive the environment in which the routers work and record the path between the routes.

Due to the nature of the complex task in the network, we will not use a single agent. In multi-agent systems one single agent does not have all the data required to achieve a task and must cooperate with other agents and, thereby, jointly perform a given task. The multi-agent environment is cooperative in the sense that the agents need to pass messages between the agents. The software agents pass significant information to the meta-level agents, which use the information for further communication and secure message passing.

The software agents follow the actual message passing through network in two steps. The first step is to carry the request from the sender, between the starting point and destination, to follow a marked route. While executing the request, the software agents collect information about the route, and the results from all agents are sent to a meta-level agent. The second step is to carry the sensitive message response through the same path that is collected from the first step avoiding the possible unexpected and eccentric routes. The second step is supervised by the meta-level agents.

Hence, the actions and the results from the software agents are monitored by an administrative meta-agent. This meta-agent only includes those software agents that have successfully completed their tasks. The software agents have to deliver the complete message in an acceptable time to be considered successful. If the software agents cannot deliver the message, the message has to be resent until it is successfully delivered.

5 Meta-Agents

Meta-agents can be used for meta-reasoning, i.e., reasoning about reasoning. Meta-agents can be designed to reason with individual agents, where the goals of the meta-agents are performing reasoning, planning actions and modelling individual agents but also classifying conflicts and resolving these [10]. The main role of the meta-agents is to shift between strategies and make decisions. These strategies can be implemented in a single agent or a group of agents.

Meta-agents constitute a special case of agents. The meta-agents are not pre-defined. Instead they are created from the ground-level agents and their actions. Meta-agents can be applied over software agents in a multi-agent system to keep track of multi-agents and control their behaviour while they move between states in graphs [4]. The meta-agents are built on the agents and used to inspect the behaviour of the agents when reaching a result, where the behaviour is due to the characteristics in the environment. The meta-agents can also perceive the reason for that result and show the status of the static and dynamic characteristics in the environment. The benefit of using meta-agents for multi-agent systems is the ability to find the fastest way between nodes under given circumstances, by taking several different approaches between the start and goal into account [3].

In this paper, we use meta-agents on networks to solve security problems. As in Håkansson and Hartung, (2007:a), meta-agents are developed from the software agents in the multi-agent system, but in our work we used the meta-agents to provide a secure route or path from monitoring the agents following message through the network. The meta-agents have the structure of several ground-level software agents connected together like a chain of actions performed by the software agents. Additionally, the meta-agents can contain information about the performance, i.e., successful/unsuccessful, computation time, information about the layers, i.e., content in the transport layer and application layer. While the software agents are moving around in the network, the meta-agents need to consider the circumstances at a given time and act upon unexpected events. All these events, the software agents need to report to the meta-agents.

Meta-agents can be used as a means to help software agents observing the environment, and evaluate the alternatives, but also to prescribe and schedule actions [10]. The meta-agent is superior to the software agents and is capable of monitoring and controlling the agents, and then evaluating them. Each set of software agents that have performed and accomplished the task is encompassed by the meta-agents.

The meta-agents are intelligent in the sense that they are autonomous, which means that the agents are to some extent capable of deciding what they need to do in order to satisfy their objectives [12]. Furthermore, the meta-agents interact with ground-level software agents and other meta-agents to carry out tasks. They cooperate with the ground-level software agents in the network by sending messages directly to the software agents but also via messages through other meta-agents. The communication with the ground-level agents consists of receiving messages from the software agents and giving commands that the software agents have to execute. The communication with other meta-agents is giving commands to be sent to the ground-level software agents connected to those meta-agents. These meta-agents can be created from earlier message passing and, therefore, stored as secure paths.

The meta-agents in the network act autonomously to some extent. They have to send correct commands to make the software agents to act according to the protocol. However, if the software agents run into a failure, the meta-agents need to decide themselves what they need to do in order to decide the failure recovery. Thus, the meta-agent has to respond to the software agents' environment and if the software agent does not reach its destination, the meta-agent must find the cause for the blockage and then decide how to precede the task.

We use a scheme to dynamically generate meta-agents. The basic agents need to be embedded into the network infrastructure. The meta-agent is generated by the request. It receives the data from each agent as the packets are passed. From the message route through the network, the meta-agent has all the information about all the software agents that have been used while working with the request. Then meta-agent is used for sending the sensitive message by steering the message through the network.

6 Secure Message Passing

Current network and transport security layer solution use traditional mechanisms such as firewalls, routers and proxies to send messages through the network. These solutions are network aware [17]. Web Services operates at the application layer of OSI stack, which means that Web Services need to be protected from the beginning, i.e., starting with the request from the application layer. The goal of SOAP message monitoring is to provide Web Services secure conversation and to secure SOAP message exchange. Security considerations include unauthorized access, unauthorized alteration of messages and man in the middle attacks that have arisen with the introduction of Web Services. These security threats are identified by the WS-I Basic Security Profile. To prevent these threats, a sophisticated and intelligent approach is needed, which will be able to support distributed applications and compliment different security mechanisms. At the same time it is important to keep security solutions simple [17].

It is important to keep in mind that SOAP supports one or more intermediaries. These intermediaries can forward or reroute SOAP messages, based on the information in the SOAP message, the HTTP header or other protocols. To secure SOAP messages, we propose the following approach: monitor SOAP messages during transport by using meta-agents. Note that we assume multi-hop message flow, where one Web Service calls another Web Service to handle one piece of functionality before the result is passed on to yet another Service in the chain.

SOAP message is intercepted with a piece of software, namely ground-level software agents. These software agents will be placed into path of the SOAP message and travel with SOAP message from consumer to the ultimate receiver through intermediaries and transport level routers. Software agents copy the routines automatically when the intermediary inserts a new URL. When one agent has reached its goal, i.e. an IP address, next agent proceeds by taking the message to the next IP address according to the information in the protocol. Because an agent performs a single task and works with the same SOAP message, we built an intelligent meta-agent upon the software agents. To be able to evaluate the message passing we use the superior meta-agents. As meta-agent has the ability to control and collect information from each software agent but also record and report information, these meta-agents can control the SOAP messages path through the network, see Figure 1.

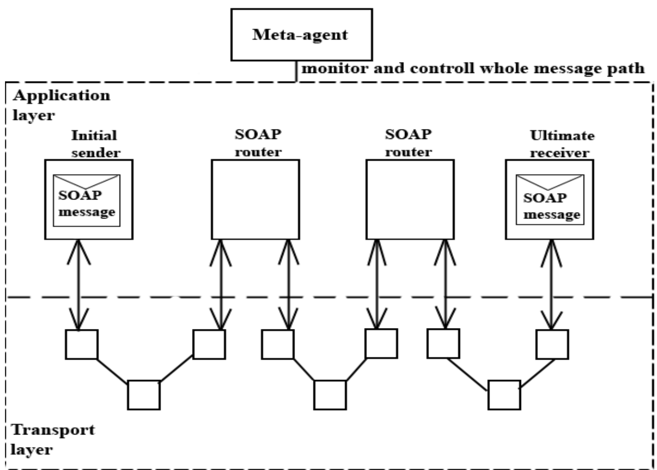


Fig. 1. Architecture of the meta-agents and software agents in the network

When the request is made the second time, the Strict Source Routing must be used on the transport level, because this time the meta-agent have the knowledge of network topology and can provide hop-by-hop specification of the path, which the message should take. Strict Source Routing is an impractical technology; however our meta-agents solve this problem since they have the ability to provide the fastest and shortest way between routers.

It is likely that the response (or return of the message) for a request will require multiple packets. If so, a second meta-agent could be used during the return. When return is made, the meta-agent will correspond with the original meta-agent to perform a validation of the message path. A validation can also be made when each packet is generated and received as a response.

7 Failure Recover

The ground-level agent must stop execution, if the time deviates too much from first execution time. Then the agent has to recognise what the problem is and respond to the problem, correctly. The SOAP message will be discharged and the ground-level agent released from the message. The meta-agent has two options when it operates on transport level. Either the meta-agent reroutes the last SOAP message to another already tried path, that is considered to be safe, or it decide to retry the complete path. Retrying the path means that the failure cannot be too serious and probably run out of time. This failure, probably, arose because of response problems rather than having an attack. An attack requires a new evaluation of the path to find a significantly different path through the network. During message travel on the application level, a meta-agent must stop execution as soon as an execution exception is generated. Execution exceptions can be different SOAP faults that provide information and explanation of why the error occurred.

Conclusively, the meta-agent can create a fault report with detailed information about the kind of fault and place, which router caused the fault both on application and transport level but also time and reason for the fault.

8 Conclusions and Further Work

We have presented some important Web Services security issues emphasized by a study based on security problems in monitoring and auditing Web Services. The result of the study is the integration of software agents and meta-agents into Web Services, which improves WS-security. In this paper, we have presented an approach to use meta-agents on top of software agents in multi-agent systems to avoid security problems.

Meta agents can be considered as a building block that will be used in combination with other Web Services, protocols, and encryption technologies. Implementing meta-agents does not mean that application cannot be attacked, but it is a step forward in making Web Services more secure.

Because of unique ability to monitor, control, collect the information from software agents, and forward the data-rich SOAP messages, the meta-agent will be a central unit that have the control over the whole message path from the initial sender to ultimate receiver through SOAP intermediaries and transport level routers.

References

1. Fasli, M.: On agent technology for e-commerce: trust, security and legal issues. The Knowledge Engineering Review 22(1), 3–35 (2007)
2. Gorodetski, V., Kotenko, I.: The multi-agent systems for computer network security assurance: frameworks and case studies. In (ICAIS 2002). IEEE International Conference on Artificial Intelligence Systems, pp. 297–302 (2002), ISBN: 0-7695-1733-1

3. Håkansson, A., Hartung, R.: Using Meta-Agents for Multi-Agents in Networks. In: Arabnia, H., et al. (eds.) Proceedings of The 2007 International Conference on Artificial Intelligence, ICAI 2007, vol. II, pp. 561–567. CSREA Press, U.S.A (2007a), http://www.world-academy-of-science.org
4. Håkansson, A., Hartung, R.: Calculating optimal decision using Meta-level agents for Multi-Agents in Networks. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 180–188. Springer, Heidelberg (2007)
5. Moradian, E., Håkansson, A.: Possible attacks on XML Web Services (IJCSNS) International Journal of Computer Science and Network Security 6(1B) (2006), Journal ISSN: 1738-7906
6. Pedireddy, T., Vidal, J.M.: A Prototype MultiAgent Network Security System. In: AAMAS 2003 Proceedings of the second international joint conference on Autonomous agents and multiagent systems, Melbourne, Australia, July 14 - 18 (2003)
7. W3C XML Protocol Working Group, http://www.w3.org/TR/
8. Nielsen, H.F., Thatte, S.: Web Services Routing Protocol (WS-Routing), http://msdn2.microsoft.com/en-us/library/ms951249.aspx
9. Adsett, C., Bernardi, A., Liu, S., Spencer, B.: Realising Weak Work Workflow with Declarative Flexible XML Routing in SOAP (DeFleX) (2004)
10. Chelberg, D., Welch, L., Lakshmikumar, A., Gillen, M.: Meta-Reasoning For a Distributed Agent Architecture. In: Proceedings of the South-Eastern Symposium on System Theory, Athens, Ohio, USA, March 18-20, 2001, pp. 377–381 (2001)
11. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Inc., Englewood Cliffs (1995)
12. Wooldridge, M.J.: Introduction To Multi-Agent Systems. John Wiley and Sons Ltd., Chichester (2002)
13. O'Neil, M.: Web Services Security. McGraw-Hill/Osborne, New York (2003)
14. Hartman, B., Flinn, D.J., Beznosov, K., Kawamoto, S.: Mastering Web Services Security. Wiley, Chichester (2003)
15. Peterson, G.: Security Concepts, Challenges, and Design Considerations for Web Services Integration, Cigital, Inc.; Howard Lipson, Software Engineering Institute (2006), https://buildsecurityin.us-cert.gov/daisy/bsi/articles/best-practices/assembly/639.html
16. Demchenko, Y.: Attacks on Web Services and Grids. White paper (2004)
17. Papazoglou, M.: Web Services: Principles and Technology. Pearson Education, Essex England (2008)
18. Rosenberg, J., Remy, D.: Securing Web Services with WS-security (2004), ISBN 0-672-32651-5

A Software Engineering Process for BDI Agents

Aaron Hector, Frans Henskens, and Michael Hannaford

University of Newcastle,
University Drive, Callaghan NSW 2308, Australia
{Aaron.Hector, Frans.Henskens,
Michael.Hannaford}@newcastle.edu.au

Abstract. Software agents provide an increasingly popular approach to software engineering, offering a fundamentally different design technique for software development based around the creation and deployment of autonomous software components. One of the leading models for the development of agent-based software is the Belief-Desire-Intention (BDI) model. In order to effectively develop BDI systems, a software development process is required, similar to those that exist for conventional object-oriented software development. This paper presents NUMAP, a development process for BDI multi-agent systems that covers the entire software development lifecycle, from requirements analysis to implementation and testing.

Keywords: Agent-Oriented Software Engineering, Design Processes, Documentation.

1 Introduction

During the last decade there has been an increasing focus on research into software agents. This field, which encompasses a broad range of approaches, has potential for developing applications related to massively distributed systems such as the Internet [1].

Agent-based software engineering shows great promise for software development due to its high degree of abstraction. It has been argued that software engineering supports increasingly complex systems via increasingly sophisticated techniques for abstraction, and that agents may provide the next advance in this regard [2].

The Belief-Desire-Intention (BDI) model [3] is a popular paradigm for development of agents. It assigns “mental attitudes” to agents. The agent’s behaviour can be specified in terms of these attitudes, providing an intuitive abstraction for the development of agent-based software.

This paper presents NUMAP (Newcastle University Multi-Agent Process), a comprehensive software engineering process for development of software agents, and in particular, BDI agents.

2 Agent Development Processes

One of the primary concerns of software engineering is developing and formalising processes to aid in constructing complex software systems. As such, agent-based

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 132–141, 2008.

software engineering requires a process to guide development and formalise best-practices, just as does any other form of Software Engineering [4]. A number of processes and methodologies have been created to achieve this goal [2, 6, 7]. Some of the more popular processes are described in the following sections.

2.1 Tropos

The Tropos methodology [5] is based on two key ideas. Firstly, the concept of agent is used throughout the development process, from early analysis to implementation. Secondly, unlike many agent development methodologies, Tropos deals with the early requirements analysis process, with the goal of attaining a better understanding of how the environment, agents and users interact. It aims to have the entire development process take place at the knowledge level, rather than introducing notions from traditional software development.

There are five main stages to the Tropos methodology: early requirements, late requirements, architectural design, detailed design, and implementation. Early requirements analysis involves: determining the stakeholders within the system and identifying their intentions; determining the goals of each actor; and determining the dependencies between actors.

The “Late requirements analysis” phase defines the system that is to be developed, and specifies how it will interact with its environment. The system is modelled as an actor, and dependencies with other actors in the organisation are defined [6]. The architectural design phase focuses on the system’s global architecture, defining the subsystems and their dependencies.

Detailed design in Tropos involves the specification of agents’ “micro-level”. Agent properties are defined, including goals, beliefs, capabilities and communication with other agents.

2.2 Gaia

The Gaia methodology [2] defines both the macro-level (societal) and micro-level (individual agent) aspects of agent-based systems. Gaia supports the definition of roles, and their assignment to agents in the system.

Gaia has two distinct phases, analysis and design. Analysis is primarily concerned with high-level understanding of the system to be developed. The system’s organisational structure is defined through roles and relationships.

Design in Gaia involves the creation of three models for the system. An *agent model*, which defines the agent types and their instances; the services provided by each role are modelled in a *service model*, and the communication between agents is defined in an *acquaintance model*. Design takes place at an abstract level, and is not tied to any specific agent model or implementation environment.

An extension to Gaia that takes an organisational approach to agent design has also been proposed [7].

2.3 Prometheus

The Prometheus methodology [4] attempts to cover the full life-cycle of agent development, from specification and definition of the system to implementation and testing.

Prometheus is divided into three main stages: *System Specification, Architectural Design, and Detailed Design.* Components of the specification and design, such as forms and diagrams, are termed artefacts. At each stage, several artefacts are defined, representing the aspects of the system to be developed at the stage, and are stored as structured data.

The artefacts defined at the system specification level include high-level concepts such as scenarios, system goals, functionality descriptors and basic actions, in addition to the system's inputs and outputs. At the architectural design stage, artefacts include agent types, along with an overview of the system's structure and a description of interactions between agents. At the detailed design level, agents are defined in more detail, along with their capabilities, plans, events they handle, and data stored by the agents.

A support tool has also been developed for Prometheus, to assist developers in following the process.

2.4 Agent OPEN

A different approach to methodology design for multi-agent systems is taken by the OPEN Process Framework [8]. OPEN defines a metamodel for creating customised methodologies, based on the individual needs of a project or organisation. OPEN is based upon the principles of situational method engineering, allowing individual, tailored methodologies to be created.

OPEN provides a repository of predefined method fragments and construction guidelines, which can be used to create personalised methodologies as instances of OPEN's methodology metamodel. The construction guidelines assist the user in selecting the method fragments to be used in the methodology.

OPEN defines various types of method fragment. The five main kinds are: Work Units, Producers, Work Products, Languages and Stages. These are described in [8]. These work units are used to proceed through the activities in a customised process, as developers progress through the life-cycle of the process.

While OPEN was originally designed for object-oriented development, it has been expanded to support agent-based software. This extension is referred to as Agent OPEN [9]. In order to extend the OPEN Process Framework to support agent-based development, a number of additional work units were identified by examining existing agent methodologies, such as Tropos, Gaia and Prometheus. A number of new agent-based method fragments were added to the OPEN repository as part of this examination [8].

3 The NUMAP Process

3.1 Aims

NUMAP is a practical design process that guides the development of agent-based systems. It covers all aspects of design, from early requirements through to implementation. NUMAP provides a set of guidelines that define the basic concepts used in each phase of the design.

In designing NUMAP, a number of desirable properties for agent-based design processes were identified. The first such property is that the process utilises existing tools and techniques, in order to take advantage of prior experience with such techniques. One important existing technique that can be used is goal-based requirements analysis [10], an approach which specifies requirements in terms of goals. Due to the proactive, goal-focused nature of multi-agent systems, goal-based requirements analysis techniques are a natural fit.

Another useful property of a design process is its ability to use a number of different implementation environments. There are a variety of different environments available for implementing multi-agent systems, such as JACK [11], Jadex [12] and Swarm [13], and a design process is evidently more useful if it can be used with a number of such environments. However, care must be taken to ensure that the concepts used in design are not generalised too much, in order to ensure a smooth transition between design and implementation.

A support tool is needed to assist with the process. This tool must allow developers to enter data as they progress through the process, print documentation based on this data, and generate code to be used as the basis for implementation.

The ability to tailor the process is also required. In particular, designers should have the ability to select different requirements techniques, low-level design approaches, and varying code generation tools needs to be supported.

3.2 Overview

Rather than taking an abstract approach to defining the design concepts, NUMAP ensures that they have parallels in real-world agent implementation environments. In doing so, NUMAP allows software engineers to produce a design specification that is closer to the actual implementation, and which takes into account the specific requirements of the agent design technique that is being used.

In order to retain flexibility and support for different agent types, NUMAP uses a modular approach, where particular phases of the process can be replaced with a different module in order to support different design approaches.

For example, the current Agent Design and Implementation modules being used with NUMAP are based upon the BDI agent philosophy. The concepts defined during design are closely related to those used in BDI-based [13] implementation environments such as JACK [11], and Jadex [12]. These could, for example, be swapped for Swarm based [13] agent design and implementation modules in order to support this different agent design approach. The NUMAP process has been carefully designed to allow for such module changes without affecting the rest of the process.

Additionally, the design itself is modular. For example, if a decision is made to change agent implementation environments after the design is complete, then only those parts of the system related to the agent implementation module need be changed. The rest of the design remains unchanged.

The key concepts that are defined in the process are goals and agent types. Agent types define the kinds of agents that can be instantiated, in the same way as classes define the objects that can be created in object-oriented software engineering. Goals define the objectives of the agent type.

NUMAP concepts are described by completing a series of forms within a design tool. Each concept has its own distinct form, which lists the attributes that need to be defined to fully describe that concept.

The process itself is divided into five distinct phases: Requirements Elicitation, Analysis, Organizational Design, Agent Design, and Implementation.

Each of these phases can be altered or replaced in order to support a variety of design approaches. In particular, the Requirements phase can be altered to support different requirements elicitation methods, and the Agent Design and Implementation phases can be replaced in order to support different agent design techniques.

The remainder of this section provides an outline of the NUMAP process and each of its phases.

3.3 Process Outline

The primary activity within each phase of NUMAP is specification of the concepts required for the system model. These concepts are defined by filling in the required information for each concept, usually by entering the information into the support tool. Each phase of the process has its own distinct phase model, and the collection of these creates the overall system model. As development progresses, the models from the previous phases are used to assist in constructing the model for the current phase. The NUMAP tool assists where possible in automating the transition between phases. The model for each phase consists of the elements defined (via forms) for that phase, and the relationships between them.

Diagrams can be generated from the system model to provide a visual representation of the model. They can also be used as an alternative mechanism for specifying details of the system, or for making minor alterations to the design.

Other processes, such as Gaia [2] omit specific design details in order to preserve generality. NUMAP uses a modular approach to allow for a more detailed design process, without restricting the entire design process to that approach. This combines the benefits of a more detailed design process with the advantages of generality.

Different modules may be used for different agents within the same system, in order to allow various forms of agents to co-exist within the same system. For example, some agents in a system may be BDI-based, while others may be Swarm [13] based.

NUMAP's support tool provides a form-based GUI for entering the data for each phase of the process. Additionally, it enforces data integrity by checking content as it is provided, assists with transitioning between each of the phases of the process, and guides the developer through each step of the process.

The support tool assists with validating the correctness of the system model; for example, ensuring the pre-conditions and post-conditions for goals exist, ensuring that two goals are not preconditions of each other, and ensuring all goals have plans associated with them.

The tool also assists with generating diagrams from forms. These diagrams help visualise the system model, and can also be edited to directly update data in the system model.

3.4 Process Phases

An overview of each of the phases of the NUMAP process is provided below. The objectives of the phase are described, and the main concepts for each phase are explained.

Requirements

The Requirements phase uses a goal-based requirements analysis method to describe the requirements for the system. The overall goals of the system are defined, along with an overview of how these goals will be achieved.

Any goal-based requirements method may be used at this stage. The requirements method currently being used with NUMAP is GBRAM [10], however this could be replaced with another approach by using a different Requirements module. Depending on the approach used, the specific concepts defined in this phase would differ.

The GBRAM-based requirements stage requires three distinct sets of data to be defined, as specified in the GBRAM process. Firstly, the *agents and stakeholders* within the system need to be defined. Next, the *goals* for the system are defined. The GBRAM requirements elicitation method provides a number of strategies for defining these goals. These strategies are followed to create a detailed model of the high-level goal hierarchy of the system. Lastly, the operationalisations, or *actions*, of each goal are defined. These provide a basic description for how each goal may be achieved.

Analysis

The analysis process is concerned with creating an abstract model of the system based upon the results of the requirements phase.

There are two primary goals of this phase. Firstly, the outputs of the requirements phase are mapped into a standard format. This is necessary due to the modular nature of the requirements process. Secondly, these design elements are expanded with more detail, and a number of additional elements are defined.

The first of these additional elements is the system's *environment*. In order to define this, relevant elements that are external to the system are identified. The *sensors* that are used to sense environmental elements, and the *effectors* which the agent uses to effect change upon the environment, are also defined.

Also defined are agent *services*. These define the functionality that an agent type makes available to its peers. Services may be grouped into *roles*. Roles are cohesive groupings of agent services that can be applied to agent types that share functionality. Each agent type may have one or more roles, defining the services that the agent provides, and also the services that it uses.

Preliminary *organisations*, used for grouping agent types into subsystems, are also defined at this phase.

At the conclusion of this phase, all of the above essential elements are defined, ready for the organisational design phase.

Organisational Design

The next stage of NUMAP is organisational design. This phase is concerned with defining the system at an inter-agent level. Concepts that were defined in the analysis phase are refined, and the interactions between agent types are defined. The internal functionality for each agent is not defined at this stage.

An organisational-level description of *agent types* is created, expanding on the more simple types that were identified during the previous two phases. This forms the final list of agent types that will be used in implementation.

High-level goals for each of these agent types are also defined at this stage. These are used for documentation purposes only, and provide a brief description of the functionality of the agent.

In order for agents to communicate, *message types* need to be defined. To assist with standards compliance, each message type is assigned a performative, based on those defined in the FIPA Communicative Act Library Specification [14].

Agent communication is further structured by *agent interaction protocols*. These define a formal method for the exchange of messages between two parties. The interaction protocols that are defined may be based on existing FIPA interaction protocol standards, for example Request Interaction Protocol [15], Query Interaction Protocol [16], or Contract Net Interaction Protocol [17].

Higher-level agent functionality is defined by refining the *service* and *role* descriptions from the analysis phase. The particular message types and interaction protocols used by each service are defined.

Organisations are also carried over from the analysis phase, and can be used as a basis for defining more complex structures for the system. For example, extensions to the organisational design phase to allow complex behaviour such as rules, norms and agent admission criteria are currently being explored.

Agent Design

The agent design phase involves defining the internal behaviour of each agent. Depending on the technique used for creating agents, different agent design modules may be used. Currently, a BDI module is being used for creating agents.

The primary concept to be modelled within the agent design phase is the *agent type*. Within the agent design phase, this defines the internal aspects of the agent.

All agent types have *goals* that define their proactive behaviour. An agent will generally have several goals that it simultaneously pursues. An agent attempts to satisfy its goals via *plans*. Each plan defines an action, or set of actions an agent may perform in pursuit of a goal or in response to a reaction.

Agent *reactions* are instant reactions to some event. Unlike goals, these do not run constantly, rather they model the reactive behaviour of the agent type. *Events* are triggered by agents in response to some situation that requires immediate attention. They can be generated by a received message, a belief state change, or they can be manually generated by a plan.

Beliefs define what an agent “thinks it knows”. They are the agent’s information about its environment and about other agents within the system. That is, beliefs are the agent’s symbolic representation of its surroundings.

The agent's *environment* defines the elements external to the system with which the agent will interact. This includes environmental elements that will be sensed, as well as elements that will be changed by the agent. *Sensors* and *effectors* are the agent's mechanism for interacting with its environment. Sensors are used to receive inputs from the environment, while *effectors* are used by the agent to effect change upon its environment.

An agent uses *plan selection rules* to select which plan is used to attempt to satisfy a goal. Similarly, when an agent must use a service, it may have to select from a number of agents that provide that service. *Delegation selection rules* are used to make that selection, and are defined in this phase.

Agent *capabilities* can be used to create a grouping of beliefs, goals, reactions and plans that can be used by any agent. Specific agent functionality can be defined in the capability, and inherited by an agent type, thus encouraging software reuse.

Implementation

Implementation focuses on writing code for the finished design, which will be run in an agent runtime environment. There are a number of such environments, including JACK [11], and Jadex [12].

The close mapping between the concepts defined in the design phases and the actual implementation environments allows for agents to be readily implemented from the design specification.

The NUMAP support tool may be used to generate code for a specific agent implementation platform. It can generate the basic code for an agent, by producing a template, and defining the agent's overall structure. The programmer uses this as a basis for implementing the agent.

3.5 Evaluation

In order to provide a qualitative evaluation of NUMAP, a multi-agent Marketplace Simulation project has been created. This system features a number of interacting Customers, Retailers and Supplier agents involved in buying and selling of goods. The system was initially implemented, using Jadex, in an ad-hoc manner, without the assistance of a design process or support tools. Subsequently, a system with the same requirements was implemented with the assistance of NUMAP.

Development with the assistance was found to be much less error-prone, particularly in defining Agent Definition Files. The final design created with the assistance of NUMAP and its support tool was found to have more consistent interaction between agents, with communications being more structured. Code reuse was increased, with an increase in the number of plans shared between agents, and an increase in shared code for communications and service handling. Documentation created with NUMAP was found to be comprehensive, and greatly assisted with debugging. Total development time was also considerably shorter when using the NUMAP process.

A comparison was made with a system developed using Prometheus. This system was developed using the process outlined in [4] and the Prometheus Design Tool (PDT) 2.5f [18]. NUMAP shows a number of advantages in this comparison. The use of an established requirements analysis technique (GBRAM) allows experience with this technique to be utilised during the requirements elicitation phase of development.

The modular nature of NUMAP provides the flexibility of changing agent design approaches late in the development life-cycle, and allows for easy mixing of different styles of agents within the same system. Additionally, NUMAP presently allows for code generation in Jadex and has preliminary support for JACK, in contrast to PDT 2.5f, which only supports JACK.

Further evaluation will be undertaken in future, to evaluate NUMAP against a number of other processes. A larger-scale problem, namely the implementation of an existing distributed e-commerce system will be used for this evaluation and will be the subject of a future report.

4 Conclusion

NUMAP is a new process with accompanying support tools that provides a modular approach to developing agent-based software, from requirements elicitation to actual implementation. The support tool currently has modules for the GBRAM requirements process [10] and the Jadex agent environment [12]. After evaluation of the current modules is complete, support will be added for additional requirements elicitation techniques and agent implementation environments.

A number of extensions to the process are under development, including an extended organisational design phase, based upon recent work in organisational theory for agent systems [19]. This extension handles enforcement of rules and norms within agent organisations and agent admission criteria.

There are plans for the NUMAP support tool to be expanded to provide additional support for multiple iterations through the process. NUMAP's support for defining links between elements in different phases will assist with this expansion.

References

1. Nwana, H.S., Ndumu, D.T., Lee, L.C., Collis, J.C.: ZEUS: A Toolkit for Building Distributed Multi-Agent Systems. Applied Artificial Intelligence 13, 129–186 (1999)
2. Wooldridge, M., Jennings, N.R., Kinny, D.: The Gaia Methodology for Agent-Oriented Analysis and Design. Autonomous Agents and Multi-Agent Systems 3, 285–312 (2000)
3. Rao, A.S., Georgeff, M.P.: BDI-agents: from theory to practice. In: Proceedings of the First Intl. Conference on Multiagent Systems (1995)
4. Padgham, L., Winikoff, M.: Developing Intelligent Agent Systems: A Practical Guide. John Wiley & Sons, Chichester (2004)
5. Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F.: Tropos: An Agent-Oriented Software Development Methodology. Autonomous Agents and Multi-Agent Sytems 8, 203–236 (2004)
6. Giunchiglia, F., Mylopoulos, J., Perini, A.: The Tropos Software Development Methodology: Processes, Models and Diagrams. In: Giunchiglia, F., Odell, J.J., Weiss, G. (eds.) AOSE 2002. LNCS, vol. 2585, pp. 162–173. Springer, Heidelberg (2003)
7. Zambonelli, F., Jennings, N.R., Wooldridge, M.: Developing Multiagent Systems: The Gaia Methodology. ACM Transactions on Software Engineering and Methodology 12, 317–370 (2003)

8. Henderson-Sellers, B.: Creating a Comprehensive Agent-Oriented Methodology. In: Henderson-Sellers, B., Giorgini, P. (eds.) Agent-Oriented Methodologies, pp. 368–386. Idea Group Publishing, Hershey, PA (2005)
9. Henderson-Sellers, B., Giorgini, P., Bresciani, P.: Enhancing Agent OPEN with concepts used in the Tropos methodology. In: Omicini, A., Petta, P., Pitt, J. (eds.) ESAW 2003. LNCS (LNAI), vol. 3071, Springer, Heidelberg (2004)
10. Anton, A.I.: Goal-Based Requirements Analysis. In: ICRE 1996, IEEE, Los Alamitos (1996)
11. Howden, N., Ronnquist, R., Hodgson, A., Lucas, A.: JACK Intelligent Agents – Summary of an Agent Infrastructure. In: The 5th International Conference on Autonomous Agents (2001)
12. Pokahr, A., Braubach, L., Lamersdorf, W.: Jadex: Implementing a BDI-Infrastructure for JADE Agents. EXP - In Search of Innovation 3, 76–85 (2003)
13. Daniels, M.: Integrating Simulation Technologies With Swarm. In: Macal, C.M., Sallach, D. (eds.) Workshop on Agent Simulation: Applications, Models and Tools, Argonne National Laboratory, Argonne, Illinois (2000)
14. FIPA: FIPA Communicative Act Library Specification (2002)
15. FIPA: FIPA Request Interaction Protocol Specification (2002)
16. FIPA: FIPA Query Interaction Protocol Specification (2002)
17. FIPA: FIPA Contract Net Interaction Protocol Specification (2002)
18. RMIT Intelligent Agents Group: Prometheus Design Tool (2007), `http://www.cs.rmit.edu.au/agents/pdt/`
19. Dignum, V., Weigand, H.: Toward an Organization-Oriented Design Methodology for Agent Societies. In: Plekhanova, V. (ed.) Intelligent Agent Software Engineering, pp. 191–212. Idea Group (2003)

Design Pattern Based Development Methodology and Support Tool for Multi Agent System*

Hyunsang Youn and Eunseok Lee

School of Information and Communication Engineering Sungkyunkwan University
300 Chunchun Jangahn Suwon, 400-746, Korea
{wizehack, eslee}@ece.skku.ac.kr

Abstract. A multi-agent system is a system composed of several software agents, collectively capable of reaching goals that are difficult to achieve by an individual agent or monolithic system. Many useful methodologies have been proposed to develop multi-agent systems. However, existing development methodologies for agent based system are not suitable for inexperienced developers. In this paper, we propose a design pattern based agent development methodology to support the inexperienced developer. Specifically, the proposed methodology covers the overall development process from requirement analysis to system deployment. We offer a CASE tool to support the methodology. We implement an E-commerce system development example to evaluate our methodology.

1 Introduction

The development of agent-based systems such as Gaia [1], CARMEN [2], offers a new and exciting paradigm for production of sophisticated programs in dynamic and open environments. Agent technology is increasingly being used to construct intelligent software and smart control modules. Useful agent development environments include FIPA-OS [3], JADE [5]. They simplify and facilitate the implementation of agent based systems. They comply with FIPA Specifications [17] and offer a Java library of classes that programmers can use to develop their agent. However, MAS developers should perform requirement analysis, agent extraction from the requirements and design before the implementation using JADE or FIPA-OS. There is a growing need for a tool that supports earlier stages of the system development life cyle (SDLC) to develop agent based systems.

Multi-agent Systems Engineering (MaSE) methodology [4] is a complete methodology for developing heterogeneous multi-agent systems. MaSE covers the complete lifecycle of the system, from the analysis to the design utilizing a number of graphical tools based on UML. The models are transformed into diagrams in order to describe agents, their communication, and the internal structure of each agent. MaSE is

* This work was supported in parts by ITRC IITA-2007-(C1090-0701-0046, Basic Research Program of the Korea Science &Engineering Foundation, and the Post-BK21 Project. Corresponding author: Eunseok Lee.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 142–151, 2008.

supported by a software engineering tool called AgentTool [11]. AgentTool allows the designer to formally specify all MaSE models. It also supports automated design transformations and automatic verification of inter-agent communication.

Tropos[10] is agent based system development methodology using the mentalistic approach. This is founded on two key features: (i) the notions of agent, goal, plan and various other knowledge level concepts are fundamental primitives used uniformly throughout the SDLC; and (ii) a crucial role is assigned to requirements analysis and specification when the system-to-be is analyzed with respect to its intended environment.

The purpose of INGENIAS [9] is the definition of a methodology for the development of MAS. This methodology is based on the definition of a set of meta-models that describe the elements that form MAS from several viewpoints, and that allows the definition of a specification language for MAS. The methodology of INGENIAS has five viewpoints: agent (definition, control and management of agent mental state), interactions, organization, environment, and goals/tasks. INGENIAS is supported by an integrated set of tools, the INGENIAS Development Kit. These tools include an editor to create and modify MAS models, and a set of modules for code generation and verification of properties from these models.

These three methodologies are difficult for inexperienced MAS developers. They may face various problems in the design stage because of their less skill and experience. This leads to inconsistent design and poor reliability of the resulting agent-based system. A design pattern is a general repeatable solution to a commonly-occurring problem in software design. Design patterns are recognized to be important in software development. Design patterns are not only a finished design phase but also can be refined directly into code; it is a description or template for how to solve a problem that can be used in many different situations. Much research has been done in the design pattern domain [6], [7], [8], [12]. Object-oriented design patterns typically show relationships and interactions between classes or objects, without specifying the final application classes or objects that are involved. The design patterns support the inexperienced designer. Jorge et al provides useful design patterns that have features of GoF pattern for multi-agent architecture [9]. A design pattern is a useful approach for agent design. In this paper, we propose an agent based system development methodology and support toolkit based on design patterns (Practically Useful agent Release Environment: PURE).

Massimo Cossentino adds pattern reuse process in their proposed methodology PASSI [14]. The patterns can be applied to an existing agent or used to produce a new one with the support of MAS. PASSI is a step-by-step requirement-to-code methodology for developing MAS. Additionally, Massimo Cossentino provides a support toolkit[13]. PASSI methodology does not consider the elicitation, analysis and specification of requirements, so it is ambiguous to elicit requirements and define agent from requirements.

In this paper, we propose the PURE methodology that covers the overall process, from requirement elicitation and analysis to system integration, for MAS development. In this methodology MAS analysts can define system requirements and constraints by using UML and designers can identify suitable design patterns by using the score metrics technique proposed in this paper. Specifically, the PURE toolkit automatically generates JADE based source code.

This paper is organized as follows. Section 2 presents the proposed methodology. In section 3, explains the experimental results of proposed methodology and support toolkit. In section 4, we draw conclusions about our methodology and tool.

2 Proposed Methodology and Support Tool

In this section, we explain the PURE process, the method for design pattern selection and support toolkit. PURE methodology covers the overall process for MAS development and the PURE toolkit handles requirement definition, design, and implementation. Specifically, design patterns support cost effective design and implementation. The CASE tool offers auto-generation of JADE and JAVA based source code.

2.1 Development Process

In order to represent PURE methodology in a more understandable fashion, we apply PURE to an E-commerce system [15]. Figure 1 shows the overall process to develop MAS based on PURE. The PURE process consists of five phases (Requirement Definition, Design Agent Community, Each Agent Development, Integration, Evolution). Each phase has one or more steps. This figure shows all of the phases and steps to develop MAS using PURE.

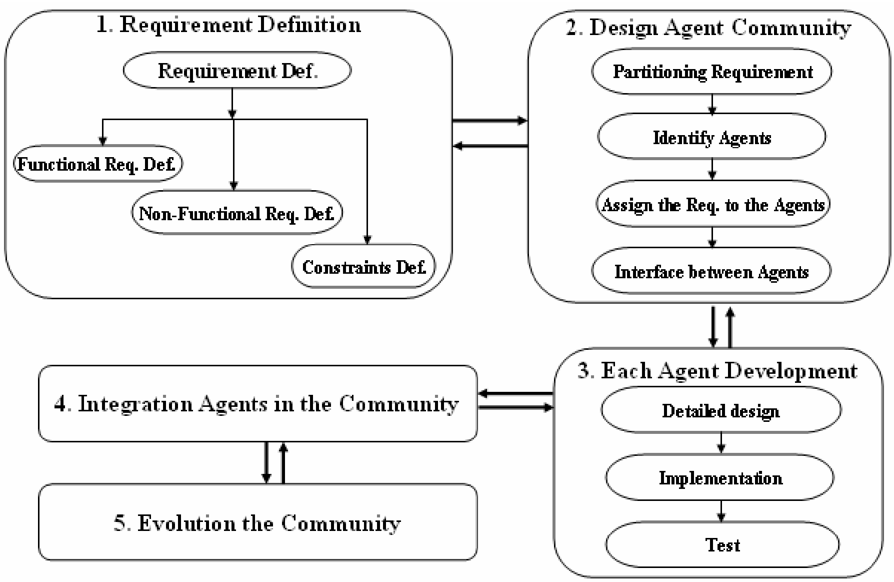


Fig. 1. Overall Process for PURE Methodology

1) Requirement Definition Phase

In this phase the requirement analyst elicits customer requirements. The analyst captures the case to be used by the actor. The Use Case diagram is generally used to capture customer requirements. To define functional and non-functional requirements

Analyst identifies system activities on a case by case basis. The Activity diagram is useful to represent system actions. Analyst defines the requirements from system activities. In the last step (Constraint Definition), UML SPT profile is used [16]. The main aims of the UML profile are to identify the requirements for system performance and methods to model physical time, timing specifications, timing services, logical and physical resources, concurrency and scheduling, software and hardware infrastructure, and their mapping. The profile provides an extension of the UML standard with specific modeling elements and their semantics. Figure 2 shows the activities to define system requirements using the PURE CASE Tool. The Use Case diagram is used to capture customer requirements and the Activity diagram is used to identify system requirements and constraints.

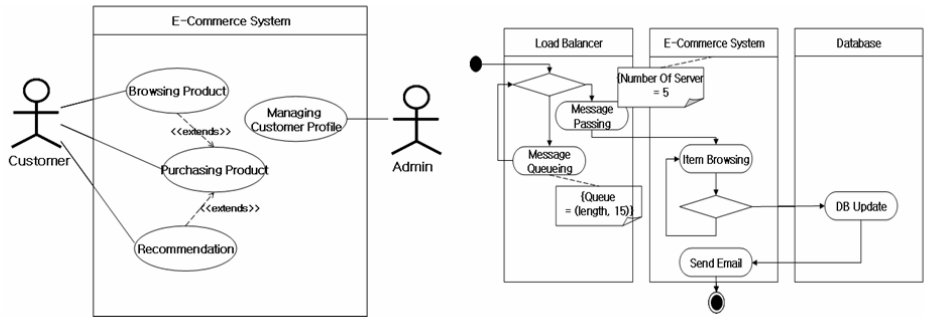


Fig. 2. Defining system requirements

2) Design Agent Community

Requirement analyst and system designer identify agents through grouping system requirements and assigning the requirements group to agents. The system designer then identifies the role of each agent from the role diagram notated by the UML sequence diagram. In this step, the designer describes communication protocol and communication language. Next the designer identifies ontology and knowledge. Finally, the system designer validates the multi-agent architecture using the requirement analyst.

3) Each Agent Development

In this phase the designer constructs the inner structure of each agent and describes their actions. The UML class diagram is used to perform this activity. In the implementation step the PURE toolkit generates JAVA based code. The code may be refined in this step. In order to test the agent the developer validates the input message, output message and state transition of the agent. JADE offers the introspect agent and the sniffer agent. The introspect agent is a tool provided by JADE. The tool offers the function that captures the state transition of agent. The sniffer agent is also tool that is a viewer to capture communication sequence of agents. This tool is useful for both black box and white box testing.

4) Integration Agent in the Community/Evolution the Community

This phase deploys the developed MAS. The developer integrates each agent on the JADE platform. Then, the developer operates the MAS for verification and validation.

2.2 Applying Design Pattern

PURE offers three types of design pattern. Table 1 shows the design pattern Taxonomy for MAS design. Architectural Patterns are used to design the abstract architecture that represents the multi-agent community. In the Design Agent Community Phase, the Software architect defines the multi-agent community to represent their intent. The roles and tasks of each agent are clearly identified using architectural patterns. Protocol Pattern defines the communication protocol between Agents in the step. FIPA proposes an interaction protocol for the standard interoperability of each agent [17]. In each agent development phase, an agent is designed using object oriented design methodology. The developer can design an agent by using a traditional object oriented design pattern such as the GoF pattern.

Table 1. Design Pattern Taxonomy for MAS design

Classifications	Description	Selection Method
Architectural Patterns	- *Two types of patterns(mobility, sociality)* - *Using UML activity diagram and Class diagram*	Score Metrics
Protocol Patterns	- *FIPA Communication Protocol* - *Using UML message sequence diagram*	Ref. FIPA Spec.
Structural Patterns	- *23 GoF Patterns* - *Using UML Class Diagram*	Ref. Documentation

Selection of a design pattern is difficult for inexperienced MAS designers. One of the contributions in this paper is the design pattern selection process. FIPA Communication Protocol Specification provides sufficient information to select suitable protocol patterns. There are also many literatures to apply GoF pattern. For example, there is less published research applied to architectural patterns. So we propose a scoring metric technique to recommend a suitable architectural pattern to the designer. Figure 3 shows the design pattern repository of the PURE toolkit. The function of the PURE CASE tool is to support architectural pattern selection using score metrics, protocol pattern selection using FIPA Spec. and structural pattern selection using GoF pattern document.

We propose the score metric method to identify a suitable architectural pattern. This method calculates the similarity between the architectural pattern and score metrics. The score metrics have characteristics (e.g. Use Legacy System, Internal Data Protection, Distributed Processing, Mediating Communication,) and score (e.g. strong positive, positive, negative). We represent the characteristic and score as $M \times 3$ matrix (M is the metric about the characteristic, and 3 is the score). The similarity calculation method is as follows:

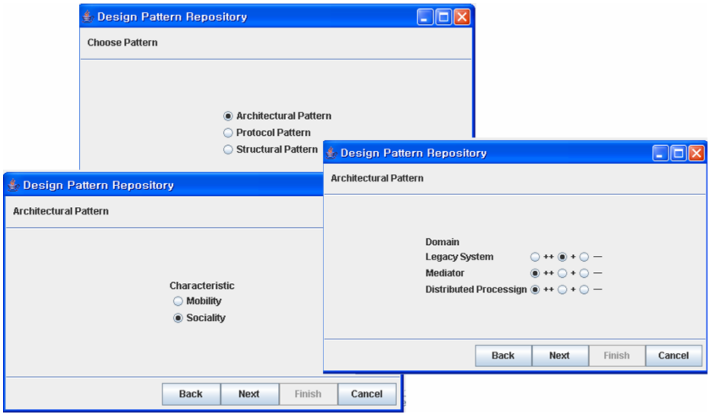


Fig. 3. Design Pattern Selection

Let us assume that: *K* is the design pattern matrix, *P* is the user's selection using the design pattern repository of PURE toolkit.

```
while(i≠j){
      count:=count+1;
      if(count=MAX_Count){
            K:=P⋂M;
            Similarity(K⋂P,K⋂M);
      }
      i:=i+1;
      j:=j+1;
}
```

`i` and `j` are the list of metrics of `K` and `P`. `count` is the number of the difference of the characteristic. We define *MAX_Count* as 0. The similarity between two matrices is calculated as follows:

Let us assume that:

$$A = K \cap P, B = K \cap M$$

First, we can calculate the distance C between A and B as follows.

$$C = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ . & . & . \\ A_{i1} & A_{i2} & A_{i3} \end{bmatrix} - \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ . & . & . \\ B_{i1} & B_{i2} & B_{i3} \end{bmatrix}$$

If the score metrics of A and B are the same, C will be zero. To derive the similarity, we calculate the sum of absolute C_{ij} and divide it by 2 as follows:

$$similarity = \sum_{i=1}^{N} \begin{bmatrix} \frac{\sum_{j=1}^{3} |C_{1j}|}{2} \\ \frac{\sum_{j=1}^{3} |C_{2j}|}{2} \\ \cdots \\ \frac{\sum_{j=1}^{3} |C_{ij}|}{2} \end{bmatrix}, \ \therefore similarity = \sum_{i=0}^{N} \left(\frac{\sum_{j=1}^{3} |C_{ij}|}{2} \right)$$

If the similarity is zero, the PURE CASE tool recommends the pattern ***P***. In this tool, we define the critical similarity of ***P*** as 0.2.

The proposed method to select the structural and protocol patterns is as follows.

First, the designer selects the appropriate design patterns to identify the design intent and problem for the design. The intent and problems have been written on the GoF specification and FIPA communication protocol specification. Second, the developer describes them. Last, the developer studies the description of the design patterns using GoF documentation or FIPA interaction protocol specification, and then the designer compares the described common problem and description of design patterns. Through this process, developers can select a suitable design pattern to solve the problem and to represent the design intent.

2.3 Source Code Generation

PURE toolkit generates JADE platform based source code in the design phase to minimize the implementation effort required. Generally CASE tools provide skeletons of the classes and methods from class diagrams. PURE toolkits not only generate the skeletons but also JAVA based algorithmic code. Figure 4 shows the structure of an agent based on JADE. In JADE, the agent has one agent class and one or more behavior classes. An agent class has a setup method to initiate the agent, and the behavior class has action(), done() method to start and finish an operation. There are several kinds of behavior in JADE such as composite behavior, and simple behavior. Each behavior inherits the Jade.Core.Behavior class. The Task class in figure 4 is the implementation pattern of PURE toolkits. This shows the structure of an agent based on JADE. In the Design Agent Community phase the system designer can describe the core algorithm using JAVA language. The algorithm is inserted in the Task class. Each class, agent, behavior and task, is related by an association.

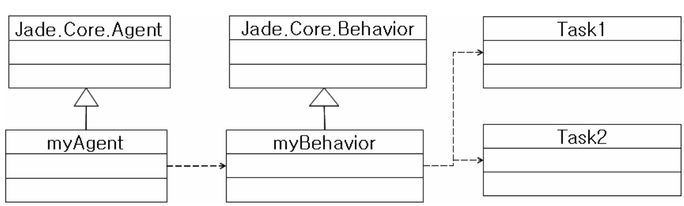


Fig. 4. JADE structure and Code Template

The PURE toolkit offers the identification of the role of each class using the right–hand side button of the mouse. The CASE tool provides the identification of the agent class and all kinds of the behavior class of JADE (Fig 5).

PURE handles the generation of source code XML Metadata Interchange Format (XMI) transformation and code insertion. XMI supports easy interchange of metadata between modeling languages. So the UML model of the PURE methodology is transformed into XML. The PURE CASE tool generates JAVA skeleton code from XML. Our CASE tool also creates the Task class inserted in the JAVA based algorithmic code.

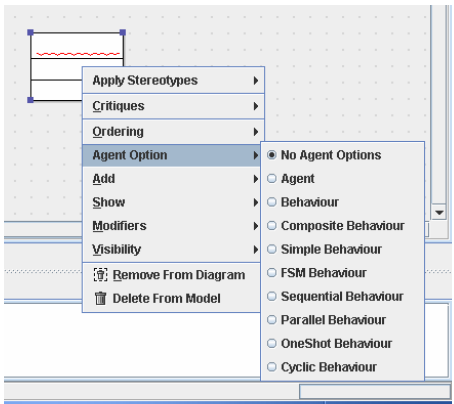


Fig. 5. Process Phase Manager

3 Experimental Results

In this section, we report the results of a practical implementation based on the PURE methodology. We implement an E-commerce system to validate the usefulness of the PURE methodology,. The agents to be configured will be identified and the suitable design pattern for the E-commerce system design will be selected in section 3. We will compare the code generation rate of the agents from the system design. The scenario is briefly described as follows.

The User Agent connects the E-commerce system to browse and purchase products. A Load Balancing agent decentralizes the workload that is assigned to the E-commerce system by many users. The E-commerce agent is running on the system. The E-commerce agent displays and recommends products by communicating with a Recommender agent.

Each agent of this scenario has been implemented using the PURE toolkit. Table 2 summarizes the results of this experiment. This system is composed of 2012 lines of code subcontracted to 4 agents. The auto-generation rate of the required is 100%. The rate of required method is 87% and the generated line of the code is 58%. This result shows our methodology and toolkit to be very useful in saving development effort.

Table 2. Design Pattern Taxonomy for MAS design

Agent	Type of Class	Implemented Result			Generation Rate		
		NOC	NOM	LOC	NOC	NOM	LOC
User Agent	Agent	1	2	22	100%	100%	63%
	Behavior	2	2	98	100%	100%	84%
	Task	8	21	532	100%	67%	21%
Load Balancing agent	Agent	1	2	27	100%	100%	54%
	Behavior	2	2	72	100%	100%	77%
	Task	5	22	219	100%	72%	31%
E-Commerce agent	Agent	1	2	28	100%	100%	70%
	Behavior	4	2	127	100%	100%	81%
	Task	12	71	474	100%	53%	26%
Recommender agent	Agent	1	2	29	100%	100%	67%
	Behavior	2	2	79	100%	100%	78%
	Task	11	76	413	100%	55%	39%

NOC: Number of Class, NOM: Number of Method, LOC: Line of Code

4 Conclusion

We proposed the MAS development methodology and support toolkit. We developed an E-commerce system based on the proposed methodology. Our methodology supported agent extraction from requirement analysis, and optimized design based on selection of an appropriate design pattern. Through the design pattern, we may solve a design problem that generally occurs at the design stage. Finally, JAVA based code was generated by our toolkit. The developers can refine the generated code and run their agent on a JADE platform.

References

1. GAIA Partners: Generic Architecture for Information Availability. Technical report (March 2000), `http://www.infowin.org/ACTS/RUS/PROJECTS/FINAL-REPORTS/fr-221.pdf`
2. Bellavista, P., Corradi, A., Montanari, R., Stefanelli, C.: Context-aware Middleware for Resource Management in the Wireless Internet. IEEE Trans. on IEEE Computer Society 29(12) (December 2003)
3. FIPA-OS Developers's Guide, `http://fipa-os.sourceforge.net/docs/Developers_Guide.pdf`
4. DeLoach, S.A., Wood, M.F., Sparkman, C.H.: Multiagent Systems Engineering. The International Journal of Software Engineering and Knowledge Engineering 11(3) (June 2001)
5. Caire, G.: Jade Tutorial Jade Programming For Beginners (December 04, 2003)

6. Weiss, M.: Pattern-Driven Design of Agent Systems: Approach and Case Study. In: Eder, J., Missikoff, M. (eds.) CAiSE 2003. LNCS, vol. 2681, pp. 711–723. Springer, Heidelberg (2003)
7. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns:Elements of Reusable Object-Oriented Software. Addison-Wesley, Reading (1995)
8. Aridor, Y., Lange, D.B.: Agent Design Patterns: Elements of Agent Application Design. In: Proc. of the Second International Conference on Autonomous Agents, Minneapolis, May 1998, pp. 108–115 (1998)
9. Gómez-Sanz, J.J., Fuentes, R., Gomez-Sanz, J.J., Fuentes, R.: Agent Oriented Software Engineering with INGENIAS. In: Fourth Iberoamerican Workshop on Multi-Agent Systems (2002)
10. Giunchiglia, F., Mylopoulos, J., Perini, A.: The tropos software development methodology: processes, models and diagrams. In: Proceedings of the first international joint conference on Autonomous agents and multiagent systems (2002)
11. DeLoach, S.A., Wood, M.: Developing Multiagent Systems with agentTool. In: Castelfranchi, C., Lespérance, Y. (eds.) ATAL 2000. LNCS (LNAI), vol. 1986, Springer, Heidelberg (2001)
12. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional Computing Series, Reading (1994)
13. Cossentino, M., Potts, C.: A case tool supported methodology for the design of multi-agent systems. In: SERP 2002, pp. 295–306 (2002)
14. Cossentino, M., Luca, S., Antonio, C.: Patterns Reuse in the PASSI Methodology. In: Omicini, A., Petta, P., Pitt, J. (eds.) ESAW 2003. LNCS (LNAI), vol. 3071, pp. 294–310. Springer, Heidelberg (2004)
15. Youn, H., Jang, S., Lee, E.: Deriving Queing Network Model for UML for Software Performance Prediction. In: SERA 2007, August 2007, pp. 125–131 (2007)
16. OMG: UML Profle for Schedulability, Performance, and Time Specification (January 2005)
17. Foundation for Intelligent Physical Agents, http://www.fipa.org/repository/ips.php3

Moving Object Detection and Classification Using Neural Network

M. Ali Akber Dewan, M. Julius Hossain, and Oksam Chae*

Department of Computer Engineering, Kyung Hee University,
1 Seochun-ri, Kiheung-eup, Yongin-si, Kyunggi-do, South Korea, 446-701
dewankhu@gmail.com, mdjulius@yahoo.com, oschae@khu.ac.kr

Abstract. Moving object detection and classification is an essential and emerging research issue in video surveillance, mobile robot navigation and intelligent home networking using distributed agents. In this paper, we present a new approach for automatic detection and classification of moving objects in a video sequence. Detection of moving edges does not require background; only three most recent consecutive frames are utilized. We employ a novel edge segment based approach along with an efficient edge-matching algorithm based on integer distance transformation, which is efficient considering both accuracy and time together. Being independent of background, the proposed method is faster and adaptive to the change of environment. Detected moving edges are utilized to classify moving object by using neural network. Experimental results, presented in this paper demonstrate the robustness of proposed method.

Keywords: Video surveillance, vision agent, motion detection, neural network.

1 Introduction

Moving object detection and classification is an important research issue in computer vision due to its applicability in diverse disciplines including automatic video monitoring, vision agent (VA) based robot navigation and multi-agent based intelligent services for home networks [1, 3]. In vision based robot navigation system, VA may be installed on the body of robot or in a fixed position outside with wireless connection with robot [2, 3]. VA installed on the body of a robot faces difficulties to recognize potential events in the environment due to the movement of viewing point while its installation in a separate fixed position is able to detect dynamic events more accurately and faster due to the availability of a single viewing point. In distributed vision based robot navigation system, each VA independently observes potential events in the environment and provides information to the robots through wireless connection [3] (Fig. 1). Multi-agent based home networking [4], intelligent monitoring, and robot navigation systems primarily engage VAs for detection and classification of moving objects from the fixed viewing point and perform application oriented higher level operations such as tracking, event detection and management, collision detection and so on.

* Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 152–161, 2008.

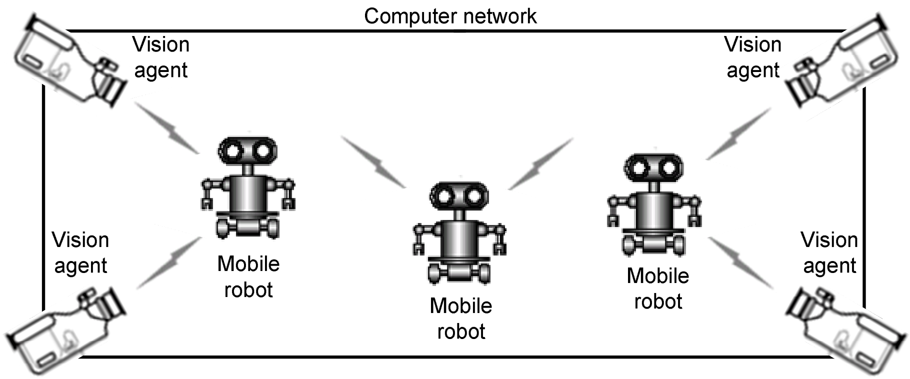


Fig. 1. Distributed vision agent for robot navigation

In this paper, we present an edge segment based approach for moving object detection and classification which can be successfully utilized in vision based monitoring and navigation systems to achieve better accuracy with considerable complexities. Extensive survey on moving object detection can be found in [1]. The most common approaches are based on background subtraction where background modeling is a necessary and unavoidable part for moving object detection. However, background modeling techniques are complex, time-consuming and suffer from poor performance if the required compensations are not made properly with the dynamism of the scene [5]. However, edge based methods show more robustness against dynamic changes (like illumination variation, quantization error etc) [6, 7]. In [6], one edge map is computed with the edges extracted from current frame (I_n) and other edge map is computed with the edges computed from difference image of I_n and I_{n-1}. These two edge maps are utilized to detect moving edges in I_n. In [7], two edge maps: one is computed from difference image of I_n and I_{n-1}, and another one is computed from I_n and I_{n+1}. Finally, these two edge maps are utilized to extract moving edges in I_n. However, due to illumination variation and quantization error, same edge pixel may change its location in consecutive frames (edge localization error). Hence, pixel based processing and exact matching result scattered moving edges in traditional edge pixel based methods. Moreover, existing three successive frame based methods cannot successfully detect moving edges of the objects having slow movement. This is due to the insignificant gradient value in the respective edge location in overlapping region of the difference image. On the other hand, edges in overlapping region in difference image deteriorate in size and shape than that is in I_n. So, it is difficult to recover actual moving edges in I_n by matching edge maps of left and right difference images. Figure 2 illustrates the problems of three successive frame based moving object detection method. Fig. 2(a)-(c) show three consecutive frames, I_{n-1}, I_n and I_{n+1}, respectively, where moving edges in I_n is shown in Fig. 2(d). Left difference and right difference image edge maps are shown in Fig. 2(e) and Fig. 2(f), respectively. For better understanding of problem, Figure 2(g) and 2(h) shows edges of left difference and right difference edge maps with in the region of moving object in I_n. From these two images, it is impossible to reconstruct the edges of moving object in I_n (Fig. 2(d)) as shape of the edges varies significantly in both images than that are in I_n.

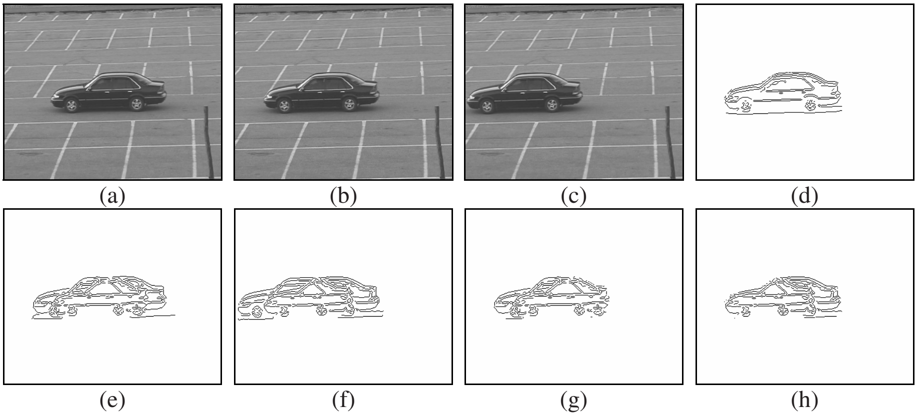


Fig. 2. Illustration of problem in the existing three successive frame based methods

To solve the scattered moving edges, we propose segment based representation of edges instead of working with each edge pixel individually like traditional edge pixel based methods. In this representation, all the edge pixels belonging to a segment are considered as a unit and processed together. Details of segment based representation and its advantages are found in [8]. To solve the problem of detecting moving edges with deteriorated shape, we utilize the edges in I_n and its background edges are removed by a distance transformed based matching with the left and right difference image edge maps. Left difference image edge map, DE_{left} and right difference image edge map, DE_{right} are generated utilizing I_{n-1} and I_n; and I_n and I_{n+1}, respectively, where current image edge map E_n is represented as segments. Insignificant edge information in the overlapping region of difference image is alleviated by incorporating distance transformation based matching and usage of both individual and combined distance transformed images. Use of most recent frames, segment based representation of edges and efficient edge matching algorithm make the system effective in detecting moving edges with reliable shape information for better classification as perfection in shape information is important cue for classification. A neural-net based classification algorithm is used to classify the detected moving objects into known patterns. For classification, we have preliminarily considered two object classes: “Human” and “Vehicle” in our implementation. The number of object classes can easily be extended to more classes with making a little change in the neural net structure and in training patterns.

2 The Proposed Method

The proposed method detects moving object utilizing three consecutive frames. Moving object detection and classification is performed through a number of consecutive steps. The processing steps are illustrated in Fig. 3 and are described in details in the following subsections:

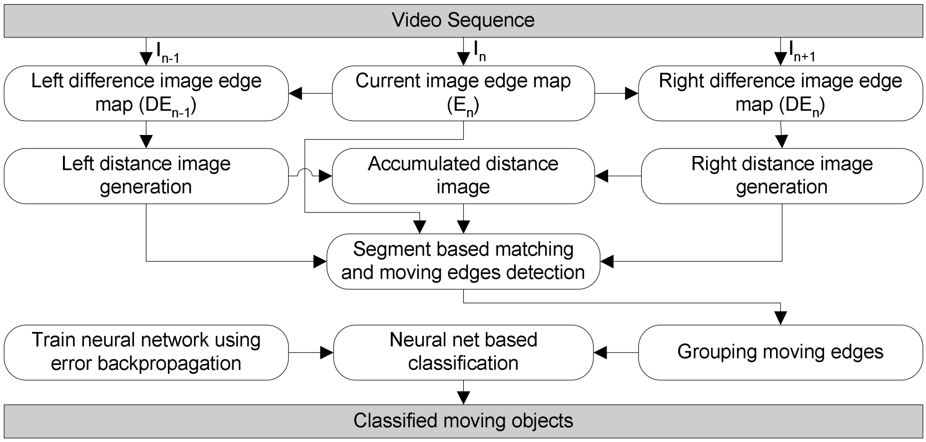


Fig. 3. Flow diagram of the proposed method

2.1 Computation of Difference Image Edge Maps

In our proposed method, the algorithm starts working with computing two difference images using three consecutive frames I_{n-1}, I_n and I_{n+1}. Instead of using edges extracted directly from input images, we have utilized difference image of successive frames for edge extraction and generate difference image edge maps. These edge maps are more noise prone to moving object detection even in dynamic environment. Two difference image edge maps, i.e. left difference image edge map (DE_{left}) and right difference image edge map (DE_{right}) are computed as follows with (1) and (2):

$$DE_{left} = \varphi\{\nabla G * (I_{n-1} - I_n)\} \quad (1)$$

$$DE_{right} = \varphi\{\nabla G * (I_n - I_{n+1})\} \quad (2)$$

Where φ, ∇ and G represent canny edge detector operator, gradient operator and Gaussian mask for filtering on difference images, respectively. Canny edge operators are also applied on I_n for edge extraction and represent the edges as segments (E_n) with an efficiently designed edge class [8]. These edge maps are utilized to extract moving edges. Fig. 4(a)-(c) show I_{n-1}, I_n and I_{n-1}, respectively, where DE_{left}, DE_{right}, and E_n are shown in Fig. 4(d)-(f).

2.2 Matching and Moving Edge Detection

An efficient edge matching algorithm based on integer distance transformation using Chamfer ¾ approximation [9] is used to detect moving edges. In matching one edge map is converted into Distance Transformed Image (DT) and another edge map is superimposed on it to compute the matching confidence. DT is generated using a linear algorithm having two sequential passes. Initially the edge pixels are set to zero and rest of the position is assigned with a very high value. The first pass (forward) modifies the distance image as follows:

$$DT_{i,j} = \min(DT_{i-1,j-1}+4, DT_{i-1,j}+3, DT_{i-1,j+1}+4, DT_{i,j-1}+3, DT_{i,j}) \quad (3)$$

The second pass (backward) works as follows:

$$DT_{i,j} = \min(DT_{i,j}, DT_{i,j+1}+3, DT_{i+1,j-1}+4, DT_{i+1,j}+3, DT_{i+1,j+1}+4) \quad (4)$$

where, $DT_{i,j}$ is the minimum distance at pixel position (i, j) from the nearest edge pixel. With this formulation, we compute two distance transformation images, DT_{left} and DT_{right} using DE_{left} and DE_{right}, respectively. Another distance transformation image (DT_n), is computed with equation (5).

$$DT_n = \min(DT_{left}, DT_{right}) \quad (5)$$

The advantage of using DT_n during matching is that it contains more edge information of moving objects of frame I_n. However, due to contribution of DT_{left} and DT_{right}, it also contains moving edge information of I_{n-1} and I_{n+1}, which may incorporate noise in the detection result. Hence, a two-tier matching approach is utilized in our method. In the first tier, coarse edge map of moving object is computed by employing matching between E_n and DT_n. In the second tier, each of the detected edges of coarse edge map is compared with DE_{left} and DE_{right} separately, and perform a variability test between them. The edge segments of coarse edge map that can satisfy the constraints are finally selected as moving edges, otherwise filters out as noise.

In coarse moving edge map generation, E_n is superimposed on DT_n to accumulate the corresponding distance values. A normalized average of these values (root mean square) is the measure of matching confidence (MC) of the edge segment in E_n, is shown in the following equation:

$$MC[l] = \frac{1}{3}\sqrt{\frac{1}{k}\sum_{i=1}^{k}\{dist(l_i)\}^2} \quad (6)$$

where k is the number of edge points in l^{th} edge segment of E_n; dist(l_i) is the distance value at position i of edge segment l. The average is divided by *3* to compensate for the unit distance *3* in the chamfer ¾-distance transformation. Edge segments are removed from E_n if matching confidence is comparatively higher. Existence of a similar edge segments in E_{n-1} and DT_n produces a low MC value for that segment. We allow some flexibility by introducing a disparity threshold, τ and empirically we set $\tau = 1.3$ in our implementation. We consider a matching occurs between edge segments, if $MC[l] \le \tau$. The corresponding edge segment is considered as moving edge and consequently enlisted to the coarse moving edge map. Finally, the resultant edge map contains the edge segments of E_n that belong to moving objects in I_n. Fig. 4(h) illustrates the procedure of computing matching confidence using DT_n image. In second tier of the algorithm, we utilize DT_{left} and DT_{right} to filter out noise from coarse moving edge map. The noise filtering approach is illustrated as follows:

Step 1: Select an edge segment from coarse moving edge map which has not considered yet.

Step 2: Compute matching confidence for both DT_{left} and DT_{right} which are represented as MC_{left} and MC_{right}, respectively.

Step 3: If ($|MC_{left}$-$MC_{right}|$) $< \sigma$, the edge segment is considered as moving edge. Otherwise, it is discarded from the coarse edge map.

Step 4: Repeat Step 1 to 3 until all edge segments of coarse map are not considered.

Figure 4(g) shows the moving edge segments of I_n using the proposed method. In our implementation, we empirically set $\sigma = 1.5$.

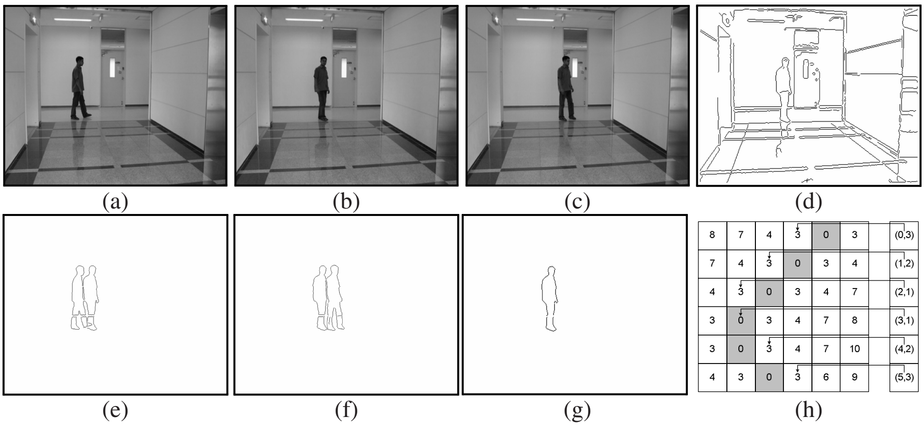


Fig. 4. Moving edge detection. (a) I_{n-1}; (b) I_n; (c) I_{n+1}; (d) E_n; (e) DE_{left}; (f) DE_{right}; (g) Detected moving edge; (h) Edge matching using *DT* image. Here, $MC = 0.91287$.

2.3 Grouping of Moving Edges

For recognition, it is important to identify which edge segments belong to which moving objects in the scenario. For this reason, we have employed chain method of clustering [10] and group edge segments into a number of clusters of moving objects considering their density in a certain region. According to this method, at first, one edge segment is assigned to cluster one. Then, the distance, d of next edge segment to previous one is computed and compared it to d_0 (predefined threshold value). If $d < d_0$, the second edge segment is assigned to same cluster; otherwise, a new cluster is formed for the new sample. In the same way, distances d from next sample to all existing clusters are computed and among them, d_{min} is selected. If $d_{min} < d_0$, then this new edge segment is assigned to corresponding cluster. Otherwise, it again forms a new cluster. In this way, all detected edge segments are grouped into moving objects. In our method, we have utilized value of *MC* described in previous section as d for clustering the edge segments. Finally, the region of interest (ROI) is located by computing a bounding box around each of the clusters of edges which are used in later subsection for classification.

2.4 Neural-Net Based Classification

A feed-forward neural-net based approach is utilized in this section for classification of moving objects. In the proposed method, our neural net is consisted with four layers- one input layer with one neuron per feature, two hidden layers with seven neurons each and one output layer with one neuron per class. In our method, sigmoid activation functions are used for the neurons. Since shape information is utilized in our method for classification, we have normalized the size of ROI into the size of training patterns and then each pixel of normalized ROI is used as input feature in the neural-net for classification. We have employed 25 human and 25 vehicle patterns with 15,000 iterations to train the network for recognition. In the training phase, we have manually picked edge segments of moving objects observing a number of frames utilizing our detection method and then normalized it to train the system. In the training phase we have employed error-back propagation method. As for example, some training patterns are shown in Fig. 5. In the output layer two thresholds are used: a decision for one class is only made if the value of the output neuron representing this class is above the upper threshold and simultaneously the value of the other output neuron is below the lower threshold. Otherwise the object in question is rejected. Thus, decisions for the wrong class are highly unlikely.

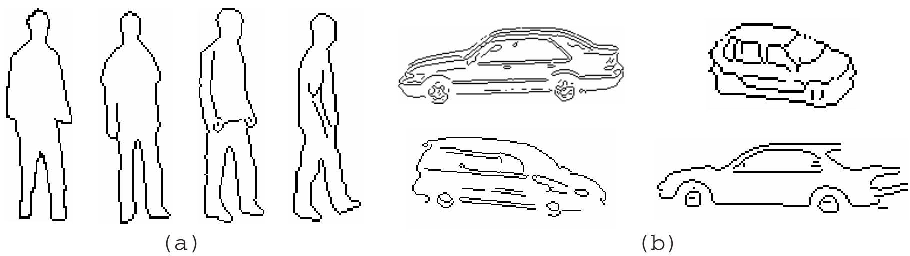


Fig. 5. Sample of training patterns that are used to train the neural-net. (a) Training patterns for human detection; (b) Training patterns for vehicle detection.

3 Experimental Results

Experiments were carried out with several video sequences captured from indoor corridor and outdoor environments to verify the effectiveness of the proposed method. We applied our proposed method on video formats of size 640 x 520 and used Intel Pentium IV 1.5 GHz processor and 512 MB of RAM. Visual C++ 6.0 and MTES [11], an image processing environment tool were used as of our working environment tool for implementation.

Fig. 6 shows experimental result for moving edge detection in a video sequence. Here, Fig. 6(a) and 6(b-d) show background image and three consecutive frames I_{272}, I_{273}, and I_{274} having different illuminations. The output from Kim and Hwang [6] approach is shown in Fig. 6(e), where double edge maps are utilized to detect moving edge pixels. In their method, the difference between background and current frame incorporates most of the noise pixels. Fig. 6(f) shows the result obtained after applying Dailey and Cathey [7] approach on frame I_{272}, I_{273}, and I_{274}. Result obtained from this method is much robust against illumination changes as it is also using most recent successive frame differences for moving edge detection. However, this method suffers from scatter edge pixels as it uses logical AND operation for pixel based matching of edges in difference edge maps to detect moving objects. But due to presence of random noise the same edge pixels did not appear in the same position in both of the edge maps. As a result, some portion of the same edge segment matched and some did not. This created lots of breaks in edges of detected moving objects (Fig. 6(f)). Our method does not suffer from this kind of problem as we have applied a flexible matching between edge segments. The result obtained using our proposed method is shown in Fig. 6(g).

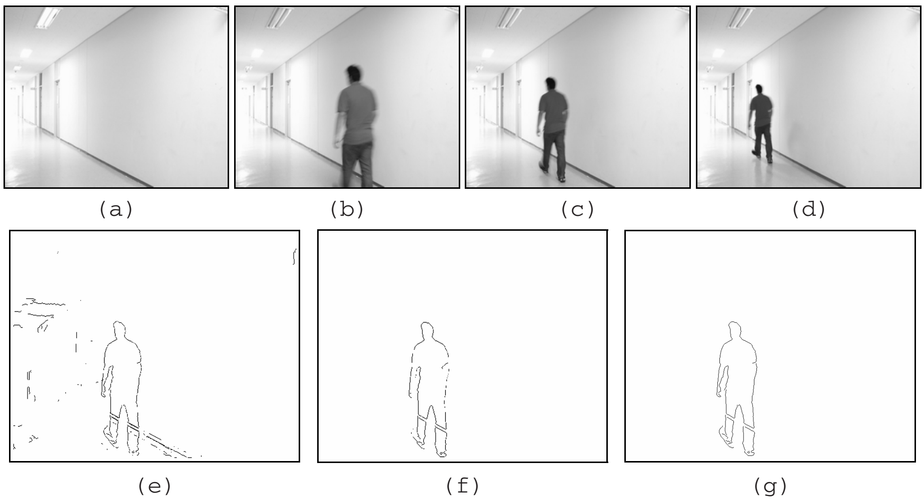


Fig. 6. (a) Background image; (b) I_{172}; (c) I_{173}; (d) I_{174}; (e) Detected moving edges of I_{173} using [6]; (f) Detected moving edges of I_{173} using [7]; (g) Moving edges of I_{173} using proposed method

Fig. 7 illustrates classification results of some experiments captured in indoor and outdoor environments. Here, the image frames are characterized with un-occluded objects and in each case detected moving objects are classified correctly. In such kind of situation, the classification result is robust for different shape of humans as well as vehicles. Moreover, classification does not depends on object size as the size of the training patterns as well as detected moving objects are normalized into same scale before using it for recognition. However, in case of occluded objects, classification module can still work effectively if occlusion occurs in small portion of the moving objects. Otherwise, it will reject the detected moving object from classification and will consider it as an unknown pattern.

In order to comprehend the computational efficiency of the algorithm, it should be mentioned that, with the processing power and the processing steps described above, execution time for the object detection on grayscale images was approximately 106 ms. Therefore, the processing speed is 9.43 frames per second. Although this frame rate is about 0.3772 of the real time frame rate (i.e. 25 frames per second), but this speed is quite reasonable for real time processing. However, using computers with higher CPU speeds which are available this day and in future as well, this frame rate can be improved. Table 1 depicts approximate times required to execute different modules of the proposed method.

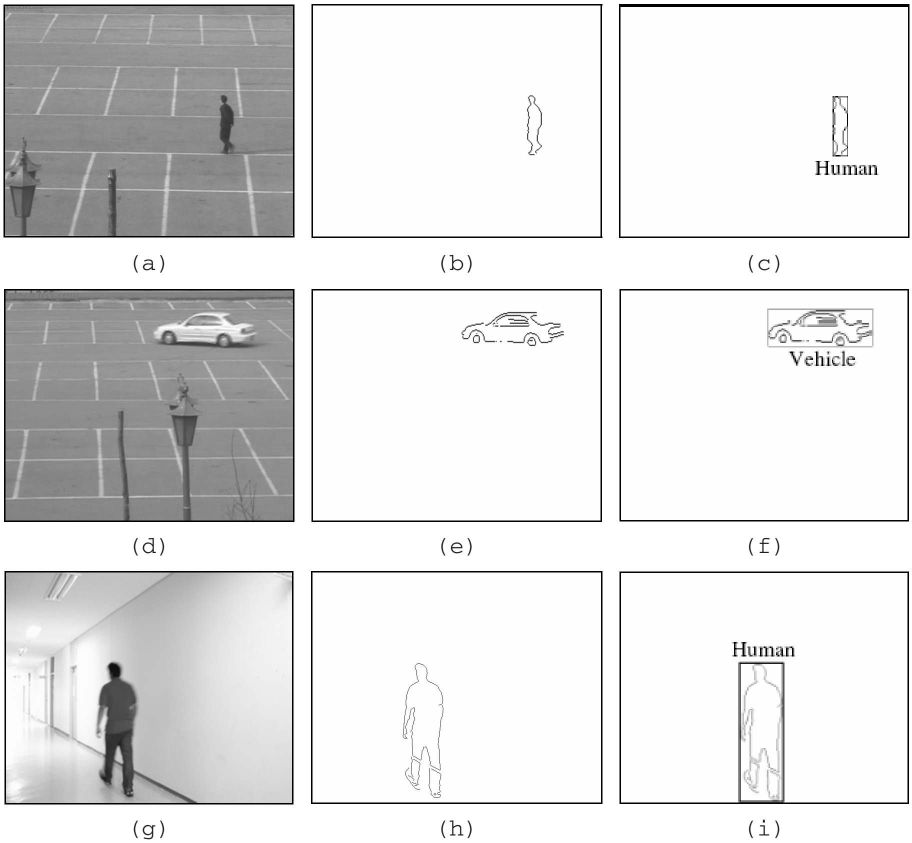


Fig. 7. Moving object classification. (a), (d) and (g) are current frames; where (b), (e) and (h) are respective detection results; (c), (f) and (i) are classification results of (a), (d) and (g), respectively.

Table 1. Mean processing time (in ms) for each module

Processing steps	**Mean time (ms)**
Computation of difference images	5
Edge map generation from difference image	39
DT image generation	11
Computation of matching confidence and moving edge detection	19
Grouping of edges into moving objects and locating ROI	26
Neural net based classification	6
Total time required	106

4 Conclusions and Future Works

This paper presents a robust method for automatic detection and classification of moving objects which can be effectively utilized in intelligent video monitoring system as well as in vision agent based home networking and robot navigation systems. Since we propose edge based method for detection, it is robust against illumination variation and random noise. Moreover, segment based representation of edges increases the robustness with compare to traditional edge pixel based methods by allowing use of geometric shape information of it during matching. Since the proposed method does not require any complex background generation and update method with it for detection, this method is faster and efficient with compare to background subtraction based methods. For classification, we utilize fed-forward neural network where detected moving edge information is used as shape features in the network. Neural-net based algorithms have already proven to be very useful and efficient for classification. This method can classify the moving objects very accurately if the network is trained properly with the sufficient training sets and if the data to be classified contains reasonable shape information with it. Since our edge segment based approach generates very accurate shape information of moving object, neural-net can classify it successfully. Moreover, use of double threshold makes the neural net decision more reliable. However, occlusion might create some problem for classification. Hence, in our future work we are trying to incorporate uncertainty management based approach which would be able to retrieve the occluded part of the moving objects and can classify it into known patterns accurately.

References

1. Radke, R., Andra, S., Al-Kohafi, O., Roysam, B.: Image Change Detection algorithms: A Systematic Survey. IEEE Transactions on Image Processing 14(3), 294–307 (2005)
2. Sogo, T., Ishiguro, H., Ishida, T.: Mobile Robot Navigation by a Distributed Vision System. Journal of New Generation Comp. 19(2), 121–138 (2001)
3. DeSouza, G.N., Avinash, C.K.: Vision for Mobile Robot Navigation: A Survey. IEEE Transaction on PAMI 24(2), 237–267 (2002)
4. Lee, M., Kim, Y., Uhm, Y., Hwang, Z., Kim, G., Park, S., Song, O.: Location-Aware Multi-agent Based Intelligent Services in Home Networks. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 178–187. Springer, Heidelberg (2006)
5. Gutchess, D., Trajkovics, M., Cohen-Solal, E., Lyons, D., Jain, A.K.: A background model initialization algorithm for video surveillance. IEEE International Conference on Computer Vision 1, 733–740 (2001)
6. Kim, C., Hwang, J.N.: Fast and automatic video object segmentation and tracking for content-based applications. IEEE Transaction on CSVT 12, 122–129 (2002)
7. Dailey, D.J., Cathey, F.W., Pumrin, S.: An algorithm to estimate mean traffic speed using un-calibrated cameras. IEEE Transactions on ITS 1(2), 98–107 (2000)
8. Hossain, M.J., Dewan, M.A.A., Chae, O.: Moving Object Detection for Real Time Video Surveillance: An Edge Segment Based Approach. IEICE Transaction on Communications E90-B(12), 3654–3664 (2007)
9. Borgefors, G.: Hierarchical chamfer matching: A parametric edge matching algorithm. IEEE Transactions on PAMI 10(6), 849–865 (1988)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley Interscience, Chichester (November 2000)
11. Lee, J., Cho, Y.K., Heo, H., Chae, O.S.: MTES: Visual Prgramming for Teaching and Research in Image Processing. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2005. LNCS, vol. 3514, pp. 1035–1042. Springer, Heidelberg (2005)

Data Integration in a System with Agents' Models*

Grażyna Brzykcy

Poznań University of Technology, Institute of Control and Information Engineering,
Pl. M. Skłodowskiej-Curie 5, 60-965 Poznań, Poland
Grazyna.Brzykcy@put.poznan.pl

Abstract. A problem of data integration in an environment consisting of a community of peer-to-peer cooperative agents is discussed. Data in the system are locally structured by the agents under different schemas. Communication and cooperation in the system are based on asking and answering queries by the agents. We show how to represent data and to merge related information from multiple data sources by means of unification operation and diverse Prolog mechanisms. The declarative approach is suitable to model data integration task and other constraints on data: from ontological to locally induced by an agent.

Keywords: P2P system, XML data, schema mapping, query rewriting, unification, Prolog-like computations.

1 Introduction

Distributed and heterogeneous nature of today's software systems is a reason for a lot of efforts in the domain of data integration [8,13,14,15,17,18]. In the paper a data integration problem is considered in systems consisted of autonomous components (agents, peers) [16] each of which manages some part of data (knowledge). An agent can independently decide how to structure its local data and can cooperate with other agents by asking them and answering queries. Peers decide when to join and leave the system, when to communicate and share their data with other partners. The SIXP2P system [6] for semantic integration of XML data, currently under development in Poznań University of Technology, is a representative example of such systems and forms a suitable testing ground for diverse solutions of data integration problem.

A P2P collection of data sources may be used as a general model of information structure. A query defined by a user and directed to a local agent in the structure is cooperatively answered by all the related peers in the P2P system. Partial answers from diverse sources are integrated and the result is presented by the local agent to the user in a form described in the query. How to model the agent's main action of integration of partial data?

We decided to take advantage of unification operation as the natural model of data integration. We also find it rational to use other Prolog-like mechanisms in the system. Logic programming environments, particularly those with types and extended

* The work was supported in part by the Polish Ministry of Science and Higher Education under Grant N516 015 31/1553.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 162–171, 2008.

records (ψ-terms, feature structures) [1,2,9], provide not only a high level of abstraction for knowledge representation, but also a pattern matching, unification and subsumption – techniques used in data integration tasks. Moreover, the systems come with metaprogramming facilities, a backtracking search, executable specifications and diverse extensions (e.g., *LogicPeer*, *LogicWeb* [11], situations [12], constraint programming [1]). We have already prepared the declarative specification of semantic-driven communication between agents and between agents and the broker in the SIXP2P system [6] and have shown how to cover context in these rules [5]. Now the focus will be on a problem of representation of data, schemas and mappings between XML data and on the tasks of asking and answering queries, data exchange and data merging. Particularly, in the paper we address the following issues: how to represent data, schemas and mappings (section 3), how to represent queries and answers (section 4), how to perform data exchange task (section 5) and how to integrate data from different sources (section 5). In section 2 we shortly characterize the SIXP2P system.

As our ideas and solutions grow out of the analysis of Yu's and Popa's work [17] on query rewriting and data integration, we have decided to follow their line of presentation.

Declarative approach means that instead of imperative algorithms (e.g., defined in Java [17]) we use logical variables to denote entities (data), Prolog-like clauses to specify algorithms and logic programming systems to perform computations. Benefits of the approach are as follows:

1. Logical variables are suitable for data representation: to express data equality, to represent partial (partially instantiated) data, to achieve more readable specifications of mappings and queries.
2. Unification operation is a suitable model of data integration task and is automatically performed in Prolog systems. We are able to match data effectively.
3. Data merging in Prolog systems is performed with built-in terms (lists) and procedures (e.g., *member*).

2 Agents and Tasks in the SIXP2P System

The SIXP2P system for semantic integration of XML data consists of autonomous agents, which manage local data sources. An agent decides how to structure its data and when to share them with other partners. The data structure is described by a peer's local schema, which contains all the information that is vital to an agent to ask and answer queries. Therefore, a schema forms a model of an agent in the system [7].

Communication between agents is defined locally as each agent in the system can "see" only a subset of other agents. However, the "seeing" relation is transitive and induces a significant extension of the set of available data sources. Thus, cooperative query evaluation is possible in the system and the task is performed together with agents indirectly connected to the enquirer.

The main goal in the system consists in answering queries by taking advantage of all the knowledge sources that store semantically related data. There are two basic forms of achieving this goal: by data exchange and by data integration.

Data exchange task [3,10] is performed to restructure data from one schema to another. In the setting a problem of cooperative evaluation of a query directed to the peer P can be solved by data transformation from all P's peers into the P's schema. Then a local query answering is possible. The data exchange task is shown in figure 1 where *mi* denotes transformation from *Pi* into *P*. It is worth noticing that additional procedure of data merging have to be executed in view of data partiality and overlapping.

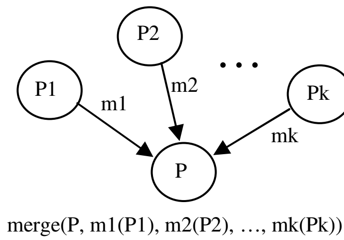


Fig. 1. Data exchange and data merging in P2P system

Data integration problem is more oriented on a query reformulation task. It rather consists of rewriting a query directed to P into queries directed towards P's peers than of transformation of the whole source data. In figure 2 *Di* denotes the answer for query *Qi,* which is directed to *Pi* and is the new (rewritten) version of the original query *Q* (directed to *P*). The data merging procedure (where *D* denotes local answer for *Q*) is valid in this case too.

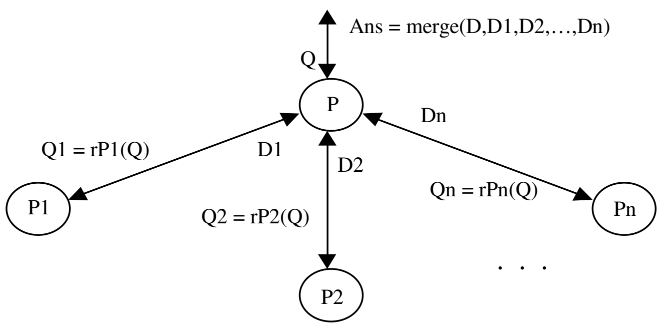


Fig. 2. Query rewriting and data fusion in P2P system

In the SIXP2P system we use, like in [17], a nested relational representation as a data model and define a schema as a set of elements with associated types. Atomic data are of simple types *(Int* or *String*), whereas repeatable elements of the same type *T* are treated as set members (complex type *SetOf T*), and groups (or sequences) of data of different types form records.

3 Representation of Schemas and Mappings

To represent sets and records in Prolog we can use lists. For a record it should be a list of a fixed length. We show such a representation for schemas recalled from [17] and depicted in figure 3.

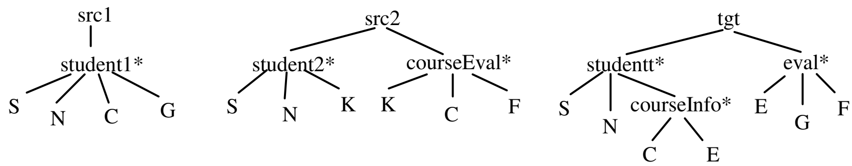


Fig. 3. Schemas (src1, src2 – source schemas, tgt – target schema)

We assume that *src1*, *src2* and *tgt* are data sources with the same domain of application. The * symbol is used to denote sets, and the symbols *S*, *N*, *C*, *G* and *F* are for, respectively, "student id", "student name", "course", "grade" and "file evaluation". The elements *K* and *E* play a role of *key* in sets *courseEval** and *eval**, and a role of *keyref* in sets *student2** and *courseInfo** respectively.

In Prolog specification one can represent any entity as a term, for instance any data of record type *student1* as *St1*, where *St1* is a variable name. The more precise representation of the record data is also possible and may be in a form of a list. In our example the *[S,N,C,G]* list is used to show explicitly components of the record. A whole set also may be named by a variable, for instance *Cs* in the record (list) *[S,N,Cs]* of the type *studentt*. To show that a pair *[C,E]* belongs to the set *Cs* it is possible to use the *member* predicate (built-in predicate in Prolog systems) and specify constraint *member([C,E],Cs)*. Surely, the empty set is represented by the empty list respectively.

Schemas from the figure 3 are defined as follows:

- data from a source *src1* are structured in the list *[St1]*, where each element of the root *St1* takes a form of record *[S,N,C,G]*,
- data from a source *src2* – in the list *[St2,Cev]* with the records *[S,N,K]* as elements of the set (list) *St2* and the records *[K,C,F]* as members of the list *Cev*,
- and data from a target *tgt* – in the list *[Stt,Ev]*, where member of the list *Stt* is a nested record *[S,N,Cs]* with the list *Cs* that consists of records *[C,E]*, and the list *Ev* is built of records *[E,G,F]*.

The Prolog language and logical variables are even more suitable for data representation than it is seen at the first glance. Logical variables denote entities (play a role of co-references) and may be used, in a natural way, to represent partial data. If we know that a student *n* takes a course *c* but a grade is not known yet, then the incomplete data may be represented as a term *[S,n,c,G]*, where *n* and *c* are constants, and *S* and *G* are variables denoting certain "student id" and "grade" respectively. Similarly, one can describe optional data by uninstantiated variables. Moreover, the same variables denote the same data, so it is rational to consider the list *[S,N,C,G]* from the set of *student1* and the list *[S,N,K]* from the set of *student2* to be two representations of the same student (with the same "student id" *S* and the same "student name" *N*). With unification operation in Prolog systems we are able to integrate partial data. For instance, taking three lists: *[S,N,C,G*] from a set *student1**, *[S,N,K]* from a set *student2** and *[K,C,F]* from a set *courseEval**, one can form the appropriate component data of the target schema, particularly, a list *[S,N,Cs]* from a set *studentt**, a list *[C,E]* from a set *Cs* and a list *[E,G,F]* from a set *eval**.

To perform cooperative query evaluation in a set of agents we have to establish mappings between their schemas (models). These relationships form correspondences between groups of elements in one schema and groups of elements from another schema. In [17] a mapping is a statement of the form:

```
foreach x₁ in g₁, …, xₙ in gₙ
        where B₁
exists y₁ in g'₁, …, yₘ in g'ₘ
        where B₂
with e'₁ = e₁ and … and e'ₖ = eₖ
```

where for each element x_i from the set g_i in the source schema, such that condition B_1 obtains, there are corresponding elements y_j from the sets g'_j in the target schema, such that condition B_2 and condition from the part <u>*with*</u> are valid. All the conditions are conjunctions of equalities between expressions built respectively over source and target variables and constants.

We define mappings in Prolog by means of the predicate *mapping*

```
mapping(Id,[X1,X2,…,Xn],[Y1,Y2,…,Ym]):-
        B.
```

where *Id* is a mapping identifier and *Xi* and *Yi* are records from the sets g_i and g'_i respectively. The mapping *m1* from *src1* data into *tgt* data, indirectly defined in the figure 3, is represented as

```
mapping(m1,[[S,N,C,G]],[[S,N,Cs],[C,E],[E,G,F]]):-
        member([C,E],Cs).
```

and the mapping *m2* from *src2* data into *tgt* data as

```
mapping(m2,[[[S,N,K],[K,C,F]],[[S,N,Cs],[C,E],[E,G,F]]):-
        member([C,E],Cs).
```

The examples show how we can design mapping definitions in Prolog. Equalities of expressions from mapping statements in [17] are used to express semantic identity of entities represented by data. In Prolog these dependencies may be reduced to variable name identity. For instance, to show that *src2.student2.K=src2.courseEval.K* in the mapping *m2* we use the same variable name *K* in *[S,N,K]* record and *[K,C,F]* record. The body *B* in mapping definition is suited to define well-formed generators from <u>*foreach*</u> and <u>*exists*</u> parts and consists of *member* constraints. In the above specified mappings *Cs* depicts the set of *[C,E]* records.

To make a mapping definition more readable we postulate an explicit specification of data schemas. Thus, in records *[X1,X2,…,Xn]* and *[Y1,Y2,…,Ym]* we give only source and target schema roots and extend a mapping body with appropriate *member* constraints for other generators. The new versions of *m1* and *m2* mappings are depicted below:

```
mapping(m1,[St1]],[Stt,Ev]):-
        member([S,N,C,G],St1),
        member([S,N,Cs],Stt),
        member([C,E],Cs),
        member([E,G,F],Ev).
mapping(m2,[St2,Cev],[Stt,Ev]):-
        member([S,N,K],St2),
        member([K,C,F],Cev)
        member([S,N,Cs],Stt),
        member([C,E],Cs),
        member([E,G,F],Ev).
```

We notice here that definition of reverse mapping is very similar to the original one. It forms a base for a query reformulation task. In our examples specification of correspondences between elements of *tgt* data and *src1* data (the *rm1* mapping) takes the following form:

```
mapping(rm1,[Stt,Ev],[St1]):-
        member([S,N,Cs],Stt),
        member([C,E],Cs),
        member([E,G,F],Ev),
        member([S,N,C,G],St1).
```

and differs from mapping *m1* on the order of second and third arguments.

It is easily seen that Prolog specification is more compact and more flexible then the one presented in [17]. At the same time more complex constraints may be defined in Prolog mappings.

4 Representation of Queries

Query answering, like data exchange and data integration tasks, is performed on data instances. An instance of data of a given schema is defined as a set of values that respect types defined in the schema. As it is stated in [17] the queries have the general form:

```
for x1 in g1, …, xn in gn
        where Bans
        return r
```

where x_i, g_i and B_{ans} are defined as with mappings and *r* is a record. Some components of the record may be defined by another, nested query .

A query is always directed to an instance of data of a given schema (source schema) and describes the answer schema and all the constraints imposed on data from the source. In Prolog we represent queries by means of predicate *query*

```
query(Id,SourceSchema,[Ans1,…,Ansk]):-
        findall(R1,B1,Ans1),
        …
        findall(Rk,Bk,Ansk).
```

where *Id* is a query identifier, *SourceSchema* is an explicit representation of schema of queried data, *Ri* expresses the record type of the set *Ansi*, which forms a part in the answer schema *[Ans1,…,Ansk]*. The built-in predicate *findall* is used to produce the list (set) *Ansi* of terms *Ri*, that fulfill the condition (goal) *Bi*. The query *q1*, defined in [17] in the way:

```
for s in tgt.studentt*, c in s.studentt.courseInfo*, e in tgt.eval*
       where c.courseInfo.E = e.eval.E
       return [name=s.studentt.N, course=c.courseInfo.C,
              grade=e.eval.G,   file=e.eval.F]
```

takes in our setting the following form:

```
query(q1,[Stt,Ev],Ans):-
       findall([N,C,G,F],(member([S,N,Cs],Stt),member([C,E],Cs),
                         member([E,G,F],Ev)),Ans).
```

where *[Stt,Ev]* is a *tgt* data schema, *[N,C,G,F]* is an answer record, and the records *[S,N,Cs]*, *[C,E*], and *[E,G,F]* correspond to variables *s*, *c* and *e* from the original query. And the nested query *q2*, defined in [17] as follows:

```
for s in tgt.studentt*, c in s.studentt.courseInfo*, e in tgt.eval*
       return [name=s.studentt.N,
              results=for s' in tgt.studentt*,
                     c' in s'.studentt.courseInfo*, e' in tgt.eval*
                     where c'.courseInfo.E=e'.eval.E
                          and s'.studentt.N=s.studentt.N
                     return [result=[grade=e'.eval.G, file=e'.eval.F]]
```

is specified as “nested” Prolog query:

```
query(q2,[Stt,Ev],Ans):-
       findall([N,Gf],(member([S,N,Cs],Stt),
                     findall([G,F],(member([C,E],Cs),
                                   member([E,G,F],Ev)),Gf),Ans).
```

with *Gf* denoting the set (list) of records *[G,F]*, which are constructed on the base of *[C,E]* records from the set *Cs* and the appropriate *[E,G,F]* records from the list *Ev*.

5 Data Exchange

Data exchange task consists in transformation of the source data into the target data. For an instance of the source schema the corresponding instance of the target schema is built, with the requirement that the mapping definition is respected. The task may be specified as an evaluation of a special query: directed to a set of source data and with the answer schema equal to the target schema. Constraints imposed on data are the same as specified in the mapping. Therefore, to transform *src1* data into the *tgt* schema the following query has to be answered:

```
query(m1,[St1],[Stt,Ev]):-
       findall([S,N,Cs],(member([S,N,C,G],St1),
                        findall([C,E],(member([S,N,C,G],St1),
                                      gensym(e,E)),Cs)),Stt),
       findall([E,G,F],(member([S,N,C,G],St1),
                      member([S,N,Cs],Stt),
                      member([C,E],Cs)),Ev).
```

where source records *[S,N,C,G]* are used together with target records *[S,N,Cs]*, *[C,E]* and *[E,G,F]*. The unique values of the *key* (*keyref*) *E* come from a buit-in predicate *gensym*.

In the SIXP2P system all the answers an agent receives should be merged in order to eliminate overlapping data and to unify partially described data. One can notice that the merging action is executed on instances of the same schema and in Prolog systems it is equivalent to the union operation on sets. Below we present the specification of the union:

```
union([],S,S):-!.
union([E|T],S,S1):-
   insert(E,S,SE),
   union(T,SE,S1).

insert(E,S,S1):-
   member(E,S),!,
   S1=S.
insert(E,S,[E|S]).
```

An element *E* is added (inserted) to a set *S* if and only if a constraint *member(E,S)* does not hold.

6 Data Integration

To exploit other data sources in the process of cooperative evaluation of local query we have to reformulate the query and to direct it at an appropriate subset of peers (a set of agent's partners). Query rewriting is an important part of a data integration task. It depends on the original query addressed to P, on P's partners and mappings from P to the partners. All the rewritten queries have the same answer schema (the answer schema of the original query), but diverse source schemas. Conditions in the queries have to be modified to cover correspondences between schemas, which are defined by mappings. Now, however, we are interested in dependences of the opposite direction, namely, how to express target data in the source schema. Therefore, in each condition from the original query we substitute elements of the target data with the corresponding elements of the source data and remove duplicated constraints. We describe the procedure by example. Let us recall the *q1* query directed towards *tgt* data:

```
query(q1,[Stt,Ev],Ans):-
       findall([N,C,G,F],(member([S,N,Cs],Stt),member([C,E],Cs),
                           member([E,G,F],Ev)),Ans).
```

and reverse mapping *rm1* (from *tgt* data to *src1* data):

```
mapping(rm1,[Stt,Ev],[St1]):-
       member([S,N,Cs],Stt),
       member([C,E],Cs),
       member([E,G,F],Ev),
       member([S,N,C,G],St1).
```

All the records *[S,N,Cs]*, *[C,E]* and *[E,G,F]* are mapped on the *[S,N,C,G]* record of *src1* data. Consequently, the rewritten query *rq11*, directed at source data *src1*, is constructed from the *q1* query and takes the following form:

```
query(rq11,[St1],Ans):-
   findall([N,C,G,F],(member([S,N,C,G],St1)),Ans).
```

If we assume, like in [17], that in the system there are mappings between the sources *src1*and *src2* and the target *tgt*, then in data integration task not only it is allowed to evaluate queries with *tgt* data, but also with the source data. As the new version of *q1* query directed at *src1* data is already constructed, we now define the rewritten version for the *src2* data. The useful mapping *rm2* is formed:

```
mapping(rm2,[Stt,Ev],[St2,Cev]):-
        member([S,N,Cs],Stt),
        member([C,E],Cs),
        member([E,G,F],Ev),
        member([S,N,K],St2),
        member([K,C,F],Cev).
```

and the *rq12* query is built by an appropriate modification of the conditions from the *q1* query:

```
query(rq12,[St2,Cev],Ans):-
   findall([N,C,G,F],(member([S,N,K],St2),member([K,C,F],Cev)),Ans).
```

With reformulated queries (e.g., queries *rq11* and *rq12* for the original query *q1*) the data integration task is executed if we merge answers from different data sources (e.g., an answer *Ans1* from the source *src1* and an answer *Ans2* from the source *src2*), including the local one (e.g., an answer *Ans*0 from *tgt* data).

```
query(q1,Ans):-
        query(q1,[Stt,Ev],Ans0),
        query(rq11, [St1],Ans1),
        query(rq12,[St2,Cev],Ans2),
        union(Ans0,Ans1,Ans01),
        union(Ans01,Ans2,Ans).
```

The declarative specification of answering queries (e.g., a query q1) is readable – to execute query q1 one has to find local answer, to execute rewritten queries (directed to partners with other schemas), and to merge partial answers with the local one. At the same time it is also executable.

7 Conclusions

In this paper we consider some problems related to data exchange and data integration tasks in P2P systems. We use declarative approach to the problems and show how to represent data and how to merge related information from multiple sources by means of unification operation and other Prolog mechanisms. Equipped with reasoning tools over partial data and metaprogramming facilities, Prolog systems are particularly suited to support the semantic-oriented distributed processing. It is known that XML data, as well as more flexible data types (e.g., ψ-terms), can be expressed in extended Prolog systems [4].

References

1. Ait-Kaci, H., et al.: The Wild LIFE Handbook, Paris Research Laboratory (1994)
2. Ait-Kaci, H., Podelski, A., Goldstein, S.: Order-sorted feature theory unification. Journal of Logic Programming 30(2), 1279–1318 (1997)
3. Arenas, M., Libkin, L.: XML Data Exchange: Consistency and Query Answering. In: PODS Conference, pp. 13–24 (2005), 2005

4. Bartoszek, J.: Przekształcanie dokumentów XML w ψ-termy. SAiI tom 26, 1–15 (2001)
5. Brzykcy, G., Bartoszek, J.: Context in rules used in P2P semantic data integration system. In: Marchiori, M., Pan, J.Z., Marie, C.d.S. (eds.) RR 2007. LNCS, vol. 4524, pp. 377–380. Springer, Heidelberg (2007)
6. Brzykcy, G., Bartoszek, J., Pankowski, T.: Semantic data integration in P2P environment using schema mappings and agent technology. In: Proceedings of AMSTA 2007 (to appear)
7. Bouquet, P., Serafini, L., Zanobini, S.: Peer-to-peer semantic coordination. Journal of Web Semantics 2(1), 81–97 (2004)
8. Calvanese, D., Giacomo, G.D., Lenzerini, M., Rosati, R.: Logical Foundations of Peer-To-Peer Data Integration. In: Proc. of the 23rd ACM SIGMOD Symposium on Principles of Database Systems (PODS 2004), pp. 241–251 (2004)
9. Carpenter, B.: The logic of typed feature structures: inheritance (in)equations and extensionality. In: Second European Summer School in Language, Logic and Information, Leuven (1990)
10. Fagin, R., Kolaitis, P., Popa, L.: Data exchange: getting to the core. ACM Trans. Database Syst. 30(1), 174–210 (2005)
11. Loke, S.W.: Declarative programming of integrated peer-to-peer and Web based systems: the case of Prolog. Journal of Systems and Software 79(4), 523–536 (2006)
12. Loke, S.W.: Representing and Reasoning with Situations for Context-Aware Pervasive Computing: a Logic Programming Perspective. School of Computer Science and Software Engineering, Monash University, Australia (2004)
13. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Improving Automatically Created Mappings Using Logical Reasoning. In: Proc. of the 1st International Workshop on Ontology Matching OM-2006. CEUR Workshop Proceedings, vol. 225 (2006), `http://CEUR-WS.org/Vol-225`
14. Pankowski, T.: Management of executable schema mappings for XML data exchange, Database Technologies for Handling XML Information on the Web. In: Grust, T., Höpfner, H., Illarramendi, A., Jablonski, S., Mesiti, M., Müller, S., Patranjan, P.-L., Sattler, K.-U., Spiliopoulou, M., Wijsen, J. (eds.) EDBT 2006. LNCS, vol. 4254, pp. 264–277. Springer, Heidelberg (2006)
15. Pankowski, T., Cybulka, J., Meissner, A.: XML Schema Mappings in the Presence of Key Constraints and Value Dependencies. In: ICDT 2007 Workshop EROW 2007, CEUR Workshop Proceedings, vol. 229, pp. 1–15 (2007), `http://CEUR-WS.org/Vol-229`
16. Wooldridge, M.: An Introduction to Multiagent Systems. John Wiley & Sons, Chichester (2002)
17. Yu, C., Popa, L.: Constraint-Based XML Query Rewriting for Data Integration. In: SIGMOD Conference, pp. 371–382 (2004)
18. Yu, C., Popa, L.: Semantic Adaptation of Schema Mappings when Schema Evolve. In: Proceedings of the 31th VLDB Conference, Trondheim, Norway (2005)

A Comparative Study of Two Short Text Semantic Similarity Measures

James O'Shea, Zuhair Bandar, Keeley Crockett, and David McLean

Department of Computing and Mathematics, Manchester Metropolitan University,
Chester St., Manchester M1 5GD, United Kingdom
{j.d.oshea, z.Bandar, k.crockett, d.mclean}@mmu.ac.uk

Abstract. This paper describes a comparative study of STASIS and LSA. These measures of semantic similarity can be applied to short texts for use in Conversational Agents (CAs). CAs are computer programs that interact with humans through natural language dialogue. Business organizations have spent large sums of money in recent years developing them for online customer self-service, but achievements have been limited to simple FAQ systems. We believe this is due to the labour-intensive process of scripting, which could be reduced radically by the use of short-text semantic similarity measures. "Short texts" are typically 10-20 words long but are not required to be grammatically correct sentences, for example spoken utterances and text messages. We also present a benchmark data set of 65 sentence pairs with human-derived similarity ratings. This data set is the first of its kind, specifically developed to evaluate such measures and we believe it will be valuable to future researchers.

Keywords: Natural Language, Semantic Similarity, Dialogue Management, User Modeling, Benchmark, Sentence.

1 Introduction

A Conversational Agent (CA) is a computer program that interacts with a human user by means of natural language dialogue. The motivation for our work is the development of a new generation of Conversational Agents (CAs) with improved techniques for dialogue management. These techniques involve highly sophisticated algorithms for the measurement of Short Text Semantic Similarity (STSS) [1, 2]. A Short Text (ST) in typical human dialogue would be a sentence in the range of 10-20 words, bearing in mind that user utterances include other forms that fail to conform to the grammatical rules of sentences. Other applications that can benefit from STSS measures are the automatic processing of text and e-mail messages [3] and natural language interfaces to databases [4]. Academic studies include health care dialogue systems [5], real estate sales [6], phone call routing [7] and intelligent tutoring [8]. CAs will be increasingly important in the future as these applications are delivered remotely via the internet.

One of the most important applications of CAs is online customer self-service, providing the user with the kind of services that would come from a knowledgeable or

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 172–181, 2008.

experienced human. In 2005 there were at least 10 major companies operating in this area, including IBM and strategic partners of Microsoft [9]. At least 28 patents have been registered concerning Conversational Agents and closely related technologies. With so much investment in R&D, where are the tangible results? Commercial CAs are basic question answering systems, incapable of genuine mixed-initiative or extended dialogue. It is now recognized that there are genuine obstacles to the transfer of CAs from the research environment to the real world [5].

Pattern matching has been identified as one of the most common and capable methods for developing dialogues that seem to be coherent and intelligent to users [5]. Patterns are grouped in rules which in turn are contained in a script file [10]. When a script is executed user utterances are compared to the patterns and the closest match results in the relevant rule firing. This generates a response to the user and passes information to other programs making up the agent for relevant action. Creating scripts is a highly skilled craft [11], requiring the anticipation of user utterances, generation of permutations of the utterances and generalization of patterns through the replacement of selected terms by wild cards. Modifications to rules containing the patterns can impact on the performance of other rules and modern pattern matching systems contain many parameters that further modify their behaviour. The main disadvantage of pattern matching systems is the labour-intensive (and therefore costly) nature of their development.

State-based systems, popular in healthcare [5], provide an alternative form of dialogue management; undergoing state transitions triggered by the content of user utterances. In simple systems tight constraints are placed on the utterances that the users can produce. This can be done with forced choice questions (e.g. yes or no answers) or the detection of a very restricted set of highly salient speech fragments. More flexible dialogue is possible, but is not trusted when high accuracy of understanding of the user intent is required [5]. Furthermore chains of NLP processes can incur a high computational overhead creating scalability problems for real-world deployment.

We propose a completely new method for CAs, which has the ability to reduce greatly the amount of effort and skill required in generating scripts. The new scripts will be composed of rules containing a few prototype sentences. The similarity measure is used to compute a match between the user utterance and the sentences, generating a firing strength for each rule. Because the match is on the meaning of the statement rather than words, scripting will largely be reduced to identifying the set of appropriate prototype statements.

Studies of semantic similarity to date have concentrated on one of two levels of detail, either single words [12] (in particular nouns) or complete documents [2].

Because STSS is a novel approach, there are no established methods for evaluating such measures. We expect future CAs to be used in applications where high accuracy of understanding the user intent is required, where the stakes are high and where the users may present adversarial or disruptive characteristics in the conversation. Therefore it is crucial that STSS measures are validated before being incorporated into systems, to do otherwise would be building on sand. Proper evaluation requires the use of appropriate statistical methods, the creation of standard benchmark datasets and a sound understanding of the properties of such datasets. Because semantic similarity is characterized by human perception there is no “ground truth” similarity rating that can be assigned to pairs of STs, the only way to obtain them is through carefully constructed experiments with human participants.

This paper uses the first of a group of benchmark data sets that we are creating. It describes the process of selecting a set of sentence pairs, obtaining human ratings based on the best practice from word similarity studies and comparing two machine measures, STASIS and LSA.

The rest of this paper is organized as follows: Section 2 discusses some relevant features of semantic similarity and the two machine measures; Section 3 describes the experiments which capture the human similarity ratings and section 4 describes the comparative study. Section 5 outlines directions for future work.

2 Measures and Models of Semantic Similarity

Semantic similarity is fundamental to many experiments in fields such as natural language processing, linguistics and psychology [13],[14],[15] and is held to be a widely understood concept. Miller and Charles[16], in a word-based study wrote " . . . subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty." This view, dating back to the 1960s, has been reinforced by other researchers such as Resnik [13] who observed that similarity is treated as a property characterised by human perception and intuition. There is an implicit assumption that not only are the participants comfortable in their understanding of the concept, but also when they perform a judgment task they do it using the same procedure or at least have a common understanding of the attribute they are measuring.

2.1 Relevant Features of Similarity

Empirical studies suggest that semantic similarity is a little more subtle than has been assumed. Some draw a distinction between "similarity" and "relatedness" [13], [17]. Resnik gives an example: cars and gasoline seem more closely related than cars and bicycles, but the latter pair is more similar. Although Resnik specifies semantic similarity as a special case of semantic relatedness, Charles has used relatedness to describe degrees of similarity in an empirical study [18].

Four forms of similarity are described by Klein and Murphy [19]: Taxonomic, Thematic, Goal-derived and Radial. Taxonomic similarity is the foundation of Noun similarity studies, following ISA relations through a structure such as Wordnet. Cars and gasoline are a good example of Thematic similarity (related by co-occurrence or function). Goal–derived items are connected by their significance in achieving some goal and Radial items are connected through a chain of similar items, possibly through some evolutionary process. The context in which the similarity judgment is made could result in any of the forms dominating the decision.

In some studies Semantic Distance (difference) is measured. Distance can be thought of as dissimilarity - the counterpart of semantic similarity. So if a study measures distance, it is taken as having measured similarity, by applying an inversion operation [20] or by looking for a negative correlation with distance instead of a positive correlation with similarity [16].

The concept of similarity may in itself be asymmetrical, depending on the circumstances in which items are presented. According to Tversky, "A man is like a tree" and "A tree is like a man" are interpreted as having different meanings [20]. Gleitman et al [21] claim that the structural position of the noun phrases set them as figure and ground or variant and referent, leading to the asymmetry.

Most studies use similarity measures on a scale running from 0 to a specified maximum value, typically 4. However this rating scale has no capacity to represent oppositeness (antonymy) as more different than having no similarity at all. Antonyms also generate high similarity values with co-occurrence measures [16].

The final interesting property of similarity is that given two pairs of identical items to rate, experimental participants appear to give higher rating to the pair with more complex features [13].

2.2 Models of Similarity, STASIS and LSA

A full description of STASIS is given in [1]. It calculates word similarity (equation 1) from a structured lexical database taking account of path length (l) and depth (h)

$$s(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{1}$$

This is combined with word position information from a short joint word set vector (the r terms in equation 2) and word frequency information from a large corpus to give the overall similarity:

$$S(T_1, T_2) = \delta \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} + (1-\delta) \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \tag{2}$$

The parameters α,β and δ (which adjusts the relative contributions of semantic and word order) are all chosen empirically.

A full description of LSA is given in [22]. A large rectangular matrix is created from a set of terms and documents, then decomposed into a representation as the product of 3 other matrices (equation 3).

$$X = T_0 S_0 D_0' \tag{3}$$

S_0 is a diagonal matrix. Its members are sorted by size, small elements are set to zero leaving the k largest unchanged. Zero rows and columns in S_0 are deleted, with corresponding reductions in the size of T_0 and D_0, described by equation 4 – this reduction makes LSA computationally efficient.

$$X \approx Xhihat = TSD' \tag{4}$$

LSA has the facility to compare search terms, which we have used to compare the similarity of two sentences. Each term has a corresponding row vector in $Xhihat$ and the dot product between the rows is a measure of their similarity:

$$XhihatXhihat' = TS^2T' \tag{5}$$

The important parameter k is chosen empirically, selecting a value which gives good information retrieval performance.

2.3 Desirable Properties of a Benchmark Dataset

2.3.1 Precision and Accuracy

The data set consists of judgments by human participants. Precision requires the judgments to be in close agreement with each other. Accuracy requires the derived measures to be in close agreement with the "true" similarity. Precision is affected by both the participant's internal state (mental and physical) and the measurement instrument (for example ambiguity of instructions). Accuracy depends on a common model of similarity and also on the possibility of blunders by the participant. These problems influence the design of the measurement instrument.

2.3.2 Measurement Scale

The scale on which the similarity measures are made determines the statistical techniques that can be applied to them later [23]. Human similarity measures are at least ordinal, showing reasonably consistent ranking between individuals [1], groups [24] and over time [13]. Interval scales improve on ordinal by having consistent units of measurement and ratio scales improve over interval by having an absolute zero point on the scale. Absolute scales are used where there is only one way of making the measurement: counting occurrences. Word semantic similarity has always been treated as a ratio scale attribute for both machine measures and human data sets. Our sentence data set is intended for algorithms that run from an absolute zero point (unrelated in meaning) to a maximum (identical in meaning). Setting the upper bound of the scale is common in word similarity measures and transformation of the range for comparisons is permissible.

3 Production of the Data Set

3.1 Experimental Design

Trial 1 collected ratings from 32 graduate Native English speakers to create the initial benchmark data set. The sample had a good balance of Arts/Humanities vs. Science/Engineering backgrounds. We also conducted three smaller-scale trails to investigate the importance of the randomization and semantic anchor factors as a basis for future work.

3.2 Materials

We followed the word-based procedure used in [24]. We took the first definition for 48 nouns from the Collins Cobuild dictionary [25] which uses sentence definitions derived from a large corpus of natural English. These were combined to make 65 sentence pairs in the same combinations as in [24]. Table 1 contains some examples, including two that required minor modifications to make usable sentences, *bird* and *smile*.

Table 1. Example sentence pairs derived from Rubenstein and Goodenough

Sentence pair	Cobuild Dictionary Definitions
1.cord: smile	Cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.
42.bird: crane	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A crane is a large machine that moves heavy things by lifting them in the air.
56.coast: shore	The coast is an area of land that is next to the sea. The shores or shore of a sea, lake or wide river is the land along the edge of it.
62.cemetery: graveyard	A cemetery is a place where dead people's bodies or their ashes are buried. A graveyard is an area of land, sometimes near a church, where dead people are buried.

3.3 Experimental Procedures

Each of the 65 sentence pairs was printed on a separate sheet. We randomized both the order of presentation of sentences within a pair and the order of sentence pairs within the questionnaire to minimize asymmetry and ordering effects. Participants were instructed to work through the questionnaire in a single pass. Following [24], the participants were presented with a pair of sentences and asked to "rate how similar they are in meaning." The rating scale ran from 0 (minimum similarity) to 4.0 (maximum similarity). We also included the statement "You can use the first decimal place, for example if you think the similarity is half way between 3.0 and 4.0 you can use a value like 3.5." to emphasize the linearity of the judgment. We used the Semantic Anchors in table 2, developed by Charles [18] to establish interval scale properties. Note however, that anchor 3.0 was tested but not used by Charles.

Table 2. Semantic anchors adopted from Charles

Scale Point	Semantic Anchor
0.0	The sentences are unrelated in meaning.
1.0	The sentences are vaguely similar in meaning.
2.0	The sentences are very much alike in meaning.
3.0	The sentences are strongly related in meaning.
4.0	The sentences are identical in meaning.

The full data set can be downloaded from [26].

4 Application of the Data Set

Trial 2 compared the ratings produced by STASIS and LSA with those from the human raters in the benchmark data set.

4.1 Materials and Procedure

We used a subset of the 65 sentence pairs described in section 3. This subset was the same 30 sentence pairs used in [1]. The data set contains a large number of low-similarity sentence pairs (46 pairs in the range 0.0 to 1.0), so we sampled across the low end of the range at approximately equal intervals to counter this bias. STASIS ratings were obtained directly, by running the sentence pairs through the algorithm, LSA ratings were obtained by submitting the sentence pairs through the LSA portal [27].

Table 3. Human, STASIS and LSA similarity measures for 30 sentence pairs

Sentence Pair	Human	STASIS	LSA
1.cord:smile	0.01	0.329	0.51
5.autograph:shore	0.005	0.287	0.53
9.asylum:fruit	0.005	0.209	0.505
13.boy:rooster	0.108	0.53	0.535
17.coast:forest	0.063	0.356	0.575
21.boy:sage	0.043	0.512	0.53
25.forest:graveyard	0.065	0.546	0.595
29.bird:woodland	0.013	0.335	0.505
33.hill:woodland	0.145	0.59	0.81
37.magician:oracle	0.13	0.438	0.58
41.oracle:sage	0.283	0.428	0.575
47.furnace:stove	0.348	0.721	0.715
48.magician:wizard	0.355	0.641	0.615
49.hill:mound	0.293	0.739	0.54
50.cord:string	0.47	0.685	0.675
51.glass:tumbler	0.138	0.649	0.725
52.grin:smile	0.485	0.493	0.695
53.serf:slave	0.483	0.394	0.83
54.journey:voyage	0.36	0.517	0.61
55.autograph:signature	0.405	0.55	0.7
56.coast:shore	0.588	0.759	0.78
57.forest:woodland	0.628	0.7	0.75
58.implement:tool	0.59	0.753	0.83
59.cock:rooster	0.863	1	0.985
60.boy:lad	0.58	0.663	0.83
61.cushion:pillow	0.523	0.662	0.63
62.cemetery:graveyard	0.773	0.729	0.74
63.automobile:car	0.558	0.639	0.87
64.midday:noon	0.955	0.998	1
65.gem: jewel	0.653	0.831	0.86

4.2 Results and Discussion

The human similarity measures from trial 1 are shown with the corresponding machine measures in Table 3. STASIS produces results in the range 0 to +1. LSA produced some negative results implying that it covers a true Cosine range of -1 to +1. All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 4 illustrates the agreement of both of the machine measures with human perception by calculating the product-moment correlation coefficient between the machine rating and the average rating from the human participants over the data set. We also used leave-one-out resampling to calculate the product-moment correlation coefficient for each of the human raters with the rest of the participants to establish a normative value with upper and lower bounds for performance.

Table 4. Product-moment correlation coefficients with mean human similarity ratings

	Correlation *r*	Comment
STASIS	0.816	With average of all 32 participants, significant at 0.01 level
LSA	0.838	With average of all 32 participants, significant at 0.01 level
Average Participant	0.825	Mean of individuals with group (n=32, leave-one-out resampling). Standard Deviation 0.072
Worst participant	0.594	Worst participant with group (n=32, leave-one-out resampling).
Best participant	0.921	Best participant with group (n=32, leave-one-out resampling).

Both measures have performed well with this particular data set. The normative value from the human participants (r = 0.825) sets a realistic level of expectation for the machine measures; both are close to it and LSA slightly exceeds this level. Upper and lower bounds for the expected performance are established by the performance of the best (r = 0.921) and worst (r = 0.594) performing human participants and both perform markedly better than the worst human. LSA has performed markedly better than might have been expected, despite the fact that it makes no use of word order, syntax or morphology. Plotting graphs for the subset of sentence pairs used in this test reveals STASIS is more faithful to the human ratings at low and high similarities but is brought down by its negative correlation in the similarity range 0.2 to 0.4 (r= -0.335), where LSA has a positive, if low correlation (r= 0.344). STASIS' performance in dealing with the extremes of the similarity range (particularly at the high end of the range) could be more useful in a practical CA application.

The current data set is limited to a particular type of speech act – statements (and in particular definitions). However, one of the basic tasks of a CA is to put a question to the user which can be answered with one of a number of statements. For example "What is the nature of your debt problem?" could be answered with "I am still waiting for my student loan.", "My direct debit has not gone through." Etc. This benchmark can be considered to be a form of *pons asinorum* - a measure that does not perform well will be of little use in a CA.

Overall, we consider that humans would find similarity judgements made using these algorithms to be reasonable and consistent with human judgement. However, it might be argued that if we could improve the quality of human data, we could set a more demanding target for the machine measures.

5 Future Work

We are in the early stages of a completely new study designed to give a more comprehensive set of sentence pair ratings. The current data set is limited to a particular type of speech act – statements (and in particular definitions). We will extend this to include a fuller range of relevant acts such as questions and instructions. We will draw on a variety of techniques ranging from conventional grammar to neuropsychology to construct a dataset with better representation of the semantic space and also include linguistic properties such as polysemy, homophony, affect and word frequency. Three subsidiary trials (n=18) have provided evidence of a data collection technique, card sorting combined with semantic anchors, which should reduce the noise in the data substantially. We have also gained some insights into improving the wording of the instructions to participants.

References

1. Li, Y., et al.: Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering 18(8), 1138–1150 (2006)
2. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)
3. Lapalme, G., Lamontagne, L.: Textual Reuse for Email Response. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 242–256. Springer, Heidelberg (2004)
4. Glass, J., et al.: A Framework for Developing Conversational User Interfaces. In: Fourth International Conference on Computer-Aided Design of User Interfaces, Funchal, Isle of Madeira, Portugal (2004)
5. Bickmore, T., Giorgino, T.: Health dialog systems for patients and consumers. J. Biomed. Inform. 39(5), 556–571 (2006)
6. Cassell, J., et al.: Embodied Conversational Agents (2000)
7. Gorin, A.L., Riccardi, G., Wright, J.H.: How I help you? Speech Communication 23, 113–127 (1997)
8. Graesser, A.C., et al.: AutoTutor: An Intelligent Tutoring System With Mixed Initiative Dialogue. IEEE Transactions on Education 48(4), 612–618 (2005)
9. McGeary, Z., et al.: Online Self-service: The Slow Road to Search Effectiveness, in Customer Relationship Management (2005)
10. Sammut, C.: Managing Context in a Conversational Agent. Electronic Transactions in Artificial Intelligence Volume, 191–201 (2001)
11. Michie, D.: Return of the Imitation Game. Electronic Transactions in Artificial Intelligence Volume, 205–220 (2001)

12. Resnik, P., Diab, M.: Measuring Verb Similarity. In: Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI 2000), Philadelphia (2000)
13. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
14. Prior, A., Bentin, S.: Incidental formation of episodic associations: The importance of sentential context. Memory and Cognition 31, 306–316 (2003)
15. McNamara, T.P., Sternberg, R.J.: Processing Verbal Relations. Intelligence 15, 193–221 (1991)
16. Miller, G.A., Charles, W.G.: Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
17. Viggliocho, G., et al.: Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. Cognition 85, B1–B69 (2002)
18. Charles, W.G.: Contextual Correlates of Meaning. Applied Psycholinguistics 21, 505–524 (2000)
19. Klein, D., Murphy, G.: Paper has been my ruin: conceptual relations of polysemous senses. Journal of Memory and Language 47(4), 548–570 (2002)
20. Tversky, A.: Features of Similarity. Psychological Review 84(4), 327–352 (1977)
21. Gleitman, L.R., et al.: Similar, and similar concepts. Cognition 58, 321–376 (1996)
22. Deerwester, S., et al.: Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
23. Blalock, H.M.: Social Statistics. McGraw-Hill Inc., New York (1979)
24. Rubenstein, H., Goodenough, J.: Contextual Correlates of Synonymy. Communications of the ACM 8(10), 627–633 (1965)
25. Sinclair, J.: Collins Cobuild English Dictionary for Advanced Learners, 3rd edn. Harper Collins, New York (2001)
26. O'Shea, J.D.: http://www.docm.mmu.ac.uk/STAFF/J.Oshea/
27. Laham, D.: (October 1998) (cited 30/09/2007), http://lsa.colorado.edu/

A User Interface for a Jini-Based Appliances in Ubiquitous Smart Homes

Seong Joon Lee[1], Won Tae Kim[2], and Kwang Seon Ahn[3]

[1] Department of Computer Engineering, Kyungpook National University,
1370 Sankyuk-dong, Buk-gu Daegu, South Korea
imggaibi@hotmail.com
[2] Department of Computer Engineering, Kaya University,
Jisan-ri san120, Goryeong-gun, Kyungsangpookdo, South Korea
wtkim99@kaya.ac.kr
[3] Department of Computer Engineering, Kyungpook National University,
1370 Sankyuk-dong, Buk-gu Daegu, South Korea
gsahn@knu.ac.kr

Abstract. The goal in this paper is to provide the uniform appliances interface. One of the problems with many of existing appliances may be not used the same interface for different devices, because the respective device have their heterogeneous characteristics and limitations, even among different versions made by the same manufacturer. Therefore, if end-users can have the remote controller meaning any device that is installed UI which provide the interaction between different providers and various versions, he or she is able to more efficiently control a device. In this paper, we present a system, called REMOCON, which approaches this problem by building natural language interfaces, and by using the instant messenger. In Ubiquitous smart homes environment, this mechanism can make various services to be used among service agents. Based on the mobility of service agent on the Jini platform and instant messaging system (IMS), the mobile device can receive the rapid notify from its federation leader in order to download the necessary service, to remove the unnecessary service, or to update the newly service without reconfiguring the system. Also, using the natural language processing, the client is possible to express smoothly the order. Therefore, our proposed agent-based management system for ubiquitous smart homes provides a user-friendly interface, and can make use of add-on functionalities of the IMS such as voice chatting, SMS, and multimedia.

1 Introduction

Recently, advances in Internet and digital technologies have prompted the digital convergence for helping the Ubiquitous computing environment. Particularly, in the Ubiquitous smart home, the digital convergence is caused the form of computers to change. With a rapidly increase in a variety of digital services, the homeowner have demanded to be able to control home appliances through the same method, even outdoor. Digital information of these services for monitoring and controlling IAs is becoming increasingly important to households and consumer electronics. In fact, this

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 182–191, 2008.

display and management is that something computers can do very well. However, because homeowners have to go to the computer to get anything that they wish, it is not easy to notify residents of status information of home appliances to homeowner, which occur in real-time. So, one of the key issues in the implementation of SHs has been how to efficiently control home appliances through the same method both inside and outside the home [17]. The remote control means both to monitor the appliances and to change their status on the network. However, because the respective devices have their heterogeneous characteristics and limitations, not all devices can be possessed of the same user interface (UI) [19]. It is also difficult to keep the consistency between various versions because developers add novel faculties and drop unnecessary faculties at the device. For this reason, the more various digital devices will be popularized, the more integrated efficient universal user interfaces (UUI) will be necessary. The UUI can provide the interconnectivity and interoperability between different platforms and environments. Therefore, to investigate the interactive interface for controlling appliances is worth.

In this paper, the core challenge is to develop the user interface that can provide the efficient real-time solution. To achieve this challenge, we used Jini network, instant messenger (IM), and natural language processing (NLP) technology. Jini technology is able to provide ubiquitous computing environment by using the protocols, especially discovery protocol, and federation of devices and services. IM gives homeowners a user friendly interface, and a near real-time response. Sentence and speech is universal, because it can be utilized by almost everyone. Although the field of speech recognition has made great strides in the last ten years, speech interfaces are still limited to a short sentence, and translated into sentence for the next stage. The proposed system, combining Jini technology with IM and NLP, will provide the best solution for Ubiquitous computing environment.

The other challenge is to propose a way to name each service and each method. Almost service and method names are defined by programmers. Because these formulas increase in ambiguity and complexity, integration with NLP is very difficult to realize. To solve this problem, we suggest the naming conventions. To reduce ambiguity and to cut down the complexity, we apply a minimum constraint, use a shallow NLP method, and collect synonyms using Google URL.

2 Relate Work

2.1 Previously System

Many solutions have been proposed on the technical part of the problem [13, 25]. Uniform [25], implemented on top of the Personal Universal Controller (PUC) system, is a system that is able to automatically generate appliance interfaces. In the Uniform, homeowners communicate with appliances through remote control interfaces displayed in personal portable devices, such as a PDA or mobile phone. PUC is XML-based the specification language and is able to atomically generate user interfaces and to control real appliances. It could not control appliances because The PUC communicates with devices by peer-to-peer model. And, this method cannot still apply to Jini technology because for communicating with different services, Jini make use of an object, called service proxy.

HASMuIM [13] have proposed a novel method using an IM system that is light-weight process, and supports real-time processing. This method can receive modified state information to occur at an IA without reconnecting to Home Server. One of possible disadvantages of this approach is the production of many data packets that transfer the state information changed around the home network. Also, when home-owners reconnect to the server, they may receive many such packets. These packets can be reduced by instructing the server to send the metadata list of updated IAs, i.e., the modified IA state information, only when the homeowner requires it. In the generality of cases, when homeowners interact with their home network, most are only interested in controlling one device. Therefore, downloading the state information that takes no interest is both costly and time consuming. For that reason, we have developed new, reuseable, and robust method that can install to different devices without the reconfiguration. This system do not supports to various user interfaces such as voice recognition, and multi-modal inputs.

2.2 Relate Work

In the ubiquitous computing environment, one of the various challenges is the ability to automatically register services and devices coming online on a network, and to rapidly retrieve the most reliable contextual information. To build this environment, Jini provide spontaneous networking with self-healing and administration-freeness using various key protocols, although Jini have a limitation on memory spaces. Specially, because Jini can allow downloading services, when the client invokes, from the local or the specific lookup server, it is able to provide the best solution to be suitable for any user environment.

In the paper [4], they proposed the Instant Messaging as a Platform for the Realization of a true Ubiquitous Computing Environment (IMPRUVE) platform and Instant Messenger Information Server (IMIS) that, over heterogeneous IM infrastructure, hypermedia content can be rapidly and efficiently transferred to a designated resident. They builds IM Connectivity API providing the integrated environment for different messengers, controls disparate appliances or devices registered through MAPS (Mobile and Pervasive System) in IMIS. MAPS are the subsystem with the ability to automatically register services and devices coming online on a network. They showed that IM have the potential as UI for ubiquitous computing environment.

The Natural Language Processing (NLP) is that computer process human language and returns a reasonable response that people expect. Typically, these methods accept a sentence, generate appropriate parsing trees, and translate it into defined structures. Question Answering (QA) is defined to be an information retrieval (IR) paradigm to map users Natural Language (NL) questions into the most reliable answers. When different technology is designed together with NLP, user interface can be simple, provide powerful uniformed UI tools, and easy way to give the operation that user wants. With the progress of the ubiquitous computing environment, the adoption of NLP including voice processing will be continuously increased.

3 Proposed Technique

In this paper, we intend to develop an efficient real-time solution for home automation systems by using Jini network, instant messenger (IM), and natural language processing (NLP) technology. The proposed system provides a uniform graphic UI (GUI) for end users by using IM. It immediately sends state information from the home network to the homeowner via the Internet. Homeowners with the appropriate IM installed on their mobile phones can remotely control and monitor their home appliances, anytime and anywhere. Furthermore, the system is lightweight, flexible, and extensible. The proposed system is convenient because it can sends a completion or alarm message to a manager via push functionality (NOTIFY message), even if the manager does not reconnect to the home network after it finishes setting the device. The proposed system makes use of a simple sentence for any intention of homeowner, while a web-based interface require hierarchically clicked menus.

To achieve this system, there is three main agents in this implementation: the Mobile Messenger Agent (MMA), the Home Messenger Agent (HMA), and the Information Appliance Manager Agent (IAMA). The overview of the REmote Monitoring and CONtrol method of home appliances (REMOCON) is shown in figure 1.

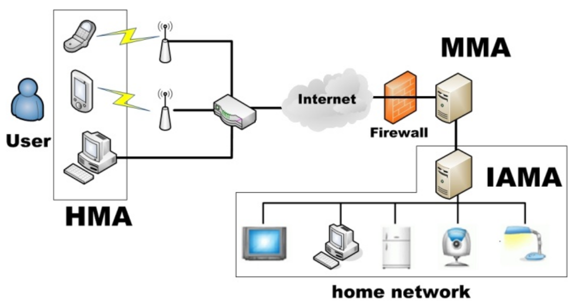


Fig. 1. Overall architecture of REMOCON

3.1 Assumptions

We assume that the home network uses the network address translation (NAT), and that the residential gateway uses ADSL with dynamic IP addressing. Recently, in the world, almost household Internet environment make use of Ipv4 and DHCP. NAT is quite well known for protecting the home server from a trespasser using the firewall provided naturally, and for providing a practical solution to the impending exhaustion of Ipv4 address space. IM is able to avoid the problem of the DMZ by using the private IP

We assume that the proposed system use sentences for controlling IAs or services. However, NLP is difficult to determine the semantic meaning of a sentence, and to make the analysis of the context to determine the true meaning and comparing that with other text, For reducing some of these problems, each sentence is assumed to have only one service object (or IA) and various attribute values. i.e., we do not consider instructions that involve multiple services.

3.2 Definition of Methods for the IAs

In the ubiquitous computing environment, devices shall be operated without Authenticated user's intervention, and replaced and added ones should not be harmful the whole system. Therefore, we need to control devices using the same method. However, almost all service and method names are defined by programmers. In the result, without any agreement, it is impossible to change and insert a product. These methods are divided into three types: direct control, indirect control, and request command.

In the case of a direct control, the method name is the verb describing that action. In the case of indirect control, to modify a member variable, the method name is the concatenation of 'set' and the full name of the member variable (e.g., setAngle, setTemperature). Similarly, when a user wants to find out the value of a member variable, the method name is the concatenation of 'get' and the full name of the member variable (e.g., getTemperature, getVolume). Table 1 lists the interface classes of example appliances.

Table 1. The List of the interface classes

Service	Method
Lamp	turnOn()
	turnoff()
Boiler	switchOn()
	switchOff()
	setTime(Calendar dateTime)
	Calendar getTime()
	setTemperature(float degrees)
	float getTemperature

3.3 Mobile Messenger Agent

The MMA is the agent that is executed on a personal computer or the homeowner's portable devices, such as a PDA, mobile phone, or smart phone, and is used to indirectly control home appliances connected to the home network. Because the proposed method is based on sentences, the MMA can use a GUI of a general IM system However, in legacy systems, there is no cryptographic algorithm for protecting text messages. To solve this problem, the proposed system is designed to encode messages with the Advanced Encryption Standard (AES) method announced by National Institute of Standards and Technology (NIST) as U.S.

This key is delivered from the home server once the client's messenger is started. AES is one of the most popular algorithms used in symmetric key cryptography. And it is fast in both software and hardware, is relatively easy to implement, and requires little memory. As a new encryption standard, it is currently being deployed on a large scale.

3.4 Home Messenger Agent

The HMA is the agent connected in home network. The role of this agent services which the MMA and the IAMA in home network are connected. Also, it manages the homeowners including any additions to the homeowner group. When the administrator is connected by wired network device in home, manager function is activated. We have presented the HMA agent to create the Order Document XML code using 5Ws and 1H. Figure 7 shows OD XML document created by the example sentence.

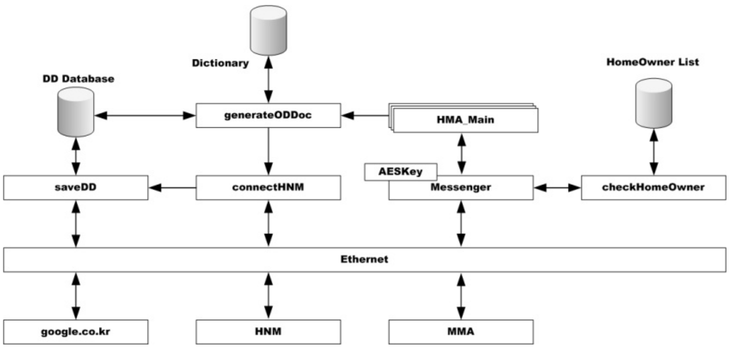


Fig. 2. The architecture of home messenger agent

When the message is delivered to HMA, the messenger function deciphers the message, transfer the plaintext with a tag to HMA_Main. HMA_Main preserve the connection with the MMA, deliver one to the generateODDoc. The generateODDoc create the parse tree from this text using the lexicalized probabilistic parser [14], and then generates the OD XML code. The following figure 5. is the translated process.

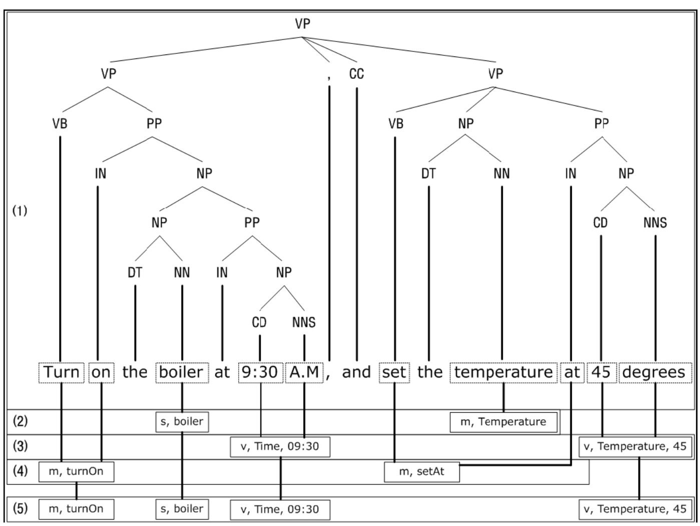


Fig. 3. Parse tree for the plaintext

For example, if the plaintext is "turns on the boiler at 9:30 AM, and set the temperature at 45 degrees," then generate the following step (1) in figure 6. In step (2), a NP+PP structure is detected, and finds 'boiler' as a service name and temperature as a method name. The service name is stored in the variable, and the method name is added to the method temporary. From the PP that has CD, generateODDoc decides that the domain of '9:30 AM' is a time method, and '45 degrees' is a temperature method, respectively. Using these detected methods, the method temporary is updated. To eliminate overlap in method names, it is deleted the record that has the repetitive method name, and attribute 'm' (step (3)). The generateODDoc finds VB tags, and put the IN in the adjacent PP tag together, then checks whether VB belongs to any of these methods. In other hands, join on to Turn, search method table for 'turnOn'. If detected, this word is added to the method temporary (step (4)). The other way, this method is disregarded because 'setAt' is not detected. Finally, generateODDoc is filled out the OD document (step (5)).

```xml
<?xml version="1.0" encoding="UTF-8"?>
<orderDiscription                    xsi:noNamespaceSchemaLocation="orderDocument.xsd"
xmlns:xsi="http://www.w3 org/2001/XMLSchema-instance">
        <who>
                <id>mggaibi@ ....</id>
                <ip>'55.230.......</ip>
        </who>
        <where>house</where>
        <what>Boiler</what>
        <when>sch</when>
        <how>
                <method>
                        <methodName>setTime</methodName>
                        <parameter>
                                <pName>"2006. 10. 19 9:30:00"</pName>
                        </parameter>
                </method>
                <method>
                        <methodName>turnOn</methodName>
                </method>
                <method>
                        <methodName>setTemperature</methodName>
                        <parameter>
                                <pName>45</pName>
                        </parameter>
                </method>
        </how>
</orderDiscription>
```

Fig. 4. The OD XML document created by the example sentence

3.5 Information Appliance Manager Agent

The IAMA is to create the new object necessary for connecting with middleware based on OD document transmitted from the HMA, and to monitor the appliance. Because Jini is centralized System, to download a certain service in the different device must be connected to Lookup server. To complete the proposed system, the IAMA is designed so as to include the Lookup server of Jini to directly manage the Lookup table. So, all IAs, when firstly connecting to Lookup server, is made into OD documents by the *createODD* method called by the modified register function, the *registerAppliance*, in the IAMA, and store local storage device. For checking a message, these documents should also send to HMA.

If a problem is encountered at the object created by the *manageControl* method in IAMA, it informs the homeowner of an urgent error message. Otherwise, if the task is completed, it must send a result message to the homeowner. These processes must be performed as quickly as possible. The Fig. 5 show the architecture of IAMA.

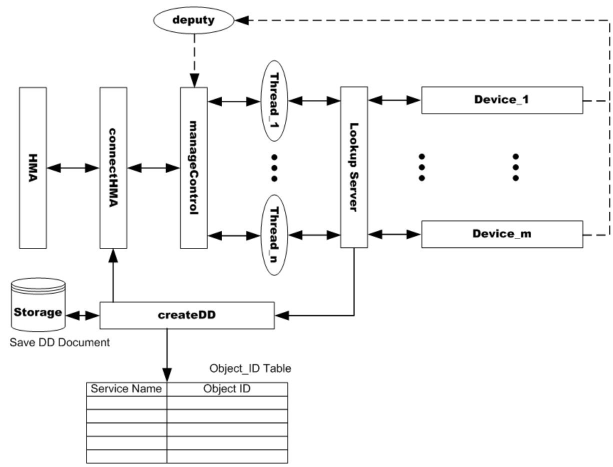


Fig. 5. The Architecture of Information Appliance Manager Agent

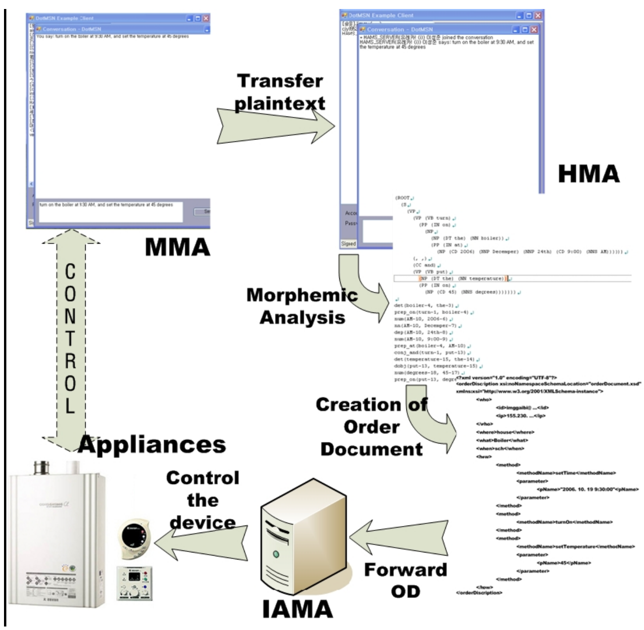


Fig. 6. Prototype system setup and experimental results

4 Implementation

The aim of this research is to develop a system for controlling and monitoring home appliances on a smart home onto which an IM was installed. To achieve this environment, we have need of each agent which must operate well even if the internal structure or the programming environment of the other two agents is changed. Therefore, we did not used technologies based on specific programming environments except JINI, so as to operate in the heterogeneous programming environments as well as platforms. The process of the proposed system is shown in the figure 6. To implement this system, we used Jini middleware to establish an efficient home network and MSN Messenger to connect to the home server from an outdoor area. The proposed system is implemented using Java for HMA and IAMA, C sharp for MMA, XML to write OD document, and MySQL for the DD Database. The HMA is porting in the Embedded Board with Intel Xscale PXA270 and embedded Linux (kernel 2.6.11), and IAMA is running on Windows XP with Java. The MMA is PDA, smart phone, portable phone, or PC.

5 Conclusions

We have described a remote monitoring and control method of home appliances combining with NLP, called REMOCON. This paper extends the meaning of a Jini service to a set of services. The proposed system is characterized by five features: first, by using IM that is one of the most popular applications on the Internet, this paper gives homeowners user-friendly interface, and secondly, the ability to make use of dynamic IP address. Finally, by using input commands written in natural language by the user, homeowners can naturally command. Moreover, input commands could potentially be given verbally by using a voice recognition interface. The next step in this study is a more detailed user interface research and evaluation.

References

1. Andrenucci, A., Sneiders, E.: Automated question answering: Review of the main approaches. In: ICITA, vol. (1), pp. 514–519 (2005)
2. Dewes, C., Wichmann, A., Feldmann, A.: An analysis of internet chat systems. In: IMC, pp. 51–64. ACM Press, New York (2003)
3. Dobrev, P., Famolari, D., Kurzke, C., Miller, B.A.: Device and service discovery in home networks with osgi. Communications Magazine, IEEE 40(8), 86–92 (2002)
4. Foley, C., Downes, B., de Leastar, E., van der Meer, S.: Instant messaging as a platform for the realisation of a true ubiquitous computing environment. In: eChallenges (2005)
5. Ha, T., Jung, J., Oh, S.: Method to analyze user behavior in home environment. Personal and Ubiquitous Computing 10(2-3), 110–121 (2006)
6. HAVi Organization. The HAVi specification (version 1.1) (2001), `http://www.havi.org/`
7. Hypothetic.org. MSN protocol documents (2003), `http://www.hypothetic.org/docs/msn/index.php`
8. Ivanov, R.S.: Controller for mobile control and monitoring via short message services. In: TELSIKS, pp. 108–111 (2003)

9. Jeronimo, M., Weast, J.: UPnP Design by Example: A Software Developer's Guide to Universal Plug and Play. Intel Press (2003)
10. Koskela, T., Väänänen-Vainio-Mattila, K.: Evolution towards smart home environments: empirical evaluation of three user interfaces. Personal and Ubiquitous Computing 8(3-4), 234–240 (2004)
11. Kumaran, I., Kumaran, S.I.: Jini Technology: An Overview. Prentice Hall PTR, Upper Saddle River (2001)
12. Lea, R., Gibbs, S., Dara-Abrams, A., Eytchison, E.: Networking home entertainment devices with havi. IEEE Computer 33(9), 35–43 (2000)
13. Lee, S.J., Ahn, K.S.: Control of Information Appliances Using Instant Messaging. In: Sha, E., Han, S.-K., Xu, C.-Z., Kim, M.-H., Yang, L.T., Xiao, B. (eds.) EUC 2006. LNCS, vol. 4096, pp. 976–986. Springer, Heidelberg (2006)
14. LI, H.Y.: Web-based remote monitoring and control for process plants. In ICMLC, pp. 936–941 (2005)
15. Liang, N.S., Fu, L.C., Wu, C.L.: An integrated, flexible, and internet-based control architecture for home automation system in the internet era. In: ICRA, pp. 1101–1106 (2002)
16. Cogkun, I., Ardam, H.: A remote controller for home and office appliances by telephone. Consumer Electronics, IEEE Transactions on 44(4), 1291–1297 (1998)
17. Lucenius, J., Suomalainen, J., Ventola, P.: Implementing mobile access to heterogeneous home environment. In: HOIT (2003)
18. Miller, B.A., Nixon, T., Tai, C., Wood, M.D.: Home networking with universal plug and play. Communications Magazine, IEEE 39(12), 104–109 (2001)
19. Nikolova, M., Meijs, F., Voorwinden, P.: Remote mobile control of home appliances. Consumer Electronics, IEEE Transactions on 49(1), 123–127 (2003)
20. Omojokun, O., Dewan, P.: Experiments with mobile computing middleware for deploying appliance uis. In: ICDCS Workshops, pp. 340–375 (2003)
21. OSGi Alliance. Osgi Service Platform (release 4) (2007), `http://www2.osgi.org/Specifications/HomePage`
22. Rahman, M., Bhattacharya, P.: Remote access and networked appliance control using biometrics features. Consumer Electronics, IEEE Transactions on 49(2), 348–353 (2003)
23. The Stanford Natural Language Processing Group. The Stanford Parser: A statistical parser (Version 1.5.1) (2006)
24. Yasumura, M., Yoshida, R., Yoshida, M.: Prototyping and Evaluation of New Remote Controls for People with Visual Impairment. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 461–468. Springer, Heidelberg (2006)
25. Nichols, J., Myers, B.A., Rothrock, B.: UNIFORM: Automatically generating consistent remote control user interfaces. In: CHI, pp. 611–620 (2006)

A Study of Meta-Search Agent Based on Tags and Ontological Approach for Improving Web Searches

Jae-Youn Jung, Hyeon-Cheol Zin, and Chonggun Kim*

Dept. of Computer Engineering, Yeungnam University
214-1 Dea-dong, Gyeongsan-si, Gyeongsangbuk-do 712-749, Korea
e-mail@ynu.ac.kr, cyberzin@ynu.ac.kr, cgkim@yu.ac.kr

Abstract. Recent increase in interest for information ranking and sharing among users with similar tastes has urged many researches towards improving relevance of search results for reducing costs and offering better quality service to users of Web search engines. Our work has focused largely on the ranking and sharing schemes of retrieved information among heterogeneous sources, whereas Web search engines need to wide crawler with just speed in short time. In this paper, we propose a meta-search agent with a URL filter, a tag-based ranking scheme, and an ontology-based sharing scheme. The meta-search agent uses vector tags to facilitate the definition of informative value and finally the maintenance of shared information. We introduce a concept of vector tag that shows a conceptual distance between retrieval interest and search results. We also compare performance of the proposed system with hyperlink-based methodologies, and analyze the pros and cons of each.

1 Introduction

The emergence of the Internet and its applications has fundamentally altered the environment in which information agencies conduct their missions and deliver services. Recently, the progress of mobile technologies has changed the some way people access Internet information. Users do not have the time to evaluate multiple search engines to knowledgeably select the best for their uses. Nor do they have the time to submit each query to multiple search engines and wade through the resulting flood of good, duplicated, or irrelevant information, and missing Web contents. To satisfy the information requirements of users, many information services have been customized for various users' interests.

This paper proposes a system that searches Web documents based on link information and vector tags and shares the search results based on an ontological approach. We can expect more quality results using link structure and more personalized results utilizing the vector tag and ontological sharing approach for more user satisfaction. Clearly, interoperability of vector tags is a key factor to make it work. The idea behind our ranking method is based on different vector tags and vector components for appropriate ranking. Quite often, many web sites successfully spam

* Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 192–202, 2008.

some of the search engines and obtain an unfair rank. By using appropriate rank aggregation strategies, we can prevent such results from appearing in the top results.

The rest of this paper is organized as follows. In section 2, we discuss prior research in the area of search agents and practical paradigm in that context. In section 3, we present the methodology and describe the potential conflicts that would need to be resolved in order to arrive at common understanding. In section 4, we present effectiveness and usefulness of the proposed strategies for Web searches as well as mechanisms for ranking to these alternate methodologies. Finally section 5 concludes the paper.

2 Related Works

Retrieval of relevant information from the Internet is a time-consuming and laborious task. Getting precise information while searching Web can be a difficult task. Since retrieval involves the analysis and separation of interesting pages from a great set of candidate pages [6]. Today's search engines are the most widely used tools for searching and retrieving information from the Internet. General purpose search engines not always fulfill user requirements, sometimes because of the lack of information about user interests. Users provide a set of key words to a search engine which in turn returns a set of links to Web pages that include those key words [11]. It is well known that users of Web search engines tend to use short queries that in fact consist of only one or two words [11; 17]. But short queries are often non-specific in nature; the results retrieved by the similarity-based method can thus bring very poor quality. This causes users to spend considerable amount of time in searching and retrieving quality information. In considering these problems, Web search personalization has been discussed from various viewpoints [1; 15].

Search result must be the most reliable site that people expect. A large-scale Web search engine solves the problem of slow speed by computing the importance of web documents before searching takes place [3]. An approach distills a large search topic on the WWW down to a size that makes sense to a human user [18]. It identifies authoritative and hub sources about user query. While authoritative and hub sources are calculated using link information, authoritative sources are the most reliable web site about specific topics and hub sources are documents that link to many authoritative sources [13].

Many agents have been developed in order to give a more appropriate solution to the problem of searching the Web [10; 16]. Several approaches have been proposed for describing user interests by receiving different kinds of feedback. Generally, this feedback is explicit and consists of a direct rating that the user makes of specific contents [5]. Explicit relevance feedback could be annoying for users and that is why most recently developed approaches have tried to avoid it [2; 4].

Some studies [14; 19] suggest systems of content labeling and filtering for schemes of Web search. Their suggestions relevant to search results can provide a guide on how to block unuseful sites. The relevant search results on Web can be built on reconsidering the semantic approach from three directions: total architectural design, detailed specification, and deployment plan.

Search engines use ranking algorithms to order the list according to a certain criteria. Alternatively, [12] proposed to learn the ranking function from clickthrough data (the logfile of results the users clicked on) of a group of users. An aggregation of the results obtained would be more useful than just dumping the raw results. For such an aggregation, [8] suggested a local Kemenization technique, which brings the results that are ranked higher by the majority of the search engines to the top of the Meta search-ranking list. In establishing an information-sharing platform, many technologies have been used. Technologies such as COM (Component Object Model), CORBA (Common Object Request Broker Architecture), and ontological agent have also been developed to solve the technical problem of information sharing [9].

With the rise of the Internet and increasing use of digital networks it has become easier to work in both an informal and ad hoc manner, and as a part of a community. Users should not be forced to use specific search engine, and the most important, search intent including their occupation and social position is a key factor in search ranking. Thus future Internet search requires new approaches.

3 The Framework and Development of Proposed System

In this section, we present an architectural design and mechanism of our meta-search agent (as shown in the Figure 1) for increasing performance on Web search. Especially, we focus on a ranking scheme using vector tags and a sharing scheme using ontological approach for accessing relevant information.

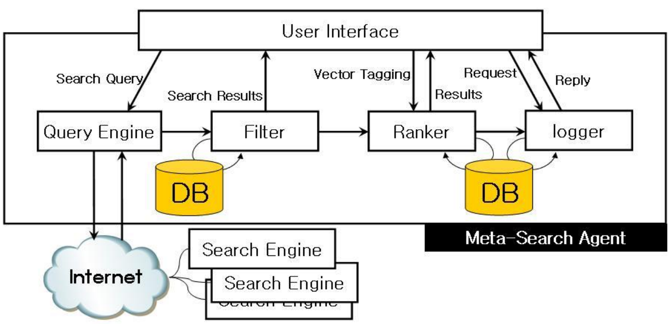


Fig. 1. The proposed Meta-Search Agent architecture

3.1 System Architecture

We first put a brief description of the meta-search task that a user is trying to search using the proposed meta-search agent. The major components of the proposed system are: 1) user interface, 2) query engine, 3) filter, 4) ranker, and 5) logger. We discuss below each component and how these components relate to each other.

The user interface is a primarily used to input the query phrases for searching task. The query engine receives users' queries from the user interface and determines how query phrases are going to be considered as a match with previous. Also, the query engine sends out queries to the several selected search engines and collects the results via Internet. The filter performs a weeding routine to eliminate irrelevant pages as duplicated Web documents from the set of documents returned by the underlying source search engines. In order to approve primary search results, we propose a ranking scheme with tags introduced by a principle of position vector. Precedence users annotate each document with vector tags, and then they can get search results in ascending or descending rank order. The ranker re-rank documents with the relevance vector value, and then the user interface browse returned results from the ranker. All of the browsing results on the user interface are kept at the logger. The logger is a database for sharing search histories of similar users' interests. For effective sharing among users, the ontology of user is deployed to categorize users' histories in the logger.

3.2 Mechanism of Pattern Filtering

This task of the filter proceeds on pattern matching the URL address of each document. The architecture of the filter (URL pattern filter) is shown in Figure 2.

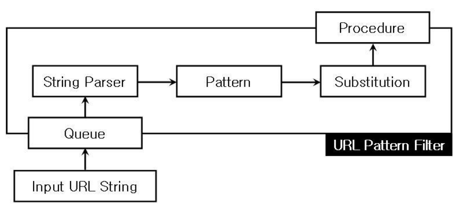


Fig. 2. URL Filter architecture of the proposed system

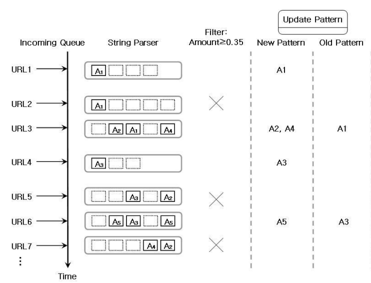


Fig. 3. Mechanism of URL filtering under Filter: Amount≥0.35

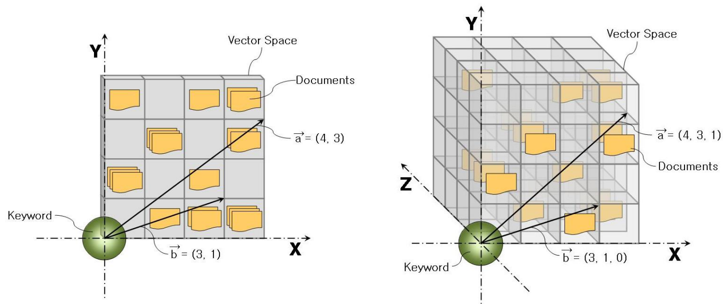
The filter eliminates duplicate documents from same source or with similar address from the returned search results via the query engine. URLs pattern elements are extracted from incoming URLs, parsed by the string parser, and matched among these extracted patterns in the subcomponent of pattern and substitution. The mechanism of URLs pattern filtering is shown in Figure 3. In order to explain how filtering works, we assume that a filtering amount sets "Filter: Amount≥0.35" as shown in Figure 3. The value of ≥0.35 means that the filter allows remaining search lists to be below 35% matching. With this filter, any withdrawn URLs that have an amount of pattern matching more than 35% do not enter the search lists and are therefore not passed to update patterns. Users can see the search lists of the remaining good pages based on the relevance pattern provided by URLs filtering results.

3.3 Principle of Rank Using Proposed Vector Tag

After filtering, the user interface shows search results returned by the underlying filter to all users. For approving of these primary search results, we suggest a vector tag method as a good ranking scheme. The proposed vector tag consists of both a keyword and a vector component. The syntax for the vector tag is as follows:

KEYWORD [X, Y, Z]

The KEYWORD, in the above syntax, expresses user's aim and becomes a criterion of categorization for classification in the logger. The vector component [X, Y, Z] represents the position of an object in space in relation to an arbitrary inertial frame of reference, and referred to as a location point that exist in a 3-dimensional space (Figure 4b). It is also used as a means of deriving displacement by the spatial comparison of distances from the keyword which is an origin of vector and represents user intention (as shown in Figure 4a and 4b).



a. 2-dimensional vector tag b. 3-dimensional vector tag

Fig. 4. Conceptual structure to define with the vector tag

The vector space of Figure 4a explained by a Cartesian coordinates system with X and Y axis represents a position vector — X axis is rank of importance based on recommendation, citation, etc; Y axis is time of information that is an indicator for creation time. In advanced, we extend the 2-dimensional vector space to Z axis direction as Figure 4b — Z axis is reference value introduced by others' evaluation for the same Web content. We give that the 3-dimensioanl vector tag (Figure 4b) is therefore more effective and detailed expression for informativeness than the 2-dimensional one (Figure 4a). A vector tag based on position vector can be used to represent any quantity that has both a magnitude and a direction. Vectors are usually shown in graphs or other diagrams as arrows, as illustrated Figure 4a and 4b. The length of the arrow makes known the vector's magnitude which means a conceptual distance between search intention and search results. So a short arrow expresses near distance with search intention, and stands for more usefulness than a long one. In Figure 4b, for an example, the vector $\vec{b}$ displays more valuable information than the vector $\vec{a}$.

A position vector of tag is made in vector space as Figure 4b. Vector's magnitude (length) R for quantitative comparison is defined with equation (1), and nearer information to search intent (origin of vector represented by keyword) has small cost. Also, weight value according to bias of α, β, and γ can be controlled.

$$R = \sqrt{\alpha x^2 + \beta y^2 + \gamma z^2} \tag{1}$$

If several users tag same search result, one document has several vector tags. Each Vector's magnitude *R* of these tags has to be accumulated, then informative value (*V*) can be written as equation (2) according to estimate cost of individual's [X, Y, Z].

$$V = \frac{\sum_{i \in T_N} \sqrt{X_t^2 + Y_t^2 + Z_t^2}}{N_{total}} \tag{2}$$

In equation (2), where T_N shows total tags which can be classified by the same category, *i* is one of total tags in a category. A tag has an eigenvalue with X_t, Y_t, and Z_t . X_t has some *x* of component's value defined at some point of importance value; Y_t has some *y* of component's value defined in concurrence of time value; Z_t has some *z* of component's value defined by other users' evaluation. The magnitude of position vector that is worked by *x*, *y*, and *z* displays informative value (*V*). The summation of value in equation (2) divides by whole tags (N_{total}) to represent mean value being accumulated.

3.4 Ontological Approach for Sharing Scheme of the Logger

The logger, a kind of database, maintains all of the browsing results on the user interface for sharing histories of Web search with similar users. For effective sharing, a request to detect similar user induce a practical suggestion as ontological approach. Especially, in order that user information could be understood and processed by machine automatically, a linkage of both the logger and not existing databases but more semantic descriptions is needed. Ontology has become a promising technology to express semantics [7].

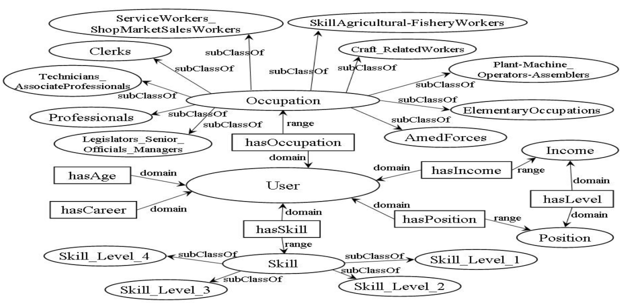


Fig. 5. A framework of User ontology

In building domain ontology people should explore the rationality in expressing domain concepts and relations between the concepts, considering the capabilities of OWL (Ontology Web Language). In Figure 5, we put forward a framework of User ontology. For the framework of User ontology, we have considered capabilities to express semantic relations provided by OWL [7]. OWL classes are interpreted as sets that contain individuals. They are described using formal descriptions that state precisely the requirements for membership of the class. Properties are binary relations on individuals. Individuals represent objects in the domain that we are interested in.

In Figure 5, classes are illustrated in ellipses and properties are in rectangle. OWL is property-centricity, and each property has domain and range. User is defined as a class. The domain of the six properties (hasAge, hasCareer, hasSkill, hasPosition, hasIncome, hasOccupation) is the class User. The range of the property hasOccupation is the class Occupation. The range of the property hasSkill is the class Skill. The range of the property hasPosition is the class Position. The range of the property hasIncome is the class Income. The domain of the properties hasLevel is the class Income and the class Position. The class Occupation has ten subclasses: Legislators_Senior_Officials_Managers, Professionals, Technicians_Associate Professionals, Clerks, ServiceWorkers_ShopMarketSalesWorkes, SkillAgricultrual-FisheryWorkers, Craft_RelatedWorkers, Plant-Machine_Operators-Assemblers, ElementraryOccupations, and AmedForces. Also, the class Skill has four subclasses: Skill_Level_1, Skill_Level_2, Skill_Level_3, and Skill_Level_4. A particular class may have a restriction on a property that at least one value for that property is of a certain type [7]. In the environment of Protégé, we have built OWL ontology of user based on Figure 5. OWL provides additional vocabulary along with a formal semantics as disjointWith, intersectionOf, unionOf, complementOf, oneOf, and relationOf.

4 Performance Evaluation for the Proposed System

4.1 Effectiveness of Pattern Filtering

We conducted experiments to evaluate the Filter implemented in the component of the proposed meta-search agent. For accomplishing our evaluation of filtering effectiveness, we compare search results (with filtering) of our system with that (without filtering) of Google.

Table 1. Search results by Google

No.	U R L	Remark
1	http://www.terms.co.kr/ontology.htm	
2	http://www.dal.kr/blog/archives/001010.html	
3	http://humbroll.springnote.com/pages/344340	
4	http://complingone.georgetown.edu/~prot/	
5	http://bklab.snu.ac.kr/blog/kwangsub/45	
6	http://europa.snu.ac.kr/index.php/Some_issues_of_ontology_building_and_collaboration	
7	http://bric.postech.ac.kr/topic/63.htm	
8	http://osdir.com/ml/misc.ontology.protege.general/2002-10/msg00025.html	
9	http://iems.net/votal/index.php?cat=215	
10	http://osdir.com/ml/misc.ontology.protege.general/2002-10/msg00024.html	*
11	http://bklab.snu.ac.kr/~kskim/wp/?cat=5	
12	http://shannon.springnote.com/tags/ontology	
13	http://shannon.springnote.com/pages/501678	*
14	http://libterm.springnote.com/pages/541582	
15	http://chord.snu.ac.kr/blog/kwangsub/44	
16	http://mar.gar.in/brandon/ontology&mp=b&msort=datedesc	
17	http://www.blogweb.co.kr/?p=31	
18	http://www.blogweb.co.kr/?cat=3	*
19	http://report.empas.com/search/index.hcam?type=total&qu=ontology	
20	http://gamedev.springnote.com/pages/467242	

Table 2. Search results by the proposal

No.	U R L	Remark
1	http://www.terms.co.kr/ontology.htm	
2	http://www.dal.kr/blog/archives/001010.html	
3	http://humbroll.springnote.com/pages/344340	
4	http://complingone.georgetown.edu/~prot/	
5	http://bklab.snu.ac.kr/blog/kwangsub/45	
6	http://europa.snu.ac.kr/index.php/Some_issues_of_ontology_building_and_collaboration	
7	http://bric.postech.ac.kr/topic/63.htm	
8	http://osdir.com/ml/misc.ontology.protege.general/2002-10/msg00025.html	
9	http://iems.net/votal/index.php?cat=215	
10	http://bklab.snu.ac.kr/~kskim/wp/?cat=5	
11	http://shannon.springnote.com/tags/ontology	
12	http://libterm.springnote.com/pages/541582	
13	http://chord.snu.ac.kr/blog/kwangsub/44	
14	http://mar.gar.in/brandon/ontology&mp=b&msort=datedesc	
15	http://www.blogweb.co.kr/?p=31	
16	http://report.empas.com/search/index.hcam?type=total&qu=ontology	
17	http://gamedev.springnote.com/pages/467242	
18	http://mar.gar.in/ironpark/ontology&mp=b&msort=datedesc00	*
19	http://coreonto.kaist.ac.kr/	*
20	http://ontology.etnews.co.kr/	*

Table 1 shows top 20 of search lists for query "ontology" provided by Google at Oct. 28th 2007, and the Table 2 shows top 20 of search lists for the same query with our system at the same date. In Table 2, we can see #18, #19, and #20 instead of #10,

#13, and #18 of the Table 1. The documents of #10, #13, and #18 in Table 1 have similar URLs patterns of #8, #12, and #17, respectively. So our system filters these documents for exclusion in the collection of the search results.

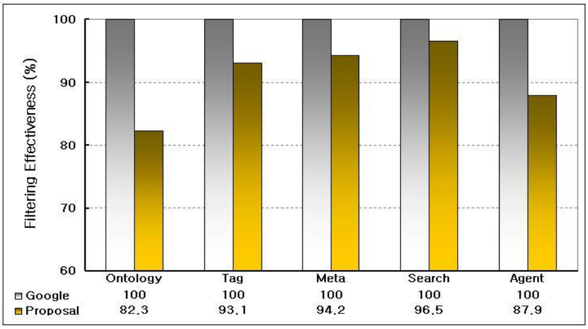


Fig. 6. Comparing results with (proposal) and without (Google) filtering

We performed experiments of filtering effectiveness for five example queries “ontology”, “tag”, “meta”, “search”, and “agent” on Web. Figure 6 shows individual filtering precision for each 1,000 crawling documents using example queries, where we use the filtering mechanism of URLs pattern matching. The X axis is the five queries with five pairs of columns which represent the percentage of filtering beside the right-contrast column (without filtering returned by Google). From this figure, we can see that the performance depends on the search results returned by Google and the proposed system. The average of filtering performance is 9.2% in the returning results for all queries. For some queries such as “ontology” and “agent”, the filtering results are heavily shown to 17.7% (remaining 82.3%) and 12.1% (remaining 87.9%) in comparing the both systems. As the results, our system seems to be effective in applying small browsers of handheld devices to achieve for fairly good filtering.

4.2 Appropriate Ranking and Time Complexity

We have chosen precision as a good measure of reasonable ranking in evaluating the appropriate performance for the document set as shown in Table 3 relating to molecular works. All the ranking work on the same set of results and try to get the most relevant ones to the top. Hence, good information that has a higher precision at the top can be rated better from the user’s prospect. Between 2 ranking results obtained by each of the proposed method (based on tag) and PageRank (based on hyperlink), we compare the ranking results processed by the both.

The Table 4 shows valid evaluation of comparison with both the proposed method and PageRank. The ranking results of two methods have wide deviations. Our interesting observation is a warning of exclusion in the collection of the Table 3 for disputed documents (document # 1, 4, 9, 16, and 19). Carrying out disputed documents seems to give a kind of effectiveness for preventing spam.

Table 3. The list of 20 documents for experimental results

No.	URL
1	http://cat.inist.fr/?aModele=afficheN&cpsidt=4181477
2	http://openwetware.org/wiki/Miniprep/Kit-free_high-throughput_protocol
3	http://people.morehead-st.edu/fs/d.peyton/protocols.html
4	http://sosnick.uchicago.edu/DNA_miniprep.html
5	http://userwww.service.emory.edu/~kressle/protocols/BAC%20miniprep%20protocol.doc
6	http://wolverton.owu.edu/lab/2006/01/eppendorf-miniprep-protocol
7	http://www.bioinformatics.vg/Methods/miniprep.shtml
8	http://www.bio.brandeis.edu/haberlab/jehsite/pdfs/ZymoPrep.pdf
9	http://www.bio.indiana.edu/~chenlab/potocols/qiagenmini.pdf
10	http://www.bio.net/hypermail/chlamydomonas/1993-December/000121.html
11	http://www.genetics.ucla.edu/labs/fan/Protocols_Miniprep.htm
12	http://www.genome.arizona.edu/agi/seq/QIAprep%20Spin%20Miniprep%20Kit%20Protocol.doc
13	http://www.genomed-dna.com/pdf/Quick-PDFs/Protocol%20Quick-Plasmid%20(Vacuum).pdf
14	http://www.gerardbiotech.com/documents/Protocols/GerardBiotech_HurricaneMiniPrep250_gbt010705.pdf
15	http://www.mbi.ufl.edu/~rowland/protocols/minipreps.htm
16	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7946306&dopt=Abstract
17	http://www.protocol-online.org/prot/Molecular_Biology/Plasmid/Miniprep/more2.html
18	http://www.tracy.k12.ca.us/thscdna/miniprep.html
19	http://www.umich.edu/~wakil/protocols/miniprep.html
20	http://www1.qiagen.com/HB/QIAprepMiniprep

Table 4. Comparison of ranking eva luation between PageRank and ours

Document #	PageRank	Proposed Rank	Accruacy	Remark
1	16	–	1.00	*
2	17	2	0.83	
3	18	1	0.89	
4	10	–	1.00	*
5	5	14	0.74	
6	11	9	0.77	
7	4	3	0.80	
8	19	4	0.86	
9	2	–	0.97	*
10	7	15	0.71	
11	14	7	0.77	
12	12	12	0.83	
13	15	13	0.83	
14	20	10	0.77	
15	1	8	0.80	
16	13	–	1.00	*
17	3	5	0.80	
18	9	6	0.71	
19	8	–	1.00	*
20	6	10	0.86	

In comparison with two results of the Table 4, we experimentally clarified the accuracies of ranking by the proposed method. Our ranking of approval rating has 79.8% agreement. And also, the decision-making process for exclusion of disputed documents has 99.4% consensus.

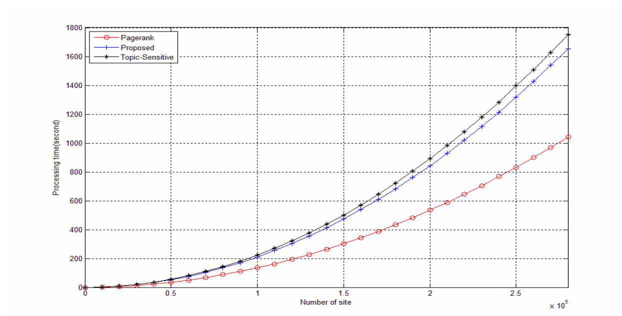


Fig. 7. Experiments for the time complexity of three algorithms (PageRank, Topic-Sensitive PageRank, and the proposal)

We performed experiments for the time complexity of PageRank, Topic-Sensitive PageRank, and the proposal as shown in Figure 7. In which the X axis stands for the number of results returned from real Web data based on Stanford's matrix [20], and the Y axis is the time spent for processing in the whole system. Figure 7 shows that our proposed method slightly outperforms than Topic-Sensitive PageRank. On other side, in comparing with PageRank, the two methods (the proposal and Topic-Sensitive PageRank) are more precipitous than PageRank because of the proposal and Topic-Sensitive ranking requested more additive calculations.

5 Conclusions

The vector tag method and pattern filtering enable Web search to increase performance. Future Web browsers seem to need coordination with a broker, software as a meta-search agent, for more powerful handling of Web contents. Our proposed vector tag is a core idea for brokers. Page titles, headers, image, and content include the vector tag into themselves in order to help optimize visibility and ranking in condition of Web browsers within limited devices or manners. It is important not to over use keywords, or hide them within inappropriate alt attributes or invisible text, as this could be perceived as spamming, resulting in exclusion by search engines. We provide a framework of User ontology for effective sharing search histories. Our exploration might not standardize User ontology. We explore with an intention to help Web search better understand the rationality in expressing concepts and relations between the concepts in user categorization. In order that proposed User ontology relating to other domain ontology is constantly improved and standardized as possible.

Acknowledgments

This research was conducted as part of National IT Ontology Infrastructure Technology Development project funded by Ministry of Information & Communication of Korean government.

References

1. Amandi, A., Godoy, D.: PersonalSearcher: An Intelligent Agent for Searching Web Pages. In: Monard, M.C., Sichman, J.S. (eds.) SBIA 2000 and IBERAMIA 2000. LNCS (LNAI), vol. 1952, pp. 43–52. Springer, Heidelberg (2000)
2. Balabanovic, M.: An Interface for Learning Multi-topic User Profiles from Implicit Feedback. In: AAAI 1998. Workshop on Recommender Systems (1998)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: The 7th International WWW Conference (1998)
4. Chen, L., Sycara, K.: WebMate: A Personal Agent for Browsing and Searching. In: Proc. of the 2nd Intl Conf. on Autonomous Agents and Multiagent Systems (1998)
5. Chen, Z., Meng, X., Zhu, B., Fowler, R.: Web Sail: From on-line learning to web search. In: Proc. of the 1st Intl Conf. on Web Information Systems Engineering, vol. 1, pp. 206–213 (2000)
6. Convey, J.: Online Information Retrieval. Library Association Publishing, London (1992)
7. McGuinness, D.L., et al.: OWL Web Ontology Language Overview. W3C (2004)
8. Dwork, C., Kumar, R., Naor, M., Siva Kumar, D.: Rank Aggregation Methods for the web. In: Proceedings of the Tenth WWW Conference (2001)
9. Filman, R.E., Korsmeyer, D.J., Lee, D.D.: A CORBA extension for intelligent software environments. Advances in Engineering Software, Eng. 31(8), 727–732 (2000)
10. Goldman, C., Langer, A., Rosenschein, J.: Musag: An agent that learns what you mean. Applied Artificial Intelligence 11 (1997)
11. Jansen, B., Spink, A., Bateman, J., Saracevic, J.: Real life information retrieval: A study of user queries on the web. SIGIR Forum 32(1), 5–17 (1998)
12. Joachims, T.: Optimizing Search Engine using Clickthrough Data. In: Proc. of the ACM Conference on Knowledge Discovery and Data Mining, ACM, New York (2002)
13. Kleinberg, J.: Authoritative sources in a hyperlinked environment. IBM RJ 10076 (1997)
14. Klyne, G., Reynolds, F., Woodrow, C., Ohto, H., Hjelm, J., Butler, M.H., Tran, L. (eds.): Composite Capability/Preference Profiles: Structure and Vocabularies 1.0. W3C (2004)

15. Ligon, G., Balachandran, M.B., Sharma, D.: Personalized search on electronic information. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3684, pp. 35–41. Springer, Heidelberg (2005)
16. Mladenic, D.: Learning word normalization using word suffix and context from unlabeled data. In: ICML 2002, pp. 427–434 (2002)
17. Silverstein, C., Henzinger, M., Hannes, M., Moricz, M.: Analysis of a very large Web search engine query log. SIGIR Forum 33(3), 6–22 (1999)
18. The Clever Search, http://www.almaden.ibm.com/cs/k53/clever.html
19. The SFS Browser: A Demonstration System That Realizes Filtering of Inappropriate Content Accessed with Mobile Phones, Internet Association Japan (2003)
20. The Stanford Web Matrix, http://nlp.stanford.edu

An Agent Using Matrix for Backward Path Search on MANET

SungSoo Lee, ZhiPeng Liu, and ChongGun Kim†

Department of Computer Engineering, Yeungnam University
lssung@chol.com, tallentbull@hotmail.com, cgkim@yu.ac.kr

Abstract. With the dynamic and mobile nature of ad hoc networks, links may fail due to topology changes. So, a major challenge in ad hoc networks is to search path from a source to a destination with an efficient routing method, which is also an important issue for delay-sensitive real-time application. A topology of the nodes interconnections in an ad hoc network can be modeled as an adjacency matrix. In this paper, we present a knowledge discovery agent to collect topology information and derive the adjacency matrix. Based on a sequence matrix calculation, we propose an effective routing algorithm Matrix-Based Backward path Search (MBBS) for route discovery from a destination to a source. The performance of MBBS and that of Breadth First Search (BFS) algorithm are compared. The research result proves that MBBS is more efficient in recognizing the route.

Keywords: a knowledge discovery agent, Ad hoc network, adjacency matrix, connectivity matrix.

1 Introduction

Since ad hoc network does not rely on existing infrastructure and is self-organized, the nodes on such network act as hosts and routers to transmit packets. With its frequent change in topology, ad hoc network does not rely on pre-established wired network, it requires special routing algorithm. Routing protocols used in ad hoc network can be divided into two categories: table-driven routing method and on-demand routing method [1] [9] [10].

Recent research in graph theory for communications is more intensive than ever. As a result, some new efficient algorithms have been studied for the parallel or concurrent data processing. However, the main concerns of graph theory for ad hoc networks are which nodes are connected to a given node, what path between two nodes is the least expensive. So, the algorithm for determining the graph connectivity (generally implemented by systematic traversal of graph beginning at a source node) can be a critical issue in the research of ad hoc networks.

A mobile ad hoc network is characterized by multi-hop wireless links. With the dynamic and mobile nature of ad hoc networks, links may fail due to topology

† Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 203–211, 2008.

changes. In ad hoc networks, the supported services are not only delay-tolerant, such as FTP and email, but also delay-sensitive real-time applications. Therefore, an efficient routing to search path from a source to a destination is very important in ad hoc networks.

A topology or the nodes interconnections in an ad hoc network can be modeled as an adjacency matrix. Several works have been conducted in order to find network topology using connectivity matrix in ad hoc networks. In [2], J. Zhang and W. Seah propose a matrix select-delete algorithm for calculating the network capacity of MANETs. In [3], N. Li, et al present a matrix-based fast calculation algorithm in terms of the maximum number of k-hop adjacency matrix of the network. All the above works concentrate on estimating network capacity, but in this work, we focus on the path discovery and propose a new path discovery algorithm.

Based on this adjacency matrix, we propose new path search algorithm through a sequence of matrix calculation. This algorithm can recognize backward path from a destination to a source using connectivity matrix.

The remainder of this paper is organized as follows. Section 2 introduces the related works of the connectivity matrix. In the section 3, we propose a knowledge discovery agent to form the connectivity matrix of network. Section 4 presents a new path search algorithm using connectivity matrix. Finally, the simulation results and conclusion are given in sections 5 and 6, respectively.

2 Network Model and Connectivity Matrix

The topology of an ad hoc network is modeled by an undirected graph G (V, A). V denotes the node set in the network and A is an adjacency matrix that describes the topology of the network. The adjacency (one-hop connectivity) matrix $A=\{a_{ij}\}$ for an N-node network, in which entry "1" at (i, j) indicates a connection from node i to node j and entry "0" at (i, j) indicates no connection from node i to node j, can be manipulated to obtain the connectivity matrix $C= \{c_{ij}\}$, for which the entry at (i, j) lists the minimum number of hops needed to connect node i to node j [4].

Connectivity matrix C is defined as follows:

First, $C^{(1)} = A$.

$$C^{(1)} = A. \quad \textbf{(1)}$$

Secondly, the elements of the square of matrix A, which we denote $A^2 = \{a_{ij}^{(2)}\}$ can be written:

$$a_{ij}^{(2)} = \sum_{k=1}^{N} a_{ik} a_{kj} = \begin{cases} 0, & \textit{no path} \\ > 0, & \textit{at least one path} \end{cases}. \quad \textbf{(2)}$$

The modification of $a_{ij}^{(2)}$ is given by:

$$b_{ij}^{(2)} = \begin{cases} 0, & i = j \\ 0, & a_{ij} = 1 \\ 2, & a_{ij}^{(2)} > 0 \quad when\ i \neq j\ and\ a_{ij} = 0 \\ 0, & otherwise\ . \end{cases} \tag{3}$$

And thirdly, $c_{ij}^{(2)}$ is found as

$$c_{ij}^{(2)} = a_{ij} + b_{ij}^{(2)} = c_{ij}^{(1)} + b_{ij}^{(2)}\ . \tag{4}$$

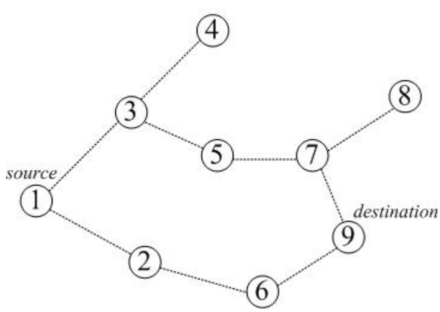
The general rule for the intermediate calculations can be stated as follows:

$$c_{ij}^{(m)} = c_{ij}^{(m-1)} + b_{ij}^{(m)}, \quad m \geq 2 \tag{5}$$

Where,

$$b_{ij}^{(m)} = \begin{cases} 0, & i = j \\ 0, & c_{ij}^{(m-1)} > 0 \\ m, & \sum_{k=1}^{N} c_{ik}^{(m-1)} a_{kj} > 0 \quad when\ i \neq j\ and\ c_{ij}^{(m-1)} = 0 \\ 0, & otherwise\ . \end{cases} \tag{6}$$

Figure 1 shows a network topology (A) and its corresponding adjacency matrix (B). Figure 2 shows 2-hop matrix (A) and connectivity matrix (B) given by matrix multiplication.



(A)

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

(B)

Fig. 1. Network model (A) and its corresponding adjacency matrix (B)

$$A^{(2)} = \begin{bmatrix} 0 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 2 \\ 1 & 2 & 0 & 1 & 1 & 0 & 2 & 0 & 0 \\ 2 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 2 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 & 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 & 0 & 1 & 0 & 2 \\ 0 & 2 & 0 & 0 & 2 & 1 & 1 & 2 & 0 \end{bmatrix}$$

(A)

$$C = A^{(4)} = \begin{bmatrix} 0 & 1 & 1 & 2 & 2 & 2 & 3 & 4 & 3 \\ 1 & 0 & 2 & 3 & 3 & 1 & 3 & 4 & 2 \\ 1 & 2 & 0 & 1 & 1 & 3 & 2 & 3 & 3 \\ 2 & 3 & 1 & 0 & 2 & 4 & 3 & 4 & 4 \\ 2 & 3 & 1 & 2 & 0 & 3 & 1 & 2 & 2 \\ 2 & 1 & 3 & 4 & 3 & 0 & 2 & 3 & 1 \\ 3 & 3 & 2 & 3 & 1 & 2 & 0 & 1 & 1 \\ 4 & 4 & 3 & 4 & 2 & 3 & 1 & 0 & 2 \\ 3 & 2 & 3 & 4 & 2 & 1 & 1 & 2 & 0 \end{bmatrix}$$

(B)

Fig. 2. 2-hop adjacency matrix (A) and connectivity matrix (B)

Algorithm to compute connectivity matrix from adjacency matrix is as follows:

1. Define two *N* * *N* matrices B and C, in addition to the *N* * *N* adjacency matrix A.
2. Initially, set C = A and B = 0.
3. For m=2 to the longest hop distance (or until the update matrix B equals zero):
   1) For all node pairs (*i, j*) = (1, 1) to (N, N):
      A. If *i* = *j*, skip to the next node pair
      B. If $c_{ij} > 0$, skip to the next node pair
      C. For k = 1 to N
         a. if $C_{ik} > 0$ and $a_{kj} > 0$ for some k,
            i. Set $b_{ij} = m$
            ii. Exit the loop and go to the next node pair (*i, j*)
   2) Set C ← C + B, the B=0:
4. At the end of these calculations, C equals the connectivity matrix and the sum of all the elements of C, divided by N(N -1), equals the average hop distance.

Fig. 3. Algorithm to compute connectivity matrix from adjacency matrix

3 The Agent for Topology Information

3.1 Creating Adjacency Matrix

The agent node gathers information of the network topology and broadcasts it to all nodes. The algorithm using adjacency matrix decreases overhead on table-driven routing. Figure 4 shows the configuration of topology discovery and distribution agent. With the adjacency matrix, a source node can easily recognize the existing route to a destination node.

All nodes periodically broadcast hello message in order to check the connectivity with their neighbor nodes [5] [11]. For collecting the information about network topology, the agent node broadcasts query message to all neighboring nodes. Then, the node receiving topology query message delivers it to its neighbor nodes. Therefore, the node flooding the topology query message becomes a parent-node; the receiver becomes a child-node. The node without child-node is a leaf node. On receiving the query message, the leaf node responds by sending topology reply message including node ID, connection and battery life-time information to its parent-node. At the end of above process, the agent node can obtain all nodes connectivity information in the network. Based on the connectivity information, the adjacency matrix of the network can be formed.

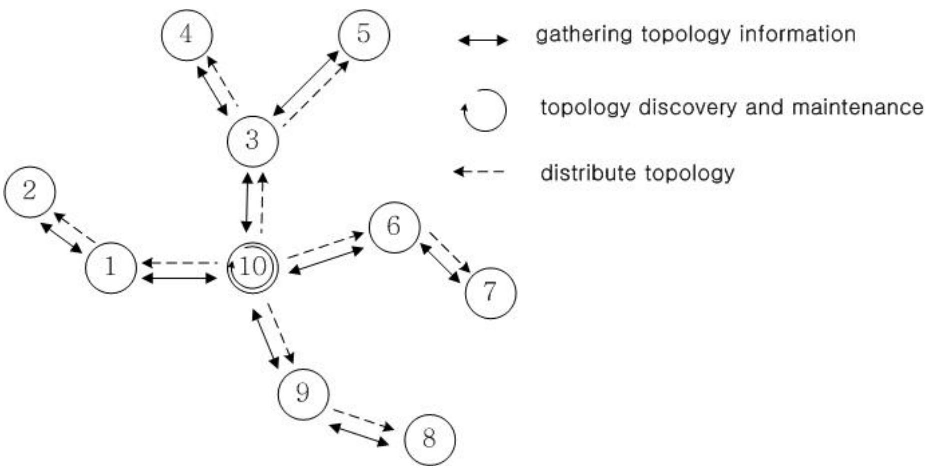


Fig. 4. Topology discovery and distribution agent

In case the agent node disappears from the network, new agent node is selected based on all nodes connection and battery-life time information [6]. Moreover, the new agent node sends adjacency matrix and new agent node information via topology advertisement message to all nodes. In this way every node in the network can receive the connectivity information of the network. An arbitrary node can be the agent node when the network is initially created. Therefore, in this initial case, selection process for the proper agent node becomes a main duty.

3.2 Agent Selection within Nodes

Initially the agent node is selected by the following rule:

a. Leaf-node can not be the agent node.

b. The agent node should have longest battery life-time.
c. The agent node should have as many as possible neighbor nodes.

The agent selection algorithm is based on the leader election algorithm [7]. We maintain a list of candidate agents instead of just one agent in every node. Each node contains an agent list of five nodes (in descending order), where the first node is considered as the active agent of the network. If the active agent becomes leaf-node by network topology change or its battery life-time drops below the threshold, it hands over the role of the agent node to the second agent node.

When the agent node disappears, the highest-valued node from the collected agent list is selected as the substitute node [8].Network partition and merging can also be managed efficiently. When network divides into two separate networks as active node disappears, the network which has the next active agent does not modify the existing agent list. Instead the network which has not active agent node, after the time out of agent nodes checking, initiator node sends the heartbeat message to all nodes of the network. To reduce the waiting time, all nodes update their agent lists by shifting the invalid agent to the end of the lists. Nodes keep the invalid agent node in the agent list for a while for use in future possible rejoin of the networks.

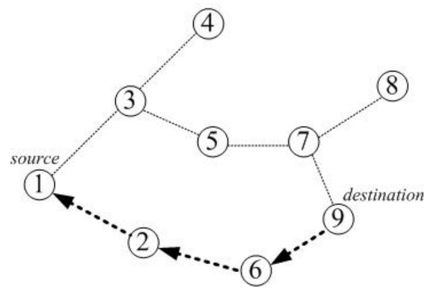
When there are more than two networks and they maintain separate agent lists, all nodes in the merged network select the highest-valued agent node as the active agent node of the network.

4 Matrix-Based Backward Path Search Algorithm

By the connectivity matrix, BFS (Breadth First Search) or DFS (Depth First Search) algorithm can be adopted to search path from a source to a destination. But in this paper, we propose a matrix-based backward route discovery algorithm which recognizes route from a destination to a source using connectivity matrix.

Connectivity matrix can be obtained with the method mentioned in Section 2. In order to decrease the complexity of calculating connectivity matrix, we calculate connectivity matrix only until the hop count between the destination and source node is obtained. We introduce a list of hop-set from a source to a destination for our algorithm. Each hop-set list includes all nodes which are the same hops from the source. First, if the hops number is k from a source to a destination, the algorithm searches k-1 hop-set list. Second, if there are neighbor nodes of destination in k-1 hop-set list, the algorithm searches its sequent neighbor nodes in *k-2* hop-set list. Then process is repeated until the source node is located.

In figure 5, the destination node 3 is 3-hop from the source node 1. With proposed algorithm, we first search destination's neighbor node in 2-hop-set list and find that node 6 is only 1-hop from the destination node. Although node 7 is also a destination's neighbor node, it is 3-hop from the source, so node 7 is excluded. And then, the 1-hop set can be traversed; as a result, node 2 is found as a neighbor node of node 6 and 1-hop from node 6. Therefore, the path is 1-2-6-9.



(A)

source	1-hop	2-hop	3-hop
①	2, 3	4, 5, 6	7, ⑨

(B)

Fig. 5. Matrix-based backward path search (A) and hop-set list (B)

The algorithm is as follows:

1. Create connectivity matrix C, and input destination hop count , k;
2. Store all n-hop nodes from the source in hop-set list (n = 1, 2, · · ·, k-1);
3. While (k ≠ 0)
    1) Select all adjacency nodes which are 1-hop distance from the destination in k-1 hop-set;
    2) Store the node in the path-set;
    3) k = k-1;
    4) The adjacency node becomes a new destination;
4. Output path-set;

Fig. 6. Algorithm of matrix-based backward path search

5 Performance Evaluation and Simulations

In this section, we evaluate the performance of MBBS algorithm and compare the results with that of BFS algorithm.

Theoretically, since the number of shortest paths should not be bigger than the minimum of number of source's and destination's neighbor nodes, our algorithm is supposed to be more efficient than algorithms which search path from a source to a destination. The simulation results verify our assumption.

In order to verify the efficiency of our algorithm, we perform a variety of simulations based different hop counts between the source and destination node. The simulation results are presented based on 64-node arbitrary network topology and the total numbers of traversal nodes versus hop counts are obtained by the simulation.

Figure 7 shows simulation results. If hop count from destination to source node is 4, the number of traversal node is 10 in MBBS algorithm comparing with 40 in BFS algorithm. BFS algorithm has to traverse all nodes in the range of destination hop count. The proposed MBBS algorithm needs to traverse less than 50% of nodes in the range of destination hop count.

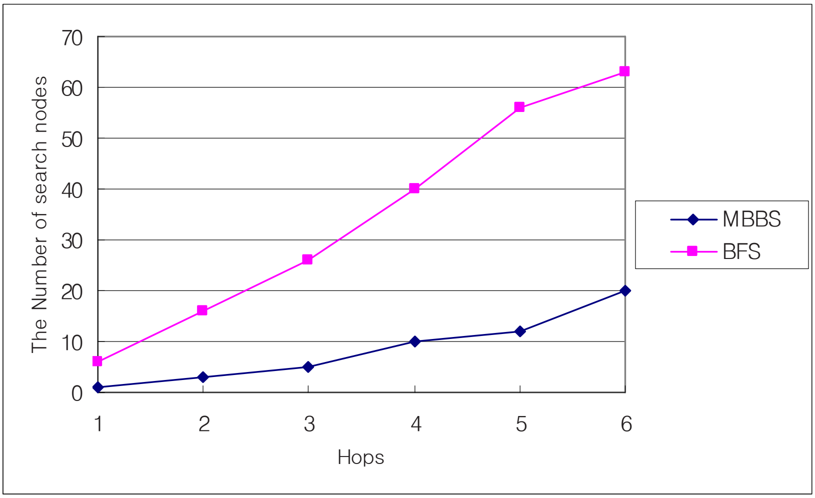


Fig. 7. Number of Total search node

6 Conclusions and Future Works

In this paper, we propose a knowledge discovery agent to form the adjacency matrix of network. Based on the adjacency matrix, we present a new algorithm, Matrix-Based Backward path Search (MBBS), to recognize the route from the destination to the source. The performance evaluation and simulations results show that the proposed algorithm is more efficient in finding route than Breadth First Search (BFS) algorithm.

Further research is to design concrete protocol and compare its performance with the existing ad hoc network routing protocols.

Acknowledgements

This research was financially supported by the Ministry of Commerce, Industry and Energy (MOCIE) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation.

References

1. Lu, Y., Wang, W., Bhargava, Y.Z.B.: Study of Distance Vector Routing Protocols for Mobile ad hoc networks. In: Proceedings of the First IEEE International Conference (2003)
2. Zhang, J., Seah, W.K.G.: Topology-based Capacity Analysis for Ad Hoc Networks with End-to-End Delay Constraints. In: Proc. of IEEE 6th CAS Symposium on Emerging Technologies: Mobile and Wireless Communications, Shanghai, China, pp. 541–544 (2004)
3. Li, N., Guo, Y., Zheng, S., Tian, C., Zheng, J.: A Matrix-Based Fast Calculation Algorithm for Estimating Network Capacity of MANETs. In: ICW/ICHSN/ICMCS/SENET 2005, pp. 407–412 (2005)
4. Miller, L.E.: Multihop connectivity of arbitrary networks (March 2001), http://w3.antd.nist.gov/wctg/netanal/ConCalc.pdf
5. Kim, H.-K., Kim, C., Lee, S.: Design of Knowledge Discovery Agent for a Bit-Map on Ad Hoc Mobile Networks. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 738–746. Springer, Heidelberg (2007)
6. Vasudevan, S., Kurose, J., Towsley, D.: Design and Analysis of a Agent Election Algorithm for Mobile Ad Hoc Networks, UMass Computer Science Technical Report 03-20
7. Vasudevan, S., Kurose, J., Towsley, D.: Design and Analysis of a Agent Election Algorithm for Mobile Ad Hoc Networks. In: Proceedings of the 12th IEEE International Conference on Network Protocols (ICNP), pp. 350–360 (2004)
8. Lee, S., Muhammad, R.M., Kim, C.: A Leader Election Algorithm within candidates on Ad Hoc Mobile Networks. In: Lee, Y.-H., Kim, H.-N., Kim, J., Park, Y.W., Yang, L.T., Kim, S.W. (eds.) ICESS 2007. LNCS, vol. 4523, Springer, Heidelberg (2007)
9. Perkins, C.E., Bhagwat, P.: Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers. In: Proceedings of ACM SIGMOMM (1994)
10. Clausen, T., Jacquet, P.: A Optimized link state routing protocol. RFC3626, Experimental (October 2003)
11. Chakeres, I.M., Belding-Royer, E.M.: The utility of hello messages for determining link connectivity. In: The 5th International Symposium on vol. 2 (October 2002)

A Personalized URL Re-ranking Methodology Using User's Browsing Behavior*

Harshit Kumar[1], Sungjoon Park[2], and Sanggil Kang[3]

[1] Department of Computer Science, University of Suwon, Hwaseong, Gyeonggi-do, South Korea
harshitkumar@suwon.ac.kr
[2] Department of Mobile Game, Kongju Communication and Arts College, Janggi-myon, Kongju City, Chungnam, South Korea
sjpark@kcac.ac.kr
[3] Computer Science and Information Engineering, Inha University, 253 Younhyun-dong, Nam-gu, Incheon, South Korea
sgkang@inha.ac.kr

Abstract. This paper proposes a personalized re-ranking of URLs returned by a search engine using user's browsing behaviors. Our personalization method constructs an index of the anchor text retrieved from the web pages that the user has clicked during his/her past searches. We propose a weight computation method that assigns different values to anchor texts according to user's browsing behaviors such as 'clicking' or 'downloading'. Experiment results show that our method can be practical for saving surfing time and effort to find users' preferred web pages.

Keywords: Personalization, Search Engine, Search API, Re-ranking.

1 Introduction

Search engines have been extensively used by the people to search for the information. The most popular search engines used these days are Google, Yahoo, MSN, and Naver (Korean Search Engine). The vast majority of queries to search engines are short and ambiguous, and different users may have completely different information needs and goals under the same query. However, a search engine always returns the same set of results; for instance, a web search on "java" returns a wide variety or range of results, which includes java island, java programming or java coffee, etc.

Google, Yahoo, and MSN have already proposed their personalization methods [1]. However, they have a couple of drawbacks such as: they need explicit input of user profile/interest, do not consider that a user's interest might change over the time, and group interest might not match individual interest in all cases. In order to solve the

* This research was supported by the Ministry of Information and Communication, Korea, under the College Information Technology Research Center Support Program, grant number IITA-2006-C1090-0603-0031.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 212–221, 2008.

problem, it is needed to automatically capture user's interest and provide personalized results based on the interests.

In this paper we propose a methodology that captures a user's browsing behavior, aimed at overcoming the above limitations. Our method dynamically constructs an index of the anchor text retrieved from the web pages that the user has clicked during his/her past searches. We assign different 'weight' to the extracted anchor text from the web pages clicked or downloaded by the user. After collecting sufficient anchor text, our method uses the weight of the anchor text to re-rank the search results from a search engine. The result from the experiment shows that the average rank reduces and test of significance is positive. The use of link texts was previously advocated by Davison [20] and Eirol et al. [15].

The rest of the paper is organized as follows: Section 2 discusses about the related work. Section 3 introduces our system. Experimental results and conclusions are presented in section 4 and 5 respectively.

2 Related Work

The following section presents related work on constructing user profile and personalization.

2.1 Infer User Interests

Automatic constructing user profile usually deals with observing user behavior. To our best knowledge, no previous work uses anchor text to model user interests. Kelly and Teevan [2] review several possible approaches for inferring user preferences. In the paper user behavior has been categorized across many dimensions such as examine, retain, and reference etc. In another research work by Agichtein et al., authors [3] have organized the user interests as a set of features. Another approach [4] by Shen et al. collects user-interest based on clicked document summaries, title, and click through history and query history that are accumulated over a session. One session is defined as a period consisting of all interaction for the same information need. This approach has one major drawback which is due to spamming. Sometimes authors of web pages deliberately add popular terms in the title to gain high relevance. Our approach is also based on click through history but we accumulate anchor text to build user interest. Teevan et al. [5] and Chirita et al. [6] use user's desktop to estimate their interests and construct his/her profile. The limitation of these approaches is that there can be lots of terms on user's desktop, which can be quite noisy or misleading for estimating user interests.

Das et al. [1, 7] use collaborative filtering (CF) for personalization. The underlying assumption of CF approach is that those who agreed in the past tend to agree again in the future. The drawback of this approach is that the group interest might not match the individual interest.

Another possible approach to personalization can be done by explicitly asking user profile. The user profile is then used to filter search results by checking content similarity between returned URLs and user profile. Early version of Google personalization asked the user to select the categories of topics of interest. Google search engine applied this information to filter search results. The inherent limitation of this approach is that the user profile may change over the time. Also studies have shown that users are quite reluctant to provide explicit input of their profile or any explicit feedback on search results. Using user's web browser cache, user profiles have been constructed by classifying web pages into appropriate concepts in the reference ontology [8, 9]. In this scenario, the user profile is constrained within the available concepts in ontology.

2.2 Personalization

The scientific literature offers various solutions to personalization of search engine. To achieve personalization researchers have either used query expansion [6, 8, 9] or filtering query results [10, 11, 12]. In query expansion user's interests are conflated with the query. For filtering query results, researches define their own formulae for filtering returned URLs and further re-ranking of the final result set. Agichtein et al. [13] have used their own measure and supervised machine learning technique for re-ranking search results. In a very recent work [14], authors have used search engine logs for constructing user profiles. They re-rank the search results by computing a personalized score for each URL returned on a query. They introduced four formulas for re-ranking: two methods closely relate to collaborative filtering and the other two relate to personal level.

3 An Overview of Our System

Fig. 1 is an outline of our system that provides personalized re-ranking of URLs. On the client side, a user issues a query and chooses a search engine from the available four options (Google, Yahoo, MSN, and Naver). The returned search

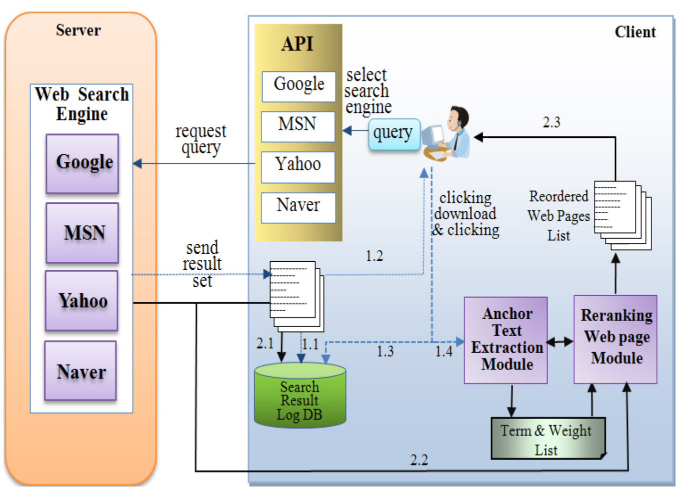


Fig. 1. Overview of our system

results (set of URLs) are logged along with the query and user *ID*. If the user clicks or downloads a URL, the system logs the selected URL along with the query and user *ID*. The anchor text extraction module extracts anchor text from the documents clicked or downloaded by the user. The extracted anchor texts are logged and used for re-ranking search results in the re-ranking web page module. The detailed description is presented in the following sections.

3.1 Compute Anchor Text Weight

A search engine returns a list of URLs (denoted as set U) having relevance with the user query. The URLs clicked by the user are denoted as a set V ($V \subseteq U$). We believe that the URLs that a user clicks or downloads are related to his/her interest. It has been reported by [15] that there is a similarity between search queries and anchor text. They also showed that anchor text is a succinct description of a web page. Therefore we extract and create index file of anchor text from the URLs clicked or downloaded by the user.

The order of web pages in a searched list indicates importance or relevance of the web pages for the query. For instance, a high ranked URL is highly relevant with the query. The anchor text in a high ranked URL can also be considered highly relevant to the query. Taking the ranking of URL into consideration, we assign the weight to each anchor text according to the rank of URL containing the anchor text and store it as an index. User behavior such as downloading and clicking a URL is used as a measure of the user's interest. Anchor text weight w_i is initialized by the rank of the URL(clicked by the user) that contained the anchor text. If a user action is downloading, we enforce the value of weight w_i by α because the downloading action is a strong indication of user interest. If an anchor text already exists in the indexed file and the same anchor text appears in another web page, the weight w_i of that anchor text is calculated as shown in equation (1). The accumulated anchor text over a period of time represents user interests. These are further used for re-ranking search results for a new query.

$$w_i = \sum_j (N - R_{i,j} + 1) / N + \alpha \tag{1}$$

In the above equation, N is the cardinality of set U, $R_{i,j}$ ($j \in N$) is the rank of the jth web page containing anchor text i. From Equation (1), the value of w_i increases exponentially as the number of high ranked web pages increases. To avoid the problem, we normalize keyword weight w_i with the log sigmoid function as shown in Fig. 2. The value of α is 0 for clicking and 2 for downloading, which was determined by empirical experiment. In general an anchor text can be extracted from multiple web pages. So, the value of w_i increases as the number of high ranked URLs containing anchor text k_i increases. The justification of equation (1) follows.

Let's assume that for a query q, a search engine returns 10 results, denoted as $U=\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$. The user clicks on u_1, u_3, u_{10}, i.e. denoted as set $V=\{u_1, u_3, u_{10}\}$. We assume that anchor text a_1, a_2, and a_1 appear in URLs corresponding to u_1, u_{10}. Let w_1 be the weight of anchor text a_1 which is computed below. Then w_1 = (10 – 1 + 1)/10 + (10 – 10 + 1)/10 = 1 + .1 = 1.1. The first component of the equation evaluates to 1 and the second component evaluates to .1.

The first component appears from higher ranked web page, hence large weight is provided. In the meantime, the second component appears from the lower ranked web page, hence lower weight is provided. For the convenience, we normalize the weight using the function as seen in Fig. 2. The variable *s* in the figure determines the slope of the function. If *s* is large then the slope is steep, otherwise gentle. The $\tilde{w}_i$ is the normalized value of w_i.

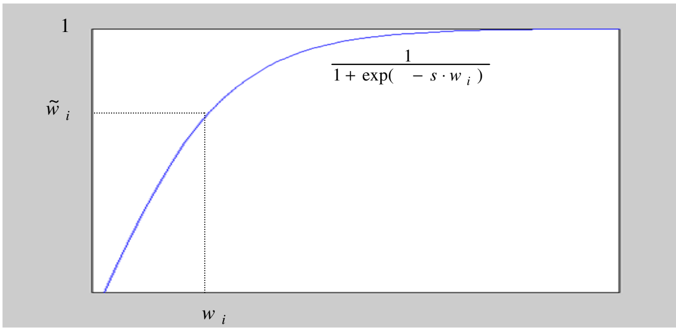


Fig. 2. The function of normalization of weight

3.2 Re-ranking Search Result

Using the index file of anchor texts along with their weights, we will re-rank the search engine results (*U*). The new rank denoted as $\tilde{R}_j$ of web page *j* is computed using the linear combination as in (2).

$$\tilde{R}_j = (1-\beta)R_j + \beta \sum_{k_i \in K_j} \tilde{w}_i \tag{2}$$

where K_j is the set of anchor text included in web page $j \in$ U and R_j is the original rank of web page j from the search engine. Since we respect the original rank of a URL returned by the search engine, we take the linear combination of the original rank during the computation of new rank. The parameter β has a value between 0 and 1. If β has a value of 0, new rank is same as original rank. If β has a value of 1, the original ranking is ignored. The variable β allows user to choose whether he/she wants personalized results or not. We consider the top 20 returned URLs for personalization, it has been shown in Speretta et al [9] that 90% of the click happen in the first page i.e. top 10 URLs. However, this would incur a considerable time whose evaluation is an interesting future work. It is interesting to note that the process of determining weight of anchor text is an offline process which can be carried out after showing the requested URL clicked by the user whereas the re-ranking process is performed before showing the returned URLs. This could have a significant effect as search engines generally return results in a fraction of second. We would like to evaluate how much is the time efficiency of re-ranking process and how can we improve it?

4 Experiment

We collected 50 volunteers' browsing behavior over a period of 3 weeks and tested our method for 1 week. The volunteers also act as assessor of URLs, they evaluate if the returned URL is relevant, highly relevant or non relevant and this information is recorded for evaluation. They are the students and faculties selected from 5 different colleges in INHA University such as Engineering, Business, Medicine, Law, and Social Science. During the test, 300 queries per person, on average, were requested. We used Google search API [16], Yahoo search API [17], and Naver search API [18] for sending/receiving user query/search results from servers of Google, Yahoo, and Naver respectively. The following section compares the proposed personalization method with Google and Yahoo search engine using evaluation metric Average Rank [5] and DCG [19]. We have tested all our results for test of significance (T-Test). The test condition is whether our personalized result set provides improvement over the non-personalized search engine.

4.1 Evaluation Metric

The metric Average Rank [9] is used for measuring the quality of personalized search. The average rank (AR) of a query q is defined as shown in equation (3).

$$AR_q = \frac{1}{|V|} \sum_{P \in V} R(P) \tag{3}$$

where $R(p)$ is the rank of URL p. The final AR over all the queries for a user is computed as shown in equation (4). Smaller value of AR indicates better placement of results.

$$AR = \frac{1}{|Q|} \sum_{P \in Q} AR_q \tag{4}$$

Second metric that we used for measuring the quality of results is the Discounted Cumulative Gain (DCG). The Average Rank fails to evaluate user preference at different relevance levels, for that reason DCG is more preferred. The DCG is particularly useful when results are evaluated at difference relevance levels (highly relevant, relevant, and not relevant) by assigning them difference gain values. For the returned result of URLs, a gain vector G is calculated. A higher value of gain vector symbolizes more relevant results and vice versa. For example: if the highest value of CG is 20 in scenario1 and 12 in scenario2, this implies scenario1 has more highly relevant or relevant results as compared to scenario 2. The formulae for calculating CG from Gain Vector G is shown in equation (5).

$$CG = \begin{cases} G(1) & \text{if } i = 1 \\ CG(i-1) + G(i) & \text{otherwise} \end{cases} \tag{5}$$

The idea behind DCG is, the greater the rank, the less accessible that URL is and hence less valuable it is to the user. The equation (6) shows the formulae used for computation of DCG,

$$DCG = \begin{cases} G(1) & \text{if } i = 1 \\ DCG(i-1) + G(i)/\log_b(i) & \text{otherwise} \end{cases} \quad (6)$$

For the purpose of this experiment, we have used 3 different relevance level G(*i*)=2 for highly relevant results and G(*i*)=1 for relevant results and G(*i*)=0 for not relevant results. Also *b* is the parameter of penalization; we have taken value 2 for *b*.

4.2 Performance Comparison

We now evaluate how the rankings of a non-personalized search engine and a personalized search engine differ based on the valuation we collected from our volunteers. We found that the personalized search engine returned more relevant results than a non-personalized search engine. However, the same query when issued by multiple users, received differed result sets and also the user's rating was better.

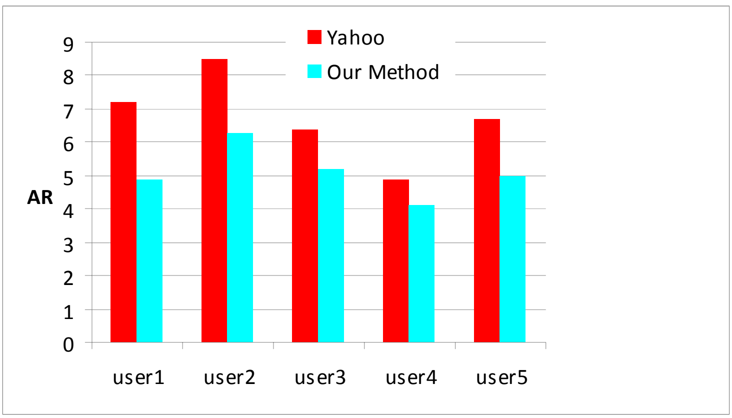


Fig. 3. Average rank of our method and Google

As seen in Fig. 3 and 4, our personalized engine gives better performance than Google and Yahoo. The groups of volunteers from different departments of Business, Medicine, Law, Engineering and Social Science are represented as User1, User2, User3, User4, and User 5 respectively. For example, for user 4, the average AR for Google is 6.5. However, our personalization method reduced the average AR to 4.1 which is an improvement of 36.9%. The overall average improvement after personalization of Google is 29.1%, whereas for Yahoo, it is 34%.

We performed a paired sample t-test to determine whether there is a significant difference between means of average AR for Google/Yahoo and that for our method. Table 1 shows the result of paired t-test for AR of Google and personalized Google. The average AR difference between Google and our method is 1.7 with t-value = 6.89 and P-value = 0.00001. This indicates that the difference of the average AR of Google and our personalization method is significantly different at the 0.01 significance level.

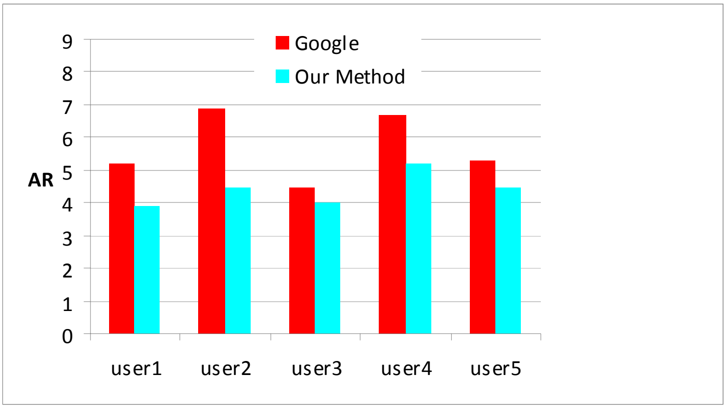


Fig. 4. Average rank of our method and Yahoo

Table 1. Paired T-Test Summary Statistics for Google and Our Method

	N	Mean	Standard Deviation	Difference of Means	t value	Pr > |*t*|
Google	50	5.92	1.37	1.7	6.89	0.00001**
Our Method	50	4.22	1.44			

$P^{*} < 0.05$, $P^{**} < 0.01$

Table 2 shows the result of paired t-test Yahoo and our method. The average AR difference of Yahoo and personalized Yahoo is 2.4 with t-value = 7.11 and P-value = 0.00001. This indicates that the difference of the average AR of Yahoo and our method is significantly different at the 0.01 significance level.

Table 2. Paired t-test summary statistics for Yahoo and our method

	N	Mean	Standard Deviation	Difference of Means	t value	Pr > |t|
Yahoo	50	7.08	1.85	2.4	7.11	0.00001**
Our Method	50	4.67	1.72			

$P^{*} < 0.05$, $P^{**} < 0.01$

The average DCG plot for Google and the personalized Google and Yahoo is shown in Fig. 5. The plot is an average plot of DCG over all the queries. It is evident from the plot that the results returned by a personalized search engine have a higher DCG, which represents better result quality.

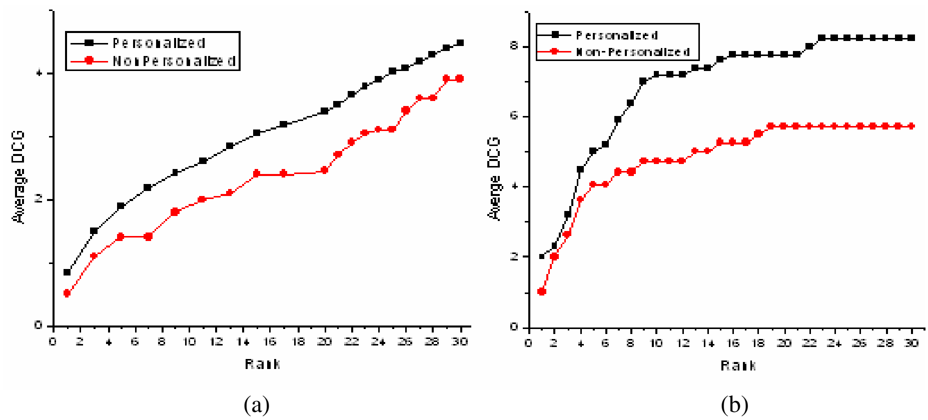


Fig. 5. (a) Average DCG plot for Google and personalized Google and (b) Average DCG plot for Yahoo and personalized Yahoo

5 Conclusion

In this paper we proposed a personalized re-ranking of URLs retrieved from an existing search engine. The re-ranking method is based on the user profile which is constructed based on user click-through behavior and assigning appropriate weights for different user actions. We evaluated the proposed method for Google, Yahoo, and Naver using DCG and Average Rank. The results in the experiment section show that our method reduces the average rank of an existing search engine. Our method can be practical for saving surfing time and effort to find users' preferred web pages. Also, the proposed system can be extended or implemented for a personalized mobile web search engine.

We would like to explore how our system can use all available search engines and integrate the results. It doesn't pose any theoretical difficulties but could significantly reduce user browsing time.

References

1. Sun, J., Zeng, H., Liu, H., Lu, Y., Chen, Z.: CubeSVD: a novel approach to personalized Web search. In: WWW 14, pp. 382–390 (2005)
2. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: A bibliography. In: SIGIR, pp. 18–28 (2003)
3. Agichtein, E., Brill, E., Dumais, S., Ragno, R.: Learning user interaction models for predicting web search result preferences. In: SIGIR, pp. 3–10 (2006)
4. Shen, X., Tan, B., Zhai, C.: Context-sensitive information retrieval using implicit feedback. In: SIGIR, pp. 43–50 (2005)
5. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: SIGIR, pp. 449–456 (2005)
6. Chirita, P.A., Firan, C., Nejdl, W.: Summarizing local context to personalize global web search. In: CIKM, pp. 287–296 (2006)

7. Das, A., Datar, M., Garg, A., Rajaram, S.: Google News Personalization: Scalable Online Collaborative Filtering. In: WWW, pp. 271–280 (2007)
8. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intelligence and Agent System 1(3-4), 219–234 (2003)
9. Speretta, M., Gauch, S.: Personalized search based on user search histories. In: WI, pp. 622–628 (2005)
10. Ferragina, P., Gulli, A.: A personalized search engine based on web-snippet hierarchical clustering. In: WWW 14, pp. 801–810 (2005)
11. http://www.vivisimo.com
12. http://www.google.com/search?hl=en&esrch=RefinementBarLhsGradientPreview&q=windows&btnG=Search
13. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: SIGIR, pp. 19–26 (2006)
14. Dou, Z.: A Large-scale Evaluation and Analysis of Personalized Search Strategies. In: WWW 16, pp. 581–590 (2007)
15. Eiron, N., McCurley, K.S.: Analysis of anchor text for web search. In: SIGIR, pp. 459–460 (2003)
16. http://code.google.com/apis/ajaxsearch/
17. http://developer.yahoo.com/search/web/
18. http://openapi.naver.com/index.nhn
19. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. In: TOIS, pp. 422–446 (2002)
20. Davison, B.D.: The Design and Evaluation of Web Pre-fetching and Caching Techniques, Ph.D. dissertation, Rutgers University, New Brunswick (2002)

Library Services as Multi Agent System

Toshiro Minami

Kyushu Institute of Information Sciences, Faculty of Management and Information Sciences, 6-3-1 Saifu, Dazaifu, Fukuoka 818-0117, Japan and
Kyushu University Library, 6-10-1 Hakozaki, Higashi, Fukuoka 812-8581, Japan
minami@kiis.ac.jp, minami@lib.kyushu-u.ac.jp
http://www.kiis.ac.jp/~minami/

Abstract. The major aim of this paper is to propose a multi-agent model that represents the natural structure of library services that are carried out in collaboration with libraries, interaction between libraries and collaboration and their patrons, collaboration among patrons, and so on. Another aim is to demonstrate its usefulness in terms of flexibility and scalability. For example, in order to achieve their missions libraries provide the service called ILL (Inter-library loan), with which a patron, or a user, can borrow books that belong to other libraries. Such a collaborative service is very common to current libraries. In the future, it would be extended to other types of collaborative services such as collaborative online referencing, collaborative community formation and information sharing, and so on. The library system organized as a multi-agent system should be very useful in order to provide such services in a flexible and scalable framework.

1 Introduction

The mission of library is well described in the well-known "The Five laws of library science" by S. R. Ranganathan [6], which was published way back in 1960s:

(1) Books are for use,
(2) Every reader his book,
(3) Every book his reader,
(4) Save the time of the reader,
(5) The library is a growing organism.

These laws are still true even now if we rephrase them by replacing "book" with "information", "material", or "service". In order to carry out such missions libraries have been introduced many tools and machines. For examples, they have installed computers and provide OPAC (Online Public Access Catalog) for book retrieval and connected them to the Internet so that their patrons are able to access library services like Web-OPAC, online referencing and other library services.

In fact, libraries and library services are a good application target for multi agent system (MAS) [3]. They are run by various organizations like universities, federal government, local government, and private companies thus they behave autonomously. At the same time, they cooperate. One typical example is the ILL

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 222–231, 2008.

(Interlibrary Loan) service which is provided by most libraries. The most important mission of libraries is to help with their patrons with providing information material, e.g. books, magazines, newspapers, and with supporting them for strengthening their information literacy skills. Suppose a patron is searching for a book and it is not in the collection of the library, the library will try to find out a library that has the book and then it will ask the library for let it borrow the book for its patron. This is a rough sketch how the ILL service is carried out.

Due to the advancement of network society, such library services are requested to be extended so that they are shifted to be more network-based, more collaborative with other libraries, hopefully providing 24 hours, and so on. In order to realize such services, libraries and library systems should change themselves to be more network-oriented automated ones so that patrons are able to access to their libraries at any time from wherever they want. Constructing the library system based on a multi-agent framework should be one of the best solutions because of the flexible and scalable nature of multi-agent systems.

The major aim of this paper is to propose a multi-agent library system model that is very appropriate to construct such systems. More precisely goals of this paper include:

1. to propose a model for describing and realizing the library services based on the multi-agent system (MAS) framework,
2. to demonstrate its usefulness and appropriateness in terms of scalability, flexibility, and other points of view, and
3. to give a grand design about how can such a MAS system be constructed.

The rest of this paper is organized as follows:

In Section 2, we will describe the library agent system in three different points of view. Firstly in Section 2.1, we present the concept model, where the participants of libraries have their corresponding agents and these agents work harmoniously so that they form a society of agents. We will give a more detailed discussion in Section 2.2, where agents are classified in two types; place and mobile agents. In this classification, an agent in the former type makes a place, or a field, where the agents in the latter type may move from one place to another one. When a mobile agent comes and joins the group of agents in this place. Mobile agents can communicate with its place agent and other mobile agents that are located in the same place. Lastly in Section 2.3, we will describe how such agent system can be implemented.

In Section 3, we will investigate more precisely about the system when we use RFID tag system [1] in a library. The discussions here are based on the model proposed in our previous work [3]. RFID is a very useful tool for automatically detecting the change of status of books and move of patrons. Firstly we will describe the principle of RFID tag system and show how this technology is used in the libraries. We will also introduce the concept of intelligent bookshelf (IBS), which is not installed in many libraries right now but it will play a very important role in the next generation RFID libraries. We will try to convince the importance of "library marketing" in collaboration with RFID. Furthermore we will show a couple of examples how events in the real world is processed in the agent society.

Lastly in Section 4, we will summarize what we have discussed in this paper and show the prospect of the agent-based library system towards the future.

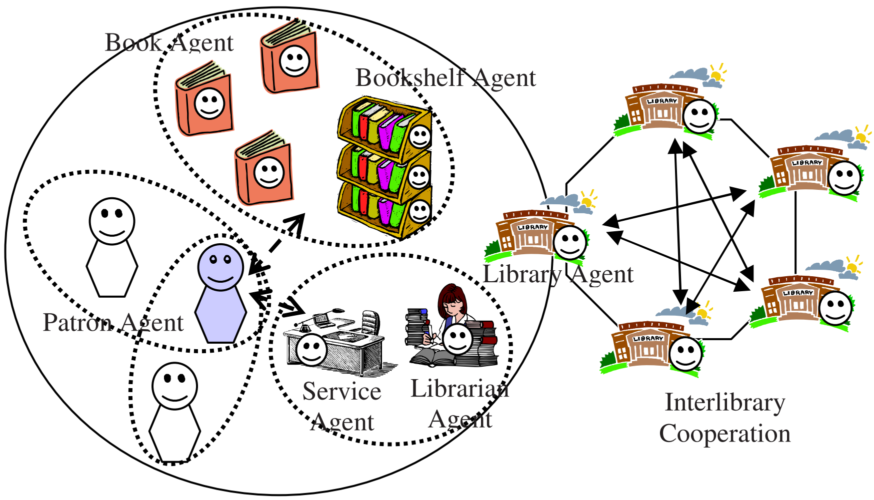


Fig. 1. Concept Model of Library Agent Systems

2 Models of Library Agent Systems

In this section we show three different types of models; each model is constructed from different view of library agent system. In Section 2.1 we deal with the concept model, where patrons, librarians, library services, books, bookshelves, libraries themselves are represented as agents and they form a heterogeneous society. They communicate each other and some of them will form various kinds of groups. Library agents form an inter-library network so that agents belonging to one library can communicate with other agents that belong to other libraries.

In Section 2.2, we discuss the inter-agent communication method more precisely by classifying agents in two types; place and mobile. Mobile agents in a place can communicate the place agent directly and they can do to other agents via the place agent.

In Section 2.3, we design a model for implementing the agent system that has been discussed before. In the conceptual, or virtual, sense book agents can be seen as if they run as they are. Actually in the real world, at least in the current technology, they are just the objects which have IDs. In the implementation in the model, book agents run in an agent server.

2.1 Concept Model of Library Agent System

The concept model is the one we consider all the participants of library system, i.e. patrons, librarians, books, bookshelves, libraries, etc. as agents. In this model, we can see the whole library system more clearly and easily than other models that come from other points of view.

Figure 1 illustrates how agents are related each other. It is separated in two areas; left and right. The left area is the group of agents belonging to a library, which is represented as the left most library in the right part that consists of library agents.

In the left area are several kinds of agents; patron, book, bookshelf, service, and librarian. Each agent represents an object, a person, a service, according to its type; a patron agent represents a patron of the library, a book agent represents a book of the library, and as well in other types of agents.

The agents in a dotted line form a group of agents. In the figure the patron agents have two groups. The groups are formed according to the patron's interests, attributes, or whatever he/she wants.

Books and bookshelves also make a group. It is a group based on the relationship between books and a tier of bookshelves; i.e. each book belongs to one of the bookshelf in the tier. So there are quite a number of such groups in a single library.

The dashed arrow between the gray patron agent and the group of books and bookshelves indicates the interactions between them. An interaction may be that the patron takes out the book from a bookshelf. This is a relationship between the patron and the book.

Other possible relationships between them include the patron reserves the book in the library system, the patron gives comments or something in a Web page of the library, or whatever action the patron takes about the book and the action is detected by the library.

The service agent and the librarian agent form a group in Figure 1. This group is formed when the librarian is assigned to the job for this service. In other words, the service agent is a representation of the system relating to the job. Suppose, for example, the job is the reference. Then the service agent represents the system for referencing. It will support the librarian who is working as a reference librarian. This group of agents represents a team for this service/job. The dashed arrow between this group and a patron agent represents a relationship; which may, for example, the action that the patron comes to the reference counter and ask a question.

Such groups and interactions are in a library. The libraries make a network for communicating and collaborating each other. One typical example is the ILL (Interlibrary Loan) activity. A patron in a library may want to read a book which is not in the library he/she is using. Then the "home library" of the patron asks another library for borrowing the book. With this ILL system, a patron can borrow whichever book that is purchased by a library in the library group, where the member libraries agree with such collaboration. One possible model for matchmaking the request and provision of information is the word-of-mouth network [5].

By seeing the library system as a MAS system, we have some advantages.

(1) We can deal with the whole participants, actions, services, etc. in a unified framework.

(2) The system becomes scalable and flexible in nature. It is easy to extend the library services and its organization in this agent-based system.

(3) Also the agent framework is very natural to represent the structure of organizations working collaboratively like the libraries, where each organization is basically independent from other organizations. Thus the autonomous nature of each library can be represented in the autonomous nature of each agent.

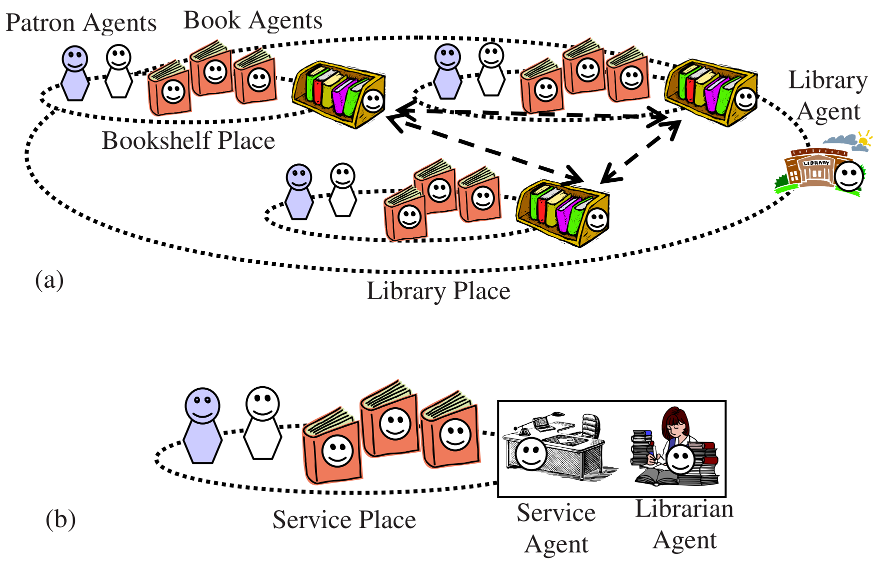


Fig. 2. Multi-Layered Models: (a) Layered Structure of Place and Agents (b) Collaborative Structure of Service Place

2.2 Two Types Classification of Library Agents

In this section we classify agents in two types; passive and active. The agents relating to patrons, books are supposed to be passive agents. In our framework passive agents can not directly communicate each other. They communicate only with active agents, which provide places, or fields, for passive agents to communicate. Active agents are stationary and relating to some kinds of places such as library building, service counter, tier of bookshelf, etc. An active agent provides a place, which is a field for passive agents communicate each other when they get together in the place. A passive agent is a mobile agent that is able to move from one place agent to another.

The general idea about this structure is illustrated in Figure 2. In Figure 2(a), patron agents and book agents are located in the same place which corresponds to a bookshelf agent. The bookshelf agents in a library make a group where they communicate each other. A passive agent who is in a bookshelf place can detect the existence of other agents and communicate with them via the network of bookshelf agents. Furthermore if an agent in a library wants to communicate with another agent who is located in another place in another library, it can do it via the networks of bookshelf places and via the network of libraries.

Figure 2 (b) illustrates how two agents collaborate. In this example, a service agent, supposed to be a reference desk, and a librarian agent work together and provide services to patron agents. The service agent knows, i.e. keep records, the history data of reference services, while the library agent similarly knows the history of jobs, skills, and other information about the librarian. The major service to a patron agent

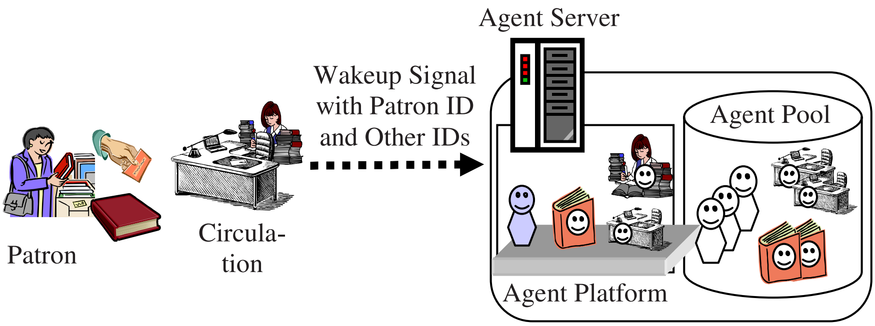


Fig. 3. Wake-up Mechanism of Agents

in this example is to take questions and give an answer, advice, etc. The reference process is automatically recorded by the service agent and some information is learned, i.e. recorded, in the librarian agent. The book agents in the figure are those material used in the reference job in order to solve the problem.

2.3 Implementation Methods of the Library Agent System

In this section we will describe in what way the agent system is realized in the real situation. The models proposed so far consist basically of abstract and virtual agents and their behaviors. In order to make such system realize, we have to show the correspondence between the real world computation and the actions in the virtual world of agents.

Figure 3 illustrates the wake-up mechanism of agents. In the current status, patron's ID card is the only candidate for the patron agent. It is only a card on which ID data is marked and it has no capability of calculation or whatsoever. Similarly a book agent has no ability of calculation. A book has a label on which the book's ID is marked. This is the reason why patron agents and book agents are classified as passive. They could not run like agents as in the abstract models.

Delegate agent mechanism is a solution to this problem. A delegate agent is an agent program in an agent server. The agent programs, or agents, are saved in a storage device of the server, or agent pool. An agent pool contains both passive and active agents. We will call that they are "sleeping" when they are just stored in an agent pool. A sleeping agent will be woken-up when a triggering event occurs. The agent starts running in the agent platform after it is waken-up.

In Figure 3, a patron arrives at a library circulation counter and asks to the librarian for borrowing a book. Then the librarian uses a barcode reader or something and reads the patron ID and book ID. Then the library system sends the wake-up signal to the agent server together with the IDs of patron, book, and circulation agent ID. Then the agent server let the corresponding agents run and let them do their jobs.

3 Library Agent System with RFID

In this section we will assume a library which has installed the IC tag system by using the RFID (Radio Frequency Identification) technology. RFID is a technology for automated detection of ID data from person and object. The contents of this section is a trial to reorganize the one proposed in our previous work in [3] based on the extended model proposed in this paper.

In libraries it is well used as a substitute to barcode for ID detection and to magnetic tag for theft-protection. We can say RFID makes the process more automated than barcode because RFID has less restriction in the position and direction of the detected objects. Thus the agents are more automatically triggered as the books, patrons and other participants just go through or stay near a reader. This is a reason why we pay much attention to RFID technology for library automation.

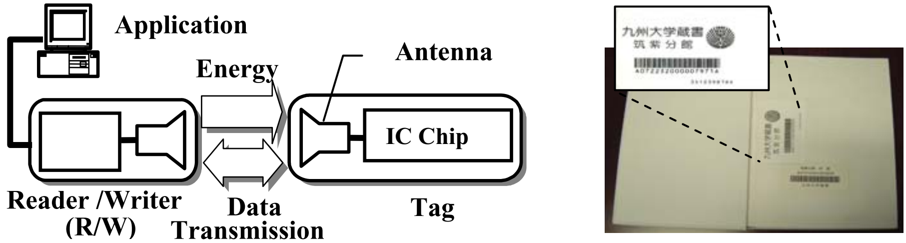
3.1 RFID Technology and Its Application to Library

The principles of RFID tag system is illustrated in Figure 4(a). The RFID tag system consists of two major components; tags and reader/writers (R/Ws). A tag is able to communicate with a reader/writer when they are located sufficiently close each other.

As is shown in Figure 4(a) the RFID tag at the right-hand side consists of an IC chip and an antenna. It has no batteries and thus cannot run stand alone. At the left-hand side is an R/W, which provides energy to the tag with its antenna. The RFID tag gets energy from the R/W with electro-magnetic induction via its antenna. It waits until sufficient energy is charged in it and when it is ready, it starts running and communicates with R/W and exchanges data such as its ID and status data by making use of the same antenna. There are two types of tags; one is read-only and the other is read-write. Normally the latter type is used in library application.

Figure 4(b) shows an example of how RFID tag is attached on a book. It is an RFID tag used in Chikushi Branch Library of Kyushu University Library [2], Japan. The tag is formed as a label on which the library name is marked together with the university logo. The material ID is also marked in barcode on the label. The barcode is supposed to be used when this material is carried to another library in the ILL (Inter-Library Loan) program, i.e. for interoperability, and when the tag has been bad and does not respond to R/Ws, i.e. for just in case.

Comparing to the barcode system which is mostly used in libraries now, RFID tag system has an advantage that it is much easier to put material in a right position. As a result self checkout machine is easier to operate so that it is easy enough for children



(a) (b)

Fig. 4. (a) Principle of RFID (b) RFID Tag Attachment on a Book

(a) (b) (c) (d)

Fig. 5. RFID Reader/Writers for Libraries: (a) Security Gate, (b) Desktop Type for Circulation Counter, (c) Handy Type for Book Inventory, and (d) Intelligent Bookshelf (IBS)

and elderly patrons to use. This is a very important point. So far the dominating reason for the libraries whey they introduce the RFID tag system is that it let the processes be more efficient; i.e. it is faster to proceed circulation, it is supposed to have less running cost, and thus the number of librarians needed will be smaller, etc., even though its initial cost is very high.

Typical usages of R/Ws in libraries are circulation counter, self checkout desk, security gate and handy R/W for inventory. Some library has reduced the circulation counter in half by adding self checkout/return machines. For inventory, the necessary job terms may reduce from a couple of weeks to a couple of days.

An intelligent bookshelf (IBS) (Figure 5(d)) is a bookshelf which is a shelf type RFID reader/writer, where its RFID antennas are installed in the bookshelves so that the R/W can detect what books are put in which shelf in real time [4][7]. There are a couple of types of antennas. The book-end type is the one the antennas are put next to books like spacers as is in Figure 5(d). Now we are starting development of such system [7]. Another one is the shelf-board type, of which the antennas are put under the books either in between the books and the shelf-board or the boards themselves.

By using such equipments we can collect the use data of books. By analyzing the use data we may extract useful knowledge, which we call “library marketing” [7]. For example, we can get the information how often a specific book is used and the differences of usage patterns according to the day of the week, time zone in a day, etc. We can provide such information to the library patrons. It may be used by library staff for planning improved services as well. Such technology should be very important for libraries in the ubiquitous environment that will come in the near future.

3.2 Agentified Library Service Model with RFID

The actual library service that uses the RFID system can be represented in the agent framework as is illustrated in Figure 6. The left part is a figure of a reference desk, where a reference librarian is taking the job. A patron comes to the desk and asks a question to the librarian. On the desk is RFID R/W, with which the IDs of books,

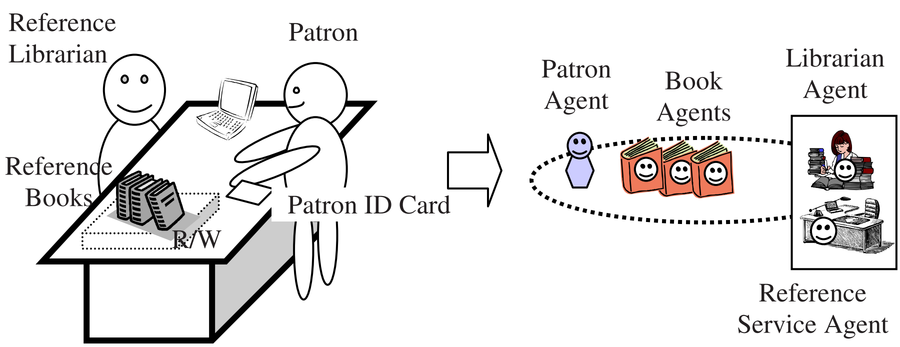


Fig. 6. Reference Desk Equipped with Reader/Writer and its Corresponding Agent System

supposed to be reference books, used in the reference process are recorded automatically by the application for the R/W. Note that the recording is done automatically so that the librarian is not supposed to do any action for this recording.

The right side of Figure 6 is the corresponding agent system. The patron agent is invoked automatically when the patron shows his/her ID card and its ID is recorded by the reference system. The date and time is also recorded by the service agent. The fact that this librarian does this reference job is also recorded automatically. The librarian agent and the reference service agent make a team when the librarian put his/her ID card near the reader. The book IDs are read by the reader, or R/W, with date and time data, i.e. timestamp, so that the reference service agent can calculate how many minutes they are put on the table, or how long time they are used in this reference job.

So far the most libraries do not put much effort for collecting the case data of referencing. Even though some library groups have tried to construct reference case database most of them failed because the cases are created mostly by library staff manually.

By using the RFID technology combined with the MAS framework such data relating to the referencing jobs are recorded automatically and later on the library agent, or maybe some other agent, will analyze the collected data and tries to extract useful knowledge.

This is a good application field of "library marketing [7]" which will be popularly used in most libraries. Thus we look forward to the future of this technology. It will innovate in the styles of services and management of libraries.

4 Concluding Remarks

In this paper we have proposed a framework for library services based on multi agent architecture. In this framework, library patrons, librarians, library materials such as books, magazines, CDs, etc. are considered agents and they communicate each other. In such a way patron agents, and thus patrons themselves, are able to share knowledge and also to work collaboratively for creating new knowledge.

We have discussed the model where agents are classified in two types; passive and active. The passive agent has no computational ability so that they actually communicate with other agents through the delegate agent. The delegate agents are in sleeping in an agent pool and they are waken-up and run as if their corresponding agents run in the virtual agent field.

The library services so far are loaning books, providing studying rooms, etc. In the coming ages more advanced services like helping patrons with sharing and crating knowledge, consulting them with finding information, or “concierge service,” teaching or training them for improving learning skills, and forming communities.

It is very appropriate to construct the library services as a multi agent system in this point of view. Then the system is scalable, robust, and flexible. You can add new members and extend services with ease in this agent-based framework for library services.

Some of the remaining problems to be pursued in this approach are (1) to investigate more precise model of agent based library services, (2) to complete the implementation of the system, which is designing and developing now, and (3) to evaluate the system and improve and refine the model so that the architecture proposed in this paper becomes applicable to many libraries.

In the future agent technology will be widely used so that the library agents will help their patrons with learning, searching information, etc. and become a useful tool.

References

1. Finkenzeller, K.: RFID Handbook, 2nd edn. John Wiley & Sons, Chichester (2003)
2. Kyushu University Library: `http://www.lib.kyushu-u.ac.jp/?skinid=7`
3. Minami, T.: Needs and Benefits of Massively Multi Book Agent Systems for u-Libraries. In: Ishida, T., Gasser, L., Nakashima, H. (eds.) MMAS 2005. LNCS (LNAI), vol. 3446, pp. 239–253. Springer, Heidelberg (2005)
4. Minami, T.: RFID Tag Based Library Marketing for Improving Patron Services. In: Hoffmann, A., Kang, B.-h., Richards, D., Tsumoto, S. (eds.) PKAW 2006. LNCS (LNAI), vol. 4303, pp. 51–63. Springer, Heidelberg (2006)
5. Ohtani, T., Minami, T.: The Word-of-Mouth Agent System for Finding Useful Web Documents. In: Proc. Asia Pacific Web Conference (APWeb 1998), pp. 369–374 (1998)
6. Ranganathan, S.R.: The Five Laws of Library Science. Bombay Asia Publishing House (1963)
7. Zhang, L., Minami, T.: Library Marketing that Boosts Education with RFID. In: Proc. 6th International Conference on Information Technology Based Higher Education and Training (ITHET 2007) (2007)

Ontology Methodology Based Context Awareness Using System Entity Structure for Network Analysis

Taekyu Kim[1], Young-Shin Han[2],Tae-young Kim[2], and Jong-Sik Lee[2]

[1] Arizona Center for Integrative Modeling and Simulation
Electrical and Computer Engineering Department
The University of Arizona, Tucson, AZ. 85721, USA
taekyuk@ece.arizona.edu
[2] School of Engineering Inha University
253 Yonghyun-Dong , Nam -Gu, Incheon , Korea 402-751
yshan95@ewhain.net, silverwild@gmail.com, jslee@inha.ac.kr

Abstract. Currently network uses, especially the number of internet users, increase rapidly. Also, high quality of service is required and this requirement results a sudden network traffic increment. As a result, an efficient management system for huge network traffic becomes an important issue. Ontology/data engineering based context awareness using the System Entity Structure (SES) concepts enables network administrators to access traffic data easily and efficiently. The network traffic analysis system, which is studied in this paper, is an effective management system with a high throughput and a low response time designed and implemented based on a model and simulation based data engineering methodology. Extensible Markup Language (XML) is used for metadata language in this system. The information which is extracted from the network traffic analysis system could be modeled and simulated in Discrete Event Simulation (DEVS) methodology for further works such as post simulation evaluation, web services, and etc.

Keywords: ontology, data engineering, System Entity Structure (SES), network traffic analysis, and Discrete Event Simulation (DEVS).

1 Introduction

Independent computers are no more valuable in current social environment. In contrast, network connected distribute and/or parallel computers maximize their computational power for keeping and processing huge data as well as sharing data among companies and organizations. Under fully utilized network environment, policy makers in companies and organizations may receive fast, accurate, and sufficient data as required.

Currently network uses, especially the number of internet users, increase rapidly. Also, high quality of service is required and this requirement causes a sudden network traffic increment. As a result, an efficient management system for huge network traffic becomes an important issue. The definition of a network traffic analysis is monitoring all the network behaviors, controlling networks and hosts, and applying

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 232–241, 2008.

network traffic behaviors to achieve an effective management. Network administrators are having trouble with lack of consistent traffic analysis and management system. Network traffic analysis systems which lack of a high throughput and a low response time require several processes of manipulating metadata. Therefore, the limitation results insufficiency of useful data for designing network capacity.

The objective of this study is to show how ontology/data engineering methodology is efficient for designing and implementing an optimal network traffic analysis system with a high throughput and a fast response time. To achieve the goal, we use the System Entity Structure (SES) for the system design, i.e., ontology(network traffic analysis) design. The SES is a theory to design structured information hierarchically and efficiently, and it is a useful method for data engineering.

This paper includes background knowledge in section 2. In section 3, we show the development processes for the network traffic analysis system. The experimental results are addressed in section 4, and finally we conclude and present future works.

2 Background

There exist modeling approaches for designing and simulating complex systems' specifications. Among the approaches, this paper examines two approaches, Entity-Relation and Unified Modeling Language. The first approach is Entity-Relation. Entity-Relation (ER) [1] is an approach to represent systems' structures such as data entities and their relationships. The other approach is Unified Modeling Language (UML) [2] and it is designed to represent objects and their relationships. Even if ER and UML enable modeling systems' specifications and describe various kinds of logical models, each approach is limited in specific modeling [3]. More specifically, ER is useful for logical model representations but not practical for visual modeling. Meanwhile, UML is good for both logical models and visual models but lacks for model persistence [4].

The System Entity Structure (SES) expressed by XML modeling framework is useful for multi-level (high level applications and low level data specifications) modeling and simulation based on ontology and data engineering concepts. The SES is a hierarchical tree structure with entities and relations. The SES approach shows better vision for logical, visual, and persistent modeling than either ER or UML.

2.1 System Entity Structure (SES) and Pruned Entity Structure (PES)

The basic idea of system entity structure is that a system entity represents the real system enclosed within a certain choice of system boundary. In real system, many system entities and the experimental frames are dealt. Thus it is necessary to organize the model and experimental frames around the structure. The entity structure is a template from which the decomposition trees of the existing models can be extracted. More than this is a template for constructing models from those already existing. The followings are the key components consisting of system entity structure.

1. Entity: An entity is intended to represent a real world object which either can be independently identified or is postulated as a component in some decomposition or a real world object.
2. Aspect: An aspect represents one decomposition out of many possible of an entity. The children of an aspect are entities representing components in a decomposition of its parents.
3. Specialization: A specialization is a mode of classifying entities and is used to express alternative choices for components in the system being modeled. The children of a specialization are entities representing variants of its parent.

To construct a desired simulation model to meet design objective, the pruning operation is used to reduce the SES to a pruned entity structure, PES [5]. The pruned entity structure can be transformed into a composition tree and eventually synthesized into a simulation model. Professor Zeigler proposed the System Entity Structure (SES) [5, 6] and the SES is a theory to design systems hierarchically and structurally. The basic idea of the SES is that a system entity represents the real system enclosed within a certain choice of system boundary. The SES includes entities and their relationships.

Fig 1 illustrates the SES basic methodology of the conceptual relationship between the SES representing ontologies and implementation in the XML. First of all, the SES which can describe the components in the source data is developed. The SES structure becomes important information to build the DTD or Schema. Entity, Aspect, Multi-Aspect, and Specialization build the primary components in DTD or Schema. At the ontology level, the modeler develops one or more SESs depending on models, and the SESs are merged to create an ontology in order to satisfy the pragmatic frames of interest in a given application domain. An SES can be specified in various ways, and then it is transformed to an XML schema or an XML document type definition (XSD or DTD) at an implementation level. Pruning operation to SESs creates pruned entity structures (PESs) and the PESs transform to simulation models.

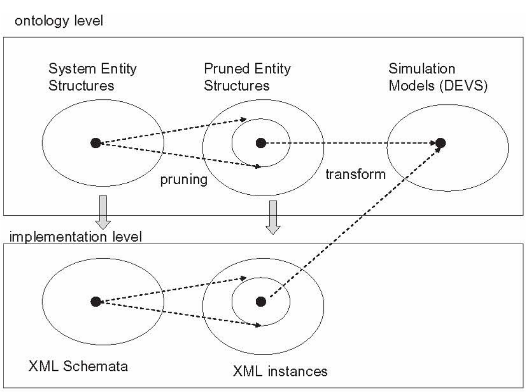


Fig. 1. Architecture for model and simulation-based data engineering methodology [7]

2.2 The DEVS Formalism

The Discrete Event System Specification (DEVS) is a formalism providing a mean of specifying a mathematical object called a system. It also allows building modular and hierarchical model compositions based on the closure-under-coupling paradigm. The DEVS modeling approach captures a system's structure from both functional and physical points of view. A system is described by a set of input/output events and internal states along with behavior functions regarding event consumption/production and internal state transitions. Generally, models are considered as either atomic models or coupled models. The Atomic model can be illustrated as a black box having a set of inputs(X) and a set of outputs(Y). It describes interface as well as data flow between the atomic model itself and other DEVS models. The Atomic model also specifies a set of internal states(S) with some operation function (i.e., external transition function (δ_{ext}), internal transition function (δ_{int}), output function (λ), and time advance function (ta())) to describe the dynamic behavior of the model. Figure 2 illustrates the system representation of an atomic model [8].

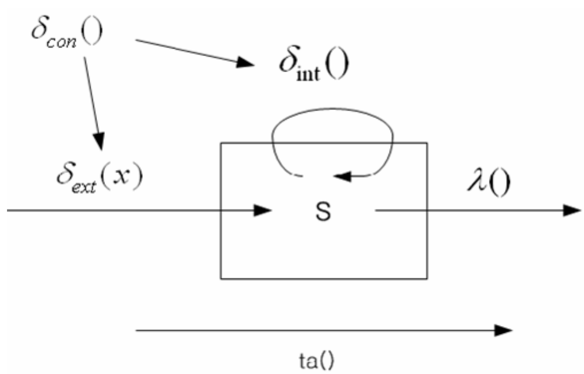


Fig. 2. System representation of atomic model

The external transition function (δ_{ext}) carries the input and changes the system states. The internal transition function(δ_{int}) changes internal variables from the previous state to the next when no events have occurred since the last transition. The output function (λ) generates an output event to outside models in the current state. The time advance (ta()) function adjusts simulation time after generating an output event.

3 Network Traffic Analysis System Based on Ontology Methodology Using SES

The goals of network traffic analysis are to help network administrators to manage complicated network topology and to increase efficiency for secure and effective data transfer. The network use, especially the number of internet users, increases rapidly. Also, high quality of service is required, and this requirement causes a sudden network traffic increment. As a result, an efficient management system for huge network traffic becomes an important issue. Here, ontology/data engineering methodology is used to build an effective system for network traffic data handling. The SES and XML modeling approaches allow this system easy to handle huge data and help modeling and simulation because the architecture of the SES is hierarchical tree structure, and an accessing data method in this hierarchical structure is one of the fast approaches. In addition, the characteristics of XML such as scalability and portability is helpful for manage metadata. In this section, we see the overview for development process of the network traffic analysis system, the way of capturing network traffic data, and the details of developing the system.

3.1 Capturing Network Traffic Data

To construct the system for a network traffic analysis, we need real network traffic data. We monitor network behaviors in a subnet of Arizona Center for Integrative Modeling and Simulation (ACIMS) lab in the department of electrical and computer engineering, The University of Arizona. We use the Ethereal[9] for capturing network behaviors and apply a filtering method for necessary data for this system.

Nine elements, which are a capturing time, a source IP address, a source MAC address, a source port number, a destination IP address, a destination MAC address, a destination port number, a protocol, and a packet length, are fundamentals for a network traffic evaluation. We also notice that the number of events is very large, i.e., the total number of packet transmitted in the subnet during one minute is about forty five thousands. However, we need monthly information or even yearly information. The study shows the reason why data engineering is so important for managing huge data.

3.2 Generation Process of Network Traffic Analysis System

This section addresses the development process of network traffic analysis for data engineering. The first step is to gather network traffic data. We use Ethereal [9] for capturing network traffic data. The next process is to design a hierarchical SES tree structure in order to access any information in a short time. Subsequently, writing a natural language which represents a designed SES is the next step. In this study, we use the SESBuilder [10] to generate XML in SES and XML in PES. The SESBuilder has its own natural language format for interpretation. Consequently, a natural language creates XML in SES and XML in PES. Before starting analysis, we port network traffic data that is captured in the first process into XML in PES, and this process produces a full XML metadata file. For this process step, we implement a porting application. The last process is to access data efficiently and examine network traffic. Also, we implement a parsing application to extract data fast. Based on the information we capture with Ethereal, we design the SES for network traffic analysis. Fig 3 represents the hierarchical SES tree of the system.

The SES hierarchically representing network traffic data is designed first, and the natural language is written as the next. We write the natural language which is mirroring to the SES in Fig 4, and we follow the format of the natural language which

Fig. 3. System Entity Structure (SES) representing the network traffic analysis system

can be interpreted in the SESBuilder. The SESBuilder transforms natural languages to SESs in an XML format. In addition to generating SESs in an XML format, the SESBuilder generates a Document Type Definition (DTD) and a schema. Subsequently, the SESBuilder produces a Pruned Entity Structure (PES).

The developing processes are (1) designing an SES ontology, (2) coding a natural language, and (3) generating xml template code. We find out that the SESBuilder is an easily usable tool for users to design systems. Once users generate a natural language, the SESBuilder produces a SES in XML and a PES in XML automatically, and these XML codes are practical for data management due to its advantages of scalability and portability. The SES design, the natural language, and the XML template code represent the same thing. In this study, we focus on a host based analysis rather than a network based analysis. Therefore, we could have a PES. PES presents the optimal number of transaction processes for a high throughput and a fast response time.

In this study, we focus on a host based analysis rather than a network based analysis. Therefore, we could have a PES. PES presents the optimal number of transaction processes for a high throughput and a fast response time. The SES in Fig 3 has multi-aspects of very top level entity, the NetworkTrafficAnalyses. It means that multiple instances of PESs for evaluating network traffic could be generated in regards to customers' various requirements. Fig 4 illustrates multi-aspects for pruning of copies of its entity. In the next section, we analyze a network throughput which requires data elements of times and packets sizes.

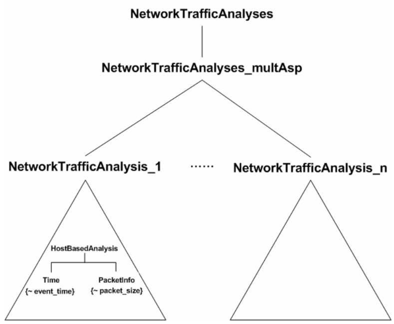


Fig. 4. MultiAspect for pruning of copies of its entity

The last step of developing process is to analysis. High throughput and low response time are essential for huge amount of data management. For analyzing application development, we uses java language, Document Object Model (DOM) library [11] for manipulating contents in XML code, and JFreeChart library [12] to show analyzed data with graphical charts. Fig 5 shows how the analyzing application works.

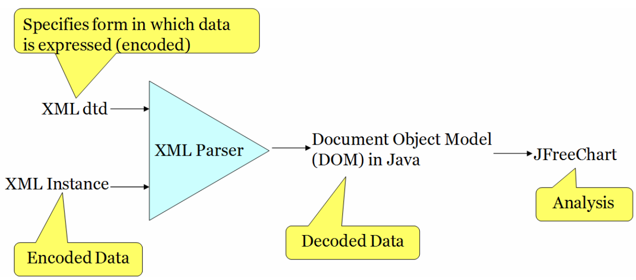


Fig. 5. Analyzing process with XML metadata

4 Experimental Results

In this paper, we model and simulate network traffic analyzer. *PacketCapture* model reads an xml instance file which is captured and transformed, and keeps the traffic data. *PacketClassifier* model classifies the traffic data, which is received from *PacketCapture* model, based on protocols. Consequently, *PacketClassifier* model sends the data to appropriate protocol analyzers, i.e., *TCP analyzers*, *NTP analyzers*, *FTP analyzers*, and *FTP-DATA analyzers*. Finally, protocol analyzer models evaluate traffics and show statistical results such as throughput, bandwidth, and etc. Fig 6 shows DEVS modeling.

The statistical results are shown as GUI reporting charts. Fig 7 shows throughput analysis. The X axis represents a simulation time which is a duration time of capturing network traffic data, and the Y axis represents a packet size at a specific

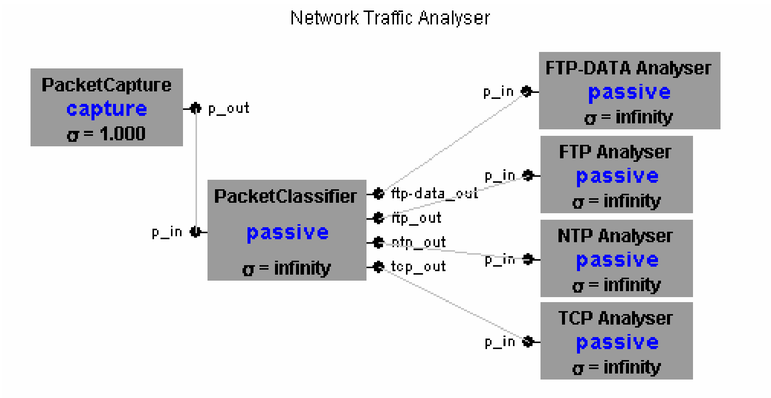


Fig. 6. DEVS modeling for network traffic analysis

time. And mouse tooltips illustrates detailed number of a packet size. We validate the reporting charts have same values as the original log data.

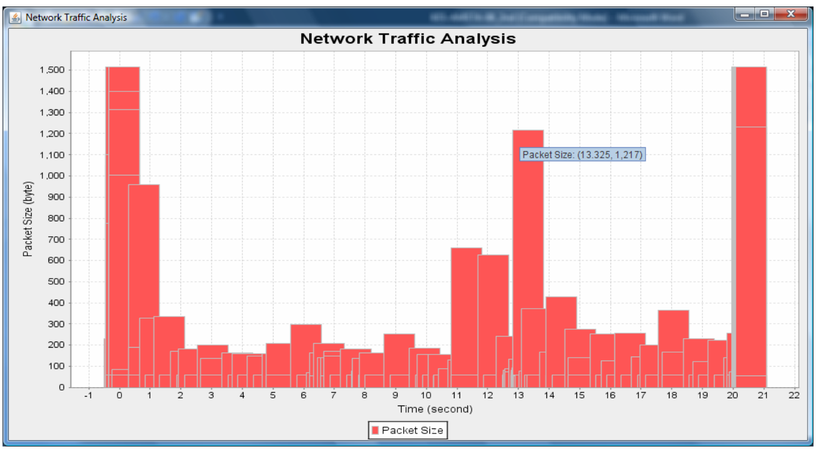


Fig. 7. Result Chart showing network throughput

The advantages of this network traffic analysis system are easy and fast information capturing, a fast response time, and a user centric schema. The fact that this system is developed based of an SES data engineering methodology which is a hierarchical tree structure, facilitates users to read, understand, and manipulate the network data easily. This user friendly graphical chart gives customers, which are network administrators, a general idea regarding a network throughput, and the graphical user interface (GUI) charts help reducing evaluating time.

In this paper we have illustrated the strength of the SES represent a large data set consisting of elementary units through the concept of multiAspect. As we have shown the SES concept of pruning employs aspects, multiAspects and specializations to allow very flexible specification of subsets of a given data set, and aggregation operations on them. Aggregation is a form of abstraction commonly employed in modeling in disciplines ranging from physics to economics. The aggregation restructuring concepts give better idea of how to select subsets of event data and how to aggregate the subsets together to integrate various ubiquitous systems in real-world applications.

5 Conclusions and Future Works

This study shows how ontology/data engineering is efficient for context awareness, and proposes an optimal solution for the difficulty of network traffic management by supplying useful data which is generated from a network traffic analysis system to network administrators. The network traffic analysis system is designed and implemented based on System Entity Structure (SES) for a high throughput and a low response time. The SES representing the network traffic ontology helps the network traffic analysis system to transform captured network traffic data to practical data expressed by Extensible Markup Language (XML). The advantages of the SES using an ontology/data engineering concept are easy and fast access to information, no necessary for multi-query, fast response time, and user centric schema. Also, the SES data engineering is a hierarchical tree structure, therefore, it is easy to read, understand, and manipulate.

Due to the rapid increment of computer network uses, the understanding of overall network knowledge as well as physical devices connected to networks are essential for efficient management to solve various problems caused in network environments. Existing network traffic management systems have several problems: lack of analyzing methods, limitation of data storing, and excessive resource consumptions. These problems cause a slow analyzing time and require big budget for maintaining.

This system using the SES data engineering is a fast and efficient data management system to capture and analyze network traffic data. The graphical user interface (GUI) reporting windows reduce analyzing time. This efficient approach is useful not only for solving problems on network traffic analysis field but also for examining research areas of mobile communication system.

As a component, the network traffic analysis system could be used for modeling and simulation for complex systems by generating network traffic. For example, cyber security system (i.e., an intrusion detection system, an anomalous behavior analyzer) can be one application. As another field in which the network traffic analysis system can be used is involved with web services. In order to use the remote models more effectively and efficiently, the network traffic analysis system using data engineering methodology seems to be a promising approach.

References

1. Teorey, T.J., Yang, D., Fry, J.P.: A logical design methodology for relational databases using the extended entity-relationship model. ACM Computing Surveys (CSUR) 18(2), 197–222 (1986)
2. OMG. Unified Modeling Language (UML), Version 2.1.1, http://www.omg.org/technology/documents/formal/uml.htm

3. Sarjoughian, H., Flasher, R.: System Modeling with Mixed Object and Data Models. In: DEVS Symposium, Spring Simulation Multiconference, Norfolk, Virginia, March, pp. 199–206
4. Jordan, D., Russell, C.: Java Data Objects. O'Reilly, Sebastopol (2003)
5. Zeigler, B.P.: Multi-Facetted Modeling and Discrete Event Simulation. Academic Press, New York (1984)
6. Zeigler, B.P., Zhang, G.: The System Entity Structure: Knowledge Representation for Simulation Modeling and Design. In: Widman, L.E., Loparo, K.A., Nielsen, N.R. (eds.) Artificial Intelligence, Simulation and Modeling, pp. 47–73. Wiley, Chichester (1989)
7. Zeigler, B.P., Hammonds, P.E.: Modeling & Simulation-Based Data Engineering: Introducing Pragmatics into Ontologies for Net-Centric Information Exchange. Elsevier, Amsterdam (2007)
8. DEVS-Java Reference Guide, http://acims.arizona.edu
9. Ethereal - The world's most popular network protocol analyzer, http://www.ethereal.com/
10. SESBuilder, http://www.sesbuilder.com/
11. Document Object Model (DOM), http://www.w3.org/DOM/
12. JFreeChart, http://www.jfree.org/jfreechart/

A Scalable Framework for Distributed Ontologies

Ngot Phu Bui[1], Young-Tack Park[2], and TaeChoong Chung[1]

[1] Department of Computer Engineering, Kyung Hee University, Republic of Korea
{bpngot, tcchung}@khu.ac.kr
[2] School of Computing, Soongsil University, Republic of Korea
park@ssu.ac.kr

Abstract. In large-scale Peer-to-Peer (P2P) systems, information is usually generated decentralized. An architecture that enables applications to have a unified view on various information sources is currently attracting enormous attention. For building that architecture, many P2P oriented applications focus on resolving two challenges. The first one is making available information sources have a uniform format and the second is to allow querying on that information without regarding the heterogeneity of the underlying systems. In this paper, we propose a framework for integrating information from heterogeneous sources by exploiting two new technologies: the Semantic Web and the Grid computing. In our scenario, we fabricate the Semantic Web service on the top of Grid Services. Each data provider becomes an autonomous node in Virtual Organization of the Grid system. The implementation and experimental result demonstrate the feasibility of our approach.

Keywords: Semantic Web – Grid computing – Virtual Organization.

1 Introduction

The Semantic Web is envisioned as new generation of the current Web where, in addition to being human-readable using WWW browsers, documents are annotated with meta-data. This meta-data defines what the information is about in a machine understandable way, better enabling computers and people to work in cooperation. The explicit representation of meta-information, accompanied by domain theories (i.e. ontologies), will enable a web that provides a qualitatively new level of service. It will weave together an incredibly large network of human knowledge and will complement it with machine processable.

Large-scale P2P applications are undertaken through the interaction of people, heterogeneous computing resources, information systems, all of which are geographically and organizationally scattered. The Grid [4] is an emerging platform to support coordinated resource sharing and problem solving on a global scale for data-intensive and compute-intensive applications [1]. Until very recently the Grid and the Semantic Web communities were separate. Both have a need for computationally accessible and sharable metadata to support automated information discovery, integration and aggregation. Both operate in a global, distributed and changeable environment.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 242–251, 2008.

The Semantic Grid is a result of the convergence between the Semantic Web and the Grid technologies. It refers to an approach to Grid computing in which information, computing resources and services are described in standard ways that can be processed by computers. This makes it easier for resources to be discovered and joined up automatically, which helps bring resources together to create Virtual Organization (VO). In this research, we will focus on the Semantic Grid in which sharing resources are computing power and meta-data in ontologies. The descriptions constitute meta-data and are typically represented using the technologies of the Semantic Web, such as the Resource Description Framework (RDF) [5]. We describe an open architecture that enables any peer of the Internet to be capable of playing a role as a Grid node. These peers have to conform themselves to the VO policy while they are fully controlled by their own organization. The sharing resources of them are the RDFS meta-data ontologies and processing power for queries. VO in our system is responsible for '*managing*' the peers and their sharing resources.

The remainder of the paper is organized as follows: Section 2 introduces to related work about existing P2P network applications and sharing data on those systems. Section 3 presents the background knowledge for building today integrating information systems. Section 4 describes the architecture, operation mechanism, and sharing meta-data management strategy in Ontology Virtual Organization. We discuss our implementation and performance evaluation in Section 5. Finally, we conclude our research in Section 6.

2 Related Work

A generic system which can dynamically integrate heterogeneous information sources on a large-scale P2P networks is now becoming one of prominent researches that leads to the success of querying mechanism over the Internet environment. When establishing such systems, we have to deal with the two main challenges. The first one is how to collect, classify, represent and search the sharing information from various sources. The second one is the network infrastructure that can make the system more self-organizing, load balanced and fault-tolerant. Currently, there are many approaches for building information integrating frameworks to utilize huge available distributed information sources of the Internet. Gnutella [6] is a pure peer-to-peer application that allows searches to flow through a network of interconnected peers. Furthermore, Gnutella is fully decentralized in the aspects of no single point of failure and unsusceptible to the denial of service. In Gnutella, each peer forwards a search to all its immediate neighbors in a breadth first manner. This way for executing query leads to the flooding queries over the network and makes its searches processing distributed but still not scalable.

Edutella ([7], [8], [9]) relies on super-peers to act as local search hubs. Each super-peer is similar to a server for a small portion of the network. In other words, one super-peer in Edutella is responsible for managing how to store, access and update a part of the whole data and meta-data in the system. RDFPeers [11], a scalable distributed RDF repository, stores each triple at three places in addressable network by using global known hash functions to the triple's value. In EII ([12], [3], [16]), data centric integration model can create either a centralized repository for searching data or a data integrating layer of a set of different data sources such as federated information systems. However, representation of data model in current EII applications is not suitable for the rich semantic data environment. AmbientDB [14] is an architecture designed to supply fully relational database functions for standalone operation in autonomous devices. A DHT is used both as a means for connection peers in a resilient way as well as supporting indexing of data.

PeerDB [15] is a P2P application supporting searching against data store on remote nodes. PeerDB differs from our system in an aspect that it is based on an unstructured P2P system, focusing on data retrieval instead of distributed querying. PIER [13] is much similar to our approach. Its query engine and storage mechanism are built on the top of distributed hash table (DHT). However, our system is only based on indexing meta-data term of semantic information while PIER is essentially indexes whatever the applications register to the system.

3 Background Knowledge

In this paper, we propose a new paradigm for managing and querying heterogeneous information sources over large-scale P2P network. Currently, no single existing technology has power for building architecture like that. In this section, we will describe technologies that are combined for building decentralized data management supporting seamless querying to heterogeneous data sources in the Data Grid that conforms to the Semantic Grid mechanism.

3.1 The Semantic Web and Ontology

The advent of the World Wide Web has significantly influenced on the information sharing among people as well as computers. In order to access the needed information, we often use software agent based on keyword search engines such as Google, Yahoo, and AltaVista. However, these agent systems have significant weaknesses such as retrieving irrelevant information, failing to integrate different information sources so they do not sufficiently support the user's various search requests. In order to overcome these drawbacks of the current Web, Tim Berners-Lee [2] proposed the Semantic Web as the next generation of Web technology. It extends the current Web to make web information understandable to computers by giving it well-defined meaning or semantic data.

Ontologies have shown to be a key enabling technology for the Semantic Web. The domain model implicit in an ontology can be taken as a unifying structure for giving information a common representation and semantics. Today, the Semantic Web and ontologies are often developed by a particular organization. Ontology is built so that it can serve for a specific domain. As a result, a major problem is that the current infrastructures of the Internet and web browsers have no generic ways for combining all related information from various web resources.

3.2 Grid Computing and Web Services

According to ([2], [4], [18]), a Grid is a system with flexible, secure, coordinated resources sharing among dynamic collections of individuals, institutions and uses standard, open, general-purpose protocols and interfaces. On a Grid system, we somehow have to take a bunch of different resources that uses a wide variety of operating systems and architectures, and somehow make them collaborate to solve computational tasks. The only way to do this is by using "*standard, open, general-purpose protocols and interfaces*" that provide a "*common language*" that everyone on the Grid can understand.

Web Services are platform-independent and language-independent since they use standard XML languages. One of the most prominent attributes of Web Services is the fact they provide themselves naturally to fabricate loosely coupled systems. Actually, plain Web Services [19] would be inadequate for building grid applications. Grid-based systems are moving away from a collection of protocols to a services-oriented approach: the Open Grid Services Architecture (OGSA) ([20], [21], [22]). OGSA is capable of facilitating interoperability between such varied, heterogeneous, and distributed resources and services as well as lessen the complication of administering heterogeneous systems.

4 Ontology Virtual Organization

The real and specific problem that underlies the Grid concept is "*coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations*". The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem-solving and resource brokering strategies emerging in industry, science, and engineering. This sharing is, necessarily, highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs. A set of individuals and/or institutions defined by such sharing rules form what we call a *virtual organization* (VO). In our case, we create Ontology Virtual Organization (OVO) in which the sharing resources are the meta-data of the ontology under triple format as known as atomic unit and the computing power for preprocessing the query and post processing for atomic query results.

4.1 Architecture

Figure 1 briefly shows the overview of OVO architecture and its operation. OVO is constituted by many sharing resources nodes from different organizations. Broker is responsible for "*managing*" all sharing resources. The first type of sharing resource is the ontology meta-data from various semantic sources over the Internet. The sharing meta-data of ontology in OVO is stored under triple format <*predicate (subject, object)*> [5]. The second type of sharing resource is the computing power for processing query to the ontology meta-data. Peer data nodes, in which Ontology Service resides, are any computers that want to share their semantic data, ontologies. Computation nodes, in which the Union Service or Relaxation Query Service resides, are the computers eager to share their computing power (i.e. CPU resources). In figure 1, we also represent an example Grid Services at work in OVO through the order of interaction among the six OVO's participants from six simple hosting environments. One that runs the user application; one that encapsulates computing and meta-data resources; two that encapsulate ontology services; one for Union Service and the last for Relaxation Query Service. We describe more details about this operation mechanism as follows:

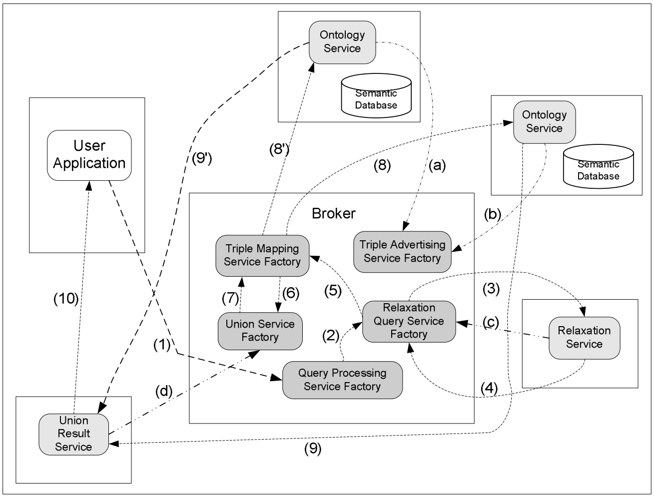


Fig. 1. The overall OVO architecture and its operation

- Host, running the user application, invokes create Service request on "*Query Processing Service Factory*" (QPSF) in the broker host of OVO, requesting the creation for "*Query Processing Service*" (QPS) (1). QPS invokes Relaxation Query Service Factory (RQSF) on the same host for creating Relaxation Query Service (RQS) (2).
- RQS on broker host does the discovery on its services registration database and determines appropriate remote "*Relaxation Service*" (RS) (3), then submitting the query as parameter to that RS. After finishing relaxation, remote RS will deliver the result to RQS of broker (4).
- RQS on the broker invokes "*Triple Mapping Service Factory*" (TMSF) on the same host for creation "*Triple Mapping Service"* (TMS) (5). TMS then invokes "*Union Service Factory*" (USF) for creation "*Union Service*" (US) (6). US will return to TMS the remote Union Result Service (UPS) (7).
- TMS continues discovering the suitable services in the triple advertising service database for answering the triple and then passing the parameters (triple, UPS, …) to the remote "*Ontology Services*" (OS) (8, 8'). After finishing processing triple, OS will deliver the results to remote UPS (9, 9'). US then carries out it uniting the partial results and send complete answer to the user application (10).

4.2 Grid Services in OVO

The Web Service Resource Framework (WSRF) [23] defines conventions for the managing state in distributed systems based on Web services. WSRF provides the stateful services that OGSA needs. In other words, WSRF is the infrastructure on which OGSA architecture is built on. On its turn, OGSA represents everything as a Grid service: a Web service that conforms to a set of conventions and supports standard interfaces for such purposes as lifetime management.

Our OVO involves several services interacting with each others. The interactions among those services can occur in same host or different remote hosts in network. Communication of services between different hosts often happens with the participating of broker, peer data, and computation nodes. Query Service, Triple Mapping Service, Triple Advertising Service, Relaxation Query Service, and Union Service are the services that reside in broker. Peer data has Ontology Service while computation node has Union Result Service and Relaxation Service. The following describes the functions of each service.

- Query Service: This service plays a role as a start access point for anonymous users who need to query the meta-data resources in OVO. It is responsible for validating the legal of the query to the system.
- Relaxation Query Service: This service is invoked by Query Service. Query relaxation [25] and query atomic triple conversion [24] are the tasks of this service. Computation node can register its computing resource to this RQSF (in figure 1 (c)).
- Triple Mapping Service: Similar to index service. TMS traverses the remote service database for processing triple to find relevant services.
- Triple Advertising Service: Peer data can advertise their ontology meta-data resource by registering triples of ontology to Triple Advertising Service Factory (in figure 1 (a) and (b)).
- Union Service (in broker): Computation nodes can announce their computing resource by registering computing capacity to USF (in figure 1 (d)).
- Ontology Service: Residing in peer data node, this service executes the query for receiving triple to its own semantic database.
- Union Result Service and Relaxation Service: Residing in the computation nodes, these services use their CPUs for executing the uniting partial results of a query and doing the relaxation for a complex query respectively.

4.3 Ontology Metadata Storing in OVO

The Semantic Web and databases technologies present enormous opportunities. A regular data model for aggregating result drawn from various sources could use RDF(S). Ontologies, the core of the Semantic Web, can present many domains of knowledge whilst being machine understandable by using OWL/ DAML+OIL/ RDF(S). In this section, we will describe how to "*manage*" multiple ontologies of diverse domains from different sources.

All actual data related to ontology is stored and fully controlled by its real owner, a peer data. One ontology often consists two RDF(S) files, schema and instance ones. RDF(S) documents are composed of a set of RDF triples. Each triple is in the form of

<*predicate (subject, object)*>. The subject is the resource about which the statement was made. The predicate is a resource representing the specific property in the statement. The object is the property value of the predicate in the statement. The object is either a resource or a literal; a resource is identified by a URI; literals are either plain or typed and have the lexical form of unicode string. In OVO, peer data will register all their triples in schema file of ontology to broker. In order to support efficient queries on distributed RDF triples, we exploit P2P routing substrates for building a distributed index for these triples. We store each triple according to the hash key value when applying SHA_1 hash function [29] to the its value. Storing strategy is as following:

- Assume we have $N = 2^k$ broker nodes in the Grid.
- We will map each triple of ontology to each broker node by using SHA_1 hash function that produces hash key with $log_2(N)$ bits. Triple is stored in broker whose node identifier (*k bits* number) is the numerically closet to the triple's hash key.
- Each entry on index table of broker node has the format (key, value, peer-list) where
  - key = SHA_1Hash(value); key is a *k bits* number
  - value = ({URI} , {URI}, { URI | literal })
  - peer-list = {id_1, id_2, …, id_n} (list of peer that contain information corresponding to the triple; identifier of peer can be peer IP address).

5 Implementation and Evaluation

In this section, we present our implementation a system for distributed ontologies based on grid service technology. Then, we will evaluate the performance of our framework by investigating the capacity of the system for query answering. In evaluation phase, we will focus on result of the system according to two factors: network size and ontology size.

5.1 Implementation Environment

We have implemented the system using Java on Linux operating system. We inherit somewhat from our previous implementation [17]. In other words, the principal technical features of the framework are highlighted as below:

- The base development and deployment platform is Globus Toolkit 4.0.2 (GT4) [26]. We got GT4 source version for Red Hat operating system.
- We build all the services in OVO based on OGSA standard.
- We use Protégé 3.1 [27] to create ontology in RDF(S) format and Sesame 2.0 [28] to store, read, process RDF(S) data.

5.2 Routing Hops for Resolving Query

Efficient factor for evaluating the performance of structure P2P network is the number of hops taken to resolve the query. For this experience, we create ontologies with the total of 800,000 concepts corresponding approximately 350,000 triples. Due to the limited number of physical nodes, we used 1024 logical nodes. There are two kinds of query used for this evaluation. The first one is atomic query and the second is complex query. [*{?University} locate {City 'New York'}*] is a kind of atomic query in which the user wants to know which *universities* locate in the *city* whose name is '*New York*'.

Table 1. Queries used for evaluation

No	Query
1	$\{C_1\}\ p_1\ \{C_2\}$
2	$\{C_1\}\ p_1\ \{C_2\}$ $p_2\ \{C_3\}$

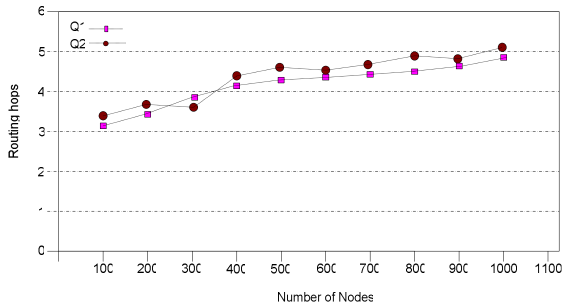


Fig. 2. The average of routing hops

Figure 2 shows the average routing hops (ARH) number when submitting query Q_1 or Q_2 to the system. According to our analysis, ARH value vacillates from 0 to log_2N (N is the number of nodes). Q_2's ARH is slight more than that of Q_1 because of relaxation cost for Q_2.

5.3 Query Processing Complete Time and Data size

The second essential factor for evaluating the efficiency of P2P networks architecture is the relationship between complete time and data size according to a specific network structure. We will investigate this factor with the network of two broker nodes.

Table 2. Triples in Ontologies

No	Query
1	$\{C_1\}\ p_1\ \{C_2\}$
2	$\{C_1\}\ p_1\ \{C_2\}$ $p_2\ \{C_3\}$

Table 3. Physical nodes used for experiment

Attr / Node type	Quantity	Memory	CPU
Broker	2	512MB	PIV 2.4 GHz
Peer Data	2	512MB	PIV 2.8 GHz
Computation	1	512MB	AMD 1.0 GHz

Figure 3 shows the performance measurement between completion time and data size in triple. In this evaluation, we also use the same kinds of query of the previous one. We investigate the completing time results for the queries with the data size from 30,000 triples to 350,000 triples of the ontologies. Time for processing Q_2 is more than that of Q_1 because of the cost for relaxation query and uniting partial results.

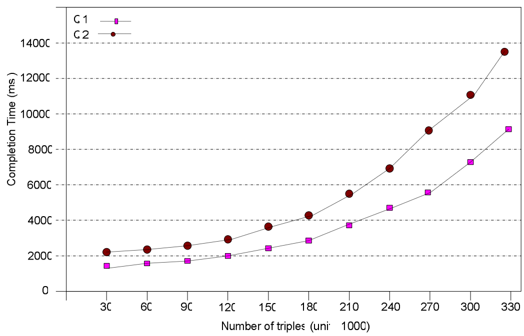


Fig. 3. Completing time vs. data size

6 Conclusion and Future Work

In this paper, we have presented a grid-based framework applied for distributed ontologies on structured P2P network. Our system has some novel aspects on both the infrastructure for P2P network and sharing resources across Grid system. We focus on the design of Ontology Virtual Organization so that it enables us to manage sharing resources efficiently. We take advantage of using Grid Services in building a powerful interaction mechanism for all nodes in Grid system, a heterogeneous P2P network environment.

Currently, our approach is suitable for routing and answering atomic triple or simple query. In other words, our framework has no feature for supporting real complex queries that require joining partial result data from peer data nodes. Our future work will be towards the development of a more reasonable routing algorithm for solving the above issue.

References

1. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organization. International J. Supercomputer Applications 15(3) (2001)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, May 2001. Scientific American (2001)
3. Paton, N., Atkinson, M., Dialani, V., Pearson, D., Storey, T., Watson, P.: Database Access and Integration Services on the Grid. UK e-Science Programme Technical Report Number UKeS-2002-03
4. Foster, I., Kesselman, C. (eds.): The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
5. Brickley, D., McBride, B. (eds.): RDF Schema, `http://www.w3.org/TR/rdf-schema/`
6. The Gnutella protocol specification (2000), http://dss.clip.com/GnutellaProtocol04.pdf
7. Neidl, W., Wolf, B., Siberski, W., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmer, M., Risch, T.: EDUTELLA: P2P Networking Infrastructure Based on RDF. In: 11th World Wide Web Conference (2002)

8. Nejfl, W., Wolpers, M., Siberski, W., Schmitz, C., Schlosser, M., Brunkhorst, I., Löser, A.: Super-Peer-Based Routing and Clustering Strategies for RDF-Based Peer-To-Peer Networks. In: Proceeding of the 12th International World Wide Web Conference (WWW), Budapest, Hungary (2003)
9. Brunkhorst, I., Dhraief, H., Kemper, A., Nejdl, W., Wienser, C.: Distributed Queries and Query Optimization in Schema-Based P2P-Systems. In: International Workshop on Databases, Information Systems and Peer-to-Peer Computing, September 2003, Berlin, Germany (2003)
10. OGSA-DAI, Open Gird Services Architecture Data Access and Integration (2006), http://www.ogsadai.org.uk/
11. Cai, M., Frank, M.: RDFPeers: A Scalable Distributed RDF Repository based on A Structure Peer-to-Peer Network. In: Proceeding of the 13th International World Wide Web Conference (WWW), New York (2004)
12. Halevy, A.Y., Ashishy, N., Bitton, D., et al.: Enterprise Information Integration: Successes, Challenges and Controversies. In: ACM SIGMOD 2005, pp. 778–787 (2005)
13. Huebsch, R., Hellerstein, J.M., Lanham, N., Loo, B.T., Shenker, S., Stoica, I.: Querying the Internet with PIER. In: Proceedings of the 29th VLDB Conference, Berlin, Germany (2003)
14. Boncz, P.A., Treijtel, C.: AmbientDB: Relational Query Processing in a P2P Network. In: Aberer, K., Koubarakis, M., Kalogeraki, V. (eds.) VLDB 2003. LNCS, vol. 2944, pp. 153–168. Springer, Heidelberg (2004)
15. Ng, W.S., Ooi1, B.C., Tan1, K.L., Zhou, A.: PeerDB: A P2P-based System for Distributed Data Sharing. In: ICDE 2003, Bangalore, India, pp. 633–644 (2003)
16. Sheth, A.P., Larson, J.A.: Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys 22(3), 183–236 (1990)
17. Bui, N.P., Lee, S., Chung, T.: Query Distributed Ontology over Grid Environment. In: Gervasi, O., Gavrilova, M.L. (eds.) ICCSA 2007, Part III. LNCS, vol. 4707, pp. 140–153. Springer, Heidelberg (2007)
18. Foster, I.: What is the Grid? A Three points checklists. Grid Today (July 20, 2002)
19. Web Services. http://www.w3.org/2002/ws/
20. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: A Open Grid Services Architecture for Distributed Systems Integration. Draft of 6/22/02, http://www.gridfrom.org/ogsi-wg/draft/ogsa_draft2.9-2002-22.pdf
21. Tuecke, S., Foster, I., Frey, J., Graham, S., Kesselman, C., Maquire, T., Sandholm, T., Snelling, D., Vanderbilt, P.: Open Grid Services Infrastructure (OGSI) Version 1.0. Global Grid Forum Draft Recommendation (6/27/2003), http://www.globus.org/alliance/publications/papers/Final_OGSI_Specification_V1.0.pdf
22. Foster, I., Kishimoto, H., Savva, A., Berry, D., Djaoui, A., Grimshaw, A., Horn, B., Maciel, F., Siebenlist, F., Subramaniam, R., Treadwell, J., Von Reich, J.: The Open Grid Services Architecture, Version 1.0. Informational Document, Global Grid Forum (GGF) (January 29, 2005), http://www.gridforum.org/documents/GWD-I-E/GFD-I.030.pdf
23. WSRF.NET, http://www.cs.virginia.edu/~gsw2c/wsrf.net.html
24. Cai, M., Frank, M.: RDFPeers: A Scalable Distributed RDF Repository based on A Structure Peer-to-Peer Network. In: Proceeding of the 13th International World Wide Web Conference (WWW), New York (2004)
25. Tempich, C., Staab, S., Wranik, A.: REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors. In: WWW 2004, May 17–22, 2004, New York (2004)
26. http://www.globus.org/toolkit/downloads/4.0.2/
27. Protégé, http://protege.stanford.edu
28. Sesame 2.0, http://www.openrdf.org
29. FIPS 180-1. Secure Hash Standard. Technical Report Publication 180-1, Federal Information Processing Standard (FIPS), National Institute of Standards and Technology, US Department of Commerce, Washington DC (April 1995)

Bridging Semantic Gap in Distant Communication: Ontology-Based Approach

Nilar Aye, Fumio Hattori, and Kazuhiro Kuwabara

College of Information Science and Engineering
Ritsumeikan University
1-1-1 Noji Higashi, Kusatsu, Shiga-ken 525-8577 Japan
gr042064@se.ritsumei.ac.jp, fhattori@is.ritsumei.ac.jp,
kuwabara@is.ritsumei.ac.jp

Abstract. This paper introduces a multi-agent system that applies semantic web technology to bridge the semantic gap between communication users with different knowledge levels. The system aims to provide smooth conversation with the assistance of semantic information to share the information by users with higher knowledge levels. In this scenario, agents act as personal assistants who monitor the user information and the exchanged contexts. Agents communicate with each another to transfer knowledge to various users depending on their background knowledge levels. Therefore, although the conversation topic is identical, the provision of information will be different. In this paper, we present a sample scenario of computer usage operations in an English composition class that can effectively be applied in a teaching environment. We create domain ontology to perform mapping with the teacher's ontology for knowledge extraction and content conversion. Bridging the semantic gap helps users to communicate easily without considering knowledge barriers.

Keywords: Semantic Gap, Multi-agent System, Ontology Mapping.

1 Introduction

Information exchange and knowledge sharing among people with the assistance of advanced communication tools that provide both audio and video functions can effectively overcome the limitations of geographical distance and the boundaries of sovereign nations. However, a semantic gap often occurs while communicating due to the different levels of education or second/foreign language ability. Although many communication tools are widely used, more attention is given to establish and maintain a connection among users to facilitate interactive communication.

We propose a framework that "fills" the semantic gaps between users with different knowledge levels. Such situations can take place in a conversation between a teacher and students who possess varying levels of knowledge on a particular conversation topic. We try to solve not only the knowledge gap problem between a teacher and students but also the dissimilar education backgrounds among the students

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 252–260, 2008.

themselves. We adjust and maintain an equivalent knowledge level of users with the help of a semantic database. Detailed information is extracted from a domain ontology and provided to beginners. Higher level information is only supplied if the user is an expert with a sufficient and corresponding knowledge level.

Our system can be applied in e-learning that supplements the necessary information for users. It is not, however, targeted as a teaching system. Rather, it is especially targeted for a teacher's assistant. For example, suppose an English composition class in which a teacher is imparting the writing assignment with a Microsoft Word application. When the teacher gives an instruction, students with less background knowledge about Microsoft Word or personal computer operations face a problem. They might have different backgrounds such as economics, information technology, medicine, or others, complicating their ability to clearly understand the teacher's instruction. In such cases, these students may require supplementary explanations to follow the teacher's instructions depending on their computer literacy. Our system can be applied in this situation so that a teacher provides the meanings of keywords selected from the lecture and students extract the information accordingly.

Filling semantic gaps benefits communication users. In other words, it is possible to enhance the understanding of users (or students in this case) on a specific topic. Depending on the user level that might range from expert to novice, relevant knowledge is extracted from the ontology and presented. In fact, the semantic gap stands out as the main barrier for smooth conversation among online users.

This paper is structured as follows. Section 2 describes the topic of semantic information support for distant communication used to clarify the approach taken in this paper. Then in Section 3 we present semantic conversion by applying ontology mapping, which is essential for knowledge representation and sharing among users. Section 4 explores a system implementation that covers multi-agent architecture. Finally, Section 5 concludes and points out issues for further developments.

2 Semantic Information Support for Distant Communication

With the help of semantic information, distant communication can be applied in various circumstances. Our system is augmented with video and chat functions for real-time instant messaging to enable conversations where users are remotely connected to each other. Initially, the same predefined conversation topics/lists [1] selected from lectures will be displayed on both user screens. This scenario includes the menu items of Microsoft Word and file operations (Fig. 1).

Suppose a teacher is instructing how to prepare a paper using a computer in an English composition class. Her instruction probably includes the following document functions: creating a new one, opening and updating existing ones, formatting, and so forth. When learning how to copy the assignment to a USB memory, some students want to know how to perform this operation. Clicking this "File Copy" keyword

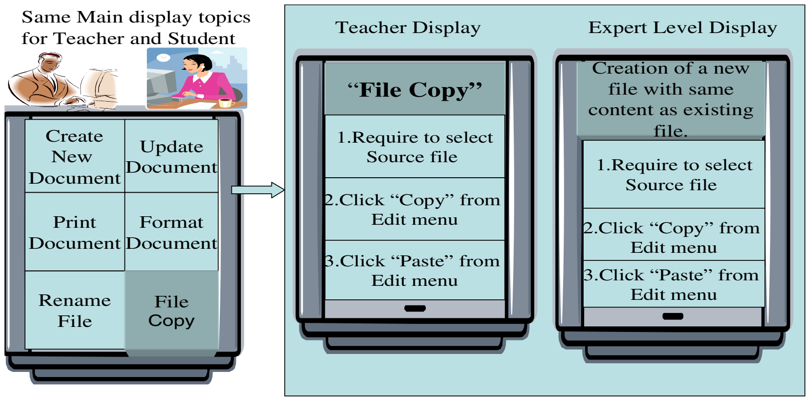


Fig. 1. Screen display for Teacher and Expert Level student after clicking "File Copy" button from Main Menu

provides such information as "Require to select Source file", "Click Copy from Edit menu," and "Click Paste from Edit menu." For an expert level student with experience in an information technology related field, these three steps are sufficient. For intermediate and novice level students, however, the teacher can provide more detailed information by alternatively clicking the displayed sub-topics on the touch screen. Semantic information combined with images is provided to beginners to enhance their understanding. Therefore, the provision of information varies from student to student (Fig.2).

Situations are possible where insufficient or inadequate information is provided to some students. In such cases, students can ask for more information by clicking a

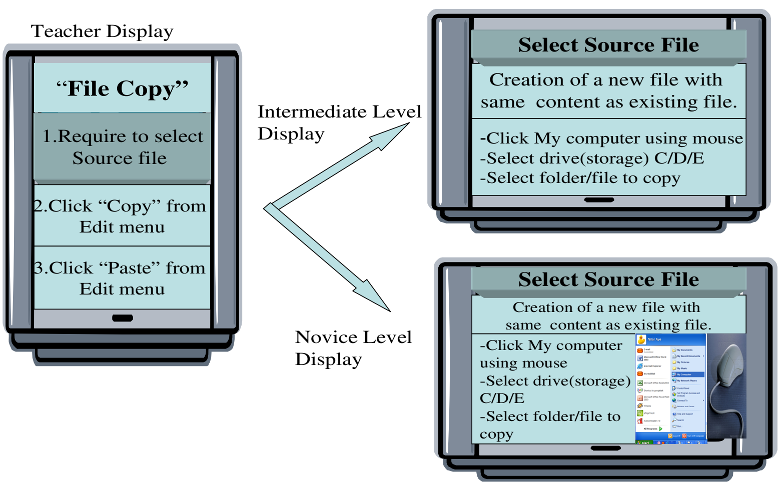


Fig. 2. Different display screen for students including meaning and operation steps of "File Copy"

detail button on the screen. The student agent passes this message to the teacher agent to support one lower level of information, for example, an intermediate level of information to experts and novice level information to intermediate students. But for beginners, a teacher agent searches for more detail information in the domain ontology. If available, this information is provided to the student agent. Otherwise, a human teacher explicitly solves this problem for a particular student after receiving a help message from his/her agent.

3 Semantic Conversion Using Ontology Mapping

In this section, we explore the domain ontology development steps and the semantic conversion processes. Ontology includes a vocabulary that provides a set of well-founded constructs to build a meaningful higher level of knowledge for specifying the semantics information in a well-defined and unambiguous manner [2]. In this framework, we create a domain ontology to carry out contents conversion for information sharing. The teacher's abstract ontology includes a list of operations of Microsoft Word and computer related operations that are used to map with the domain ontology to extract relevant metadata (Fig. 3).

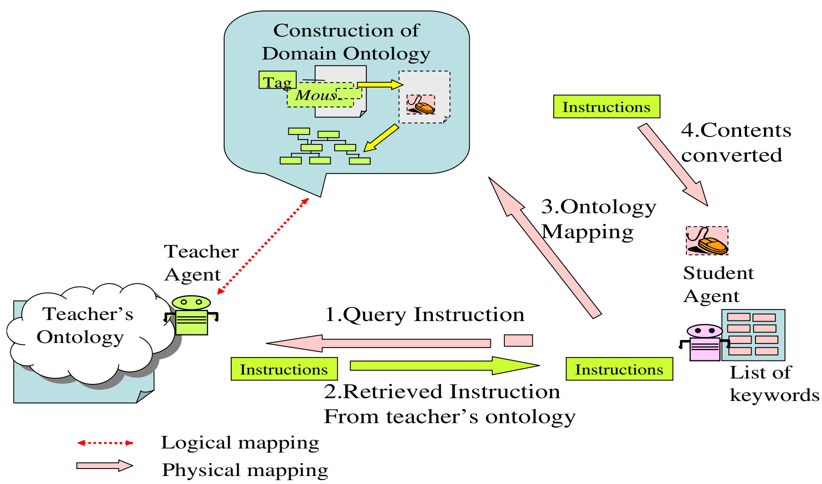


Fig. 3. Semantic conversion using ontology mapping

3.1 Developing Ontology

Developing ontology, which defines a set of data and its structure for other programs to use, is essential for the abstraction of a particular domain [3]. Required steps for ontology development can be considered as a preparation of teaching materials for a lecture. For example, the preparation may include extracting keywords from the topic, collecting definitions and associated meanings of extracted keywords, gathering related examples and supportive figures, and putting up sample scenarios. After that, combining all contents, defining relationships between keywords and classifying the hierarchical structure are essential.

Similarly in ontology creation, we first determine the scope related to computer usage, such as the Microsoft Word commands and file operations. Then, we enumerate such terms as “Create new document,” “Open document,” “File Copy,” and so on. Then we define the taxonomy (subclass) hierarchy of each term or task. The “File Copy” keyword includes three subclasses: “select source file,” “copy source file,” and “paste copied file.” After that, we identify its properties as “detail”, “meaning”, “isComposedOf”, “image”, etc. Then, instances of the classes are created. Finally, inconsistencies are detected in the ontology itself that was defined to populate the ontology.

We use Resource Description Framework (RDF) that is a simple general purpose metadata language for representing information in the web, provides a model for describing and creating relationships between resources. The structure of expression in RDF can be viewed as a directed labeled graph [4] built upon the provided keywords of the conversation topic, which has been decomposed into several steps to provide information for various levels of users (Fig. 4). Initially, the structure begins from one main keyword and continues traversing with related subtasks that include resources, attributes, and literals that may be text as well as images.

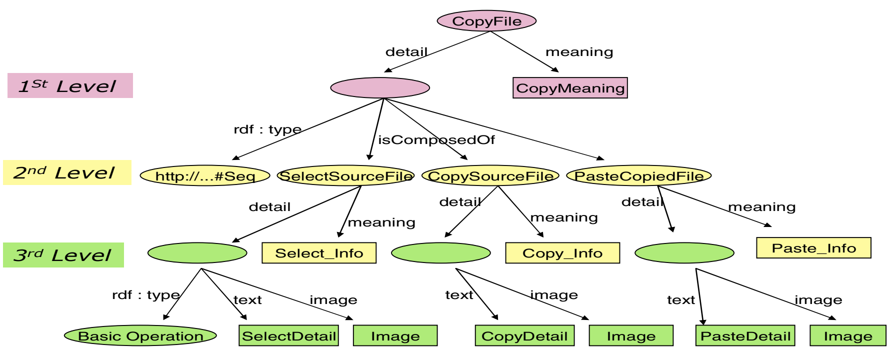


Fig. 4. Domain ontology of File Copy operation with RDF model

3.2 Contents Conversion

In this step, agents convert the teacher’s abstract instruction into relevant information with the help of the domain ontology. Contents conversion process involves an ontology mapping to check the correspondence between the teacher’s ontology and the domain ontology. Mapping helps understanding the relationship between different ontologies. Although the mapping operation is identical for student agents, the retrieved information will be different, such as first, second, or third level tasks, accordingly.

Our system presents a main display with a list of computer operations or keywords selected from lectures. Students can choose from any number of desired topics one after another. In that situation, by only clicking one button, for example, “file copy,” a student agent sends a "query" operation to get instructions or the meaning of that keyword from the teacher. An agent continues the ontology mapping process after receiving the query results. Consequently, suitable metadata are extracted from the domain ontology by a student agent. In other words, a teacher's abstract instruction that originated from pressing a button is converted into relevant explanations, meanings, or images that appear on the students' screen.

4 System Overview

In this section, we provide the design of multi-agent architecture, an agent communication protocol used for inter-agent communications, and system implementation.

4.1 Multi-agent Architecture

We apply a combination of Semantic Web and Software Agent technologies for smooth conversation of real-time communication. In our design, we create user-agents that act as personal assistants who monitor user information exchanges and contexts. Such an agent is classified as an information agent based on the agent technology roadmap [5]. Our agents record user exchanged topics and levels and trace how users handle the information resources [6] in the form of semantic tags that provide the meaning of the keyword. These tags are also used to express the relationship between data and resources.

After communication has been setup, agents send messages to each other to exchange information (Fig. 5). In this scenario, a teacher agent communicates with many student agents. For example, when a student wants to know the details of how to copy a file, he clicks this keyword. That student's agent sends a query message to the teacher agent to retrieve relevant instructions from the teacher's ontology. As a result, the teacher agent returns retrieved instruction when matches occur. Then the student agent maps with domain ontology using predefined rules to extract a suitable level of data that corresponds to the teacher's abstract instructions. The presentation of the contents of the retrieved data depends on the student agent.

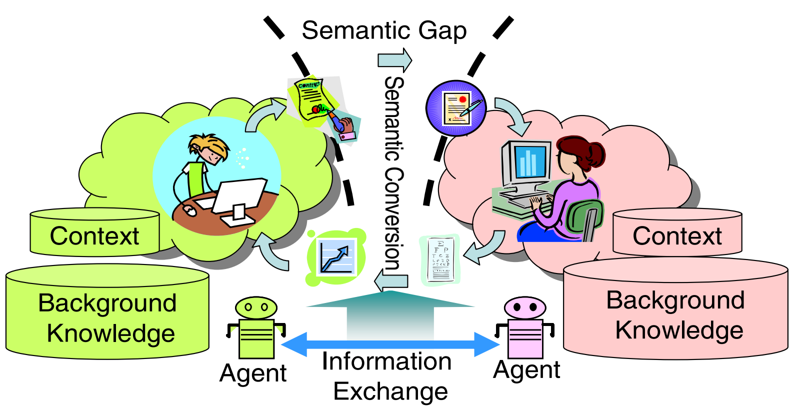


Fig. 5. Multi-agent architecture for bridging semantic gap

4.2 Agent Communication Protocol

In a multi-agent community, agents utilize a common language to understand messages while communicating with each other. Agents send and receive messages, question and inform other agents, request services, and monitor objects and values through a communication network. The messages can be of several types. In our design, the agent functions as a peer with other agents who all can perform active and passive roles in a dialog. Therefore, it can make and accept both assertions and queries [7].

Here are some examples of inter-agent communication messages utilized in our scenario (Table 1). An agent sends a "query" message to retrieve an abstract instruction from a teacher's ontology. Then the teacher agent responds with a "reply" message together with retrieved data. An "abort" message might be sent if the student clicked the wrong button to seek information. The "inform" message is used to acknowledge that the student agent successfully received the message. While processing ontology mapping and knowledge extraction, the student agent sends a "report" message to inform the teacher agent of its status. In addition, a student agent dispatches a "request" message to ask for more detailed information if the current provision is inadequate. Finally, a student agent passes the extracted information as a parameter of an "inform" message to let the teacher agent know the displayed information on the student screen.

Table 1. Summary of messages used in scenario

Message Type	Sender/Receiver of message	Meaning
QUERY	Student agent to Teacher agent	- To retrieve instruction from teacher's ontology
REPLY	Teacher agent to Student agent	- To return retrieved data
ABORT	Teacher agent to Student agent	- To reject wrong operation or information request
INFORM	Student agent to Teacher agent	- To acknowledge message was received
INFORM	Student agent to Teacher agent	- To notify displayed information
INFORM	Student agent to Teacher agent	- To let teacher agent know student's level
REQUEST	Student agent to Teacher agent	- To ask for more detailed information
REPORT	Student agent to Teacher agent	- To tell status

4.3 Implementation

In our system, we use Agent Communication Language (ACL) to communicate among agents because it provides agents with a means of exchanging information and knowledge [8]. Communication protocols are typically specified at several levels. While the lowest level of the protocol specifies the method of interconnection, the middle level specifies the format or the syntax of the information being transferred, and the top level specifies its meaning or semantics. We use an XMPP (Extensible Messaging and Presence Protocol) protocol for both instant messaging (intended for human users), and agent message transport. The aim of using this protocol is to apply the features of near-real time instant messaging (IM) [9] and agent messaging at the same time. XMPP, which is provided by the Jabber Software Foundation [10], is applicable for multi-agent systems with side by side for both simultaneous chat and agent message transport. This messaging framework creates an ideal method to transport SPARQL queries to various terminals. SPARQL is a query language [11] that is involved with RDF data retrieval.

Our system implementation (Fig. 6) mainly includes part of the user interface along with text chat and Windows NetMeeting for audio and video messages. User profiles are also included to store user personal information and past records. Additionally, the domain ontology creation module provides ontology mapping and conversion of the information contents. In this approach, a student agent extracts a keyword from the user interface module and sends a query message to the teacher agent applying a XMPP protocol. After receiving a "reply" message, a student agent extracts information from the domain ontology using SPARQL. The presence of user profiles helps to support appropriate levels of information for students.

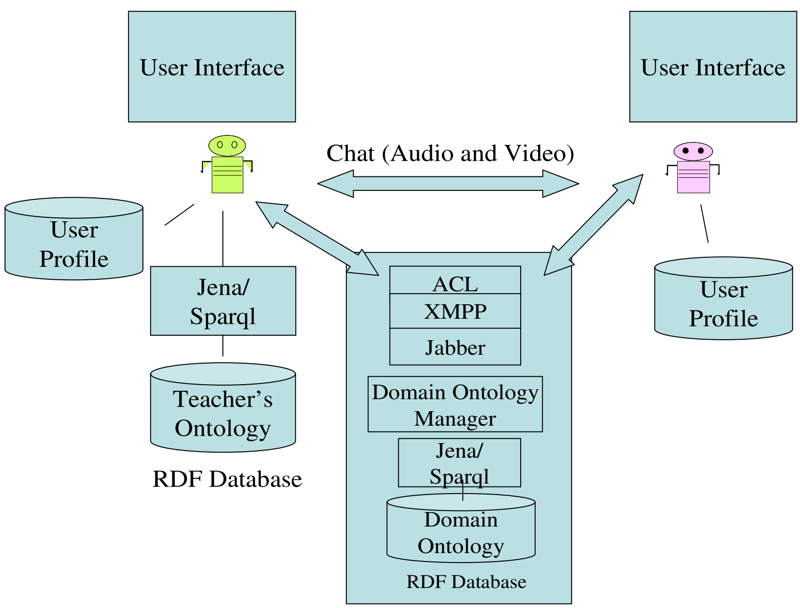


Fig. 6. System implementation

5 Conclusion and Future Plan

We proposed a system to bridge the semantic gap using a combination of Software Agent and Semantic Web technologies that are useful for effective information exchange specifically intended to fill the knowledge gap in such situations as doctor/patient in medical care [12], manager/customer at remote help desks, and teacher/student in teaching environments. The system's main advantage is transferring the required amount of information to provide a smooth conversation among the diverse knowledge backgrounds of users. We integrate communication tools with semantic information, which is greatly helpful for online users, using an XMPP protocol that can simultaneously transmit not only audio and chat but also agent messages.

To increase conversation efficiency, complete and good quality semantic knowledge must be specified. Applicable ontology mapping is also a significant challenge for successful system implementation. The relationship between keyword vocabularies must also be precisely defined.

Finally, the system leads to the development of efficient online communication by facilitating a conversation topic with additional text, images, animations, and video-clips [13] to incorporate annotations. Moreover, future work will increase the effectiveness of conversation using different languages under various knowledge backgrounds with the assistance of language translation functions integrated with semantic information.

References

1. Yasuda, K., Nemoto, T., Takenaka, K., Mitachi, M., Kuwabara, K.: Effectiveness of vocabulary data file, encyclopedia, and Internet homepages in a conversation support system for people with moderate-to-severe aphasia. Aphasiology 21(9), 867–882 (2007)
2. Cardoso, J., Sheth, A.: Semantic Web Services. Processes and Applications, 253–272 (2006)
3. Antoniou, G., Harmelen van, F.: A Semantic Web Primer, pp. 205–208 (2004)
4. http://www.w3.org/TR/rdf-primer
5. Luck, M., McBurney, P., Preist, C.: Agent Technology Enabling Next Generation Computing. AgentLink (2003), http://www.agentlink.org/roadmap/index.html
6. Kamei, K., Yoshida, S., Kuwabara, K., Akahani, J., Satoh, T.: An Agent Framework for Inter-personal Information Sharing with an RDF-based repository. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 438–452. Springer, Heidelberg (2003)
7. Weiss, G.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, pp. 83–92. MIT Press, Cambridge (2000)
8. Labrou, Y., Finin, T., Peng, Y.: Agent Communication Languages: The Current Landscape. IEEE Intelligent System 14(2), 45–52 (1999)
9. http://en.wikipedia.org/wiki/Extensible_Messaging_and_Presence_Protocol
10. http://www.jabber.org/about/overview.shtml
11. http://www.w3.org/2001/sw/DataAccess/rq23/#defn_SPARQLquery
12. Hattori, F., et al.: Socialware for People with Disabilities. In: Proc. 6th Int. Conf. on Cognitive Informatics (ICCI 2007), pp. 321–326 (2007)
13. Kuwahara, N., Abe, S., Yasuda, K., Kuwabara, K.: Networked reminiscence therapy for individuals with dementia by using photo and video sharing. In: ASSETS 2006: The 8th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 125–132 (2006)

Using Meta-agents to Reason with Multiple Ontologies

Ronald Hartung[1] and Anne Håkansson[2]

[1] Department of Computer Science,
Franklin University,
201 S. Grant Avenue,
Columbus, Ohio 43215, USA
Hartung@franklin.edu

[2] Department of Information Science, Computer Science,
Uppsala University,
Box 513, SE-751 20,
Uppsala, Sweden
Anne.Hakansson@dis.uu.se

Abstract. The semantic web uses ontologies to improve searching since ontologies provide a richer semantic model of content by expressing terms and relationships. However, a problem with the web is the large number of ontologies, which complicates searching. To maximize the capability of the search, the ontologies need to be combined to obtain complex answers. Usually, the ontologies are created without following any guidelines and, therefore, combining them is not a simple task, especially when ensuring a consistent result. We propose using meta-agents on top of software agents in a multi-agent system to resolve the use and the combination of multiple ontologies and to enable searching and reasoning. The software agents search for parts in the ontologies corresponding to the user-request and meta-agents combine the results from the agents. The meta-agents also partition the ontologies into consistent sets and then combine multiple ontologies into meaningful and logically consistent structures. For this purpose, we extend an existing mapping and alignment algorithm used for communication between agents. The use of multi-agents gives advantages since they provide a parallel approach and, thereby, efficiently handle large numbers of ontologies in order to accomplish tasks separately.

Keywords: Semantic Web 2.0, Ontology, ontology mapping, reasoning, software agents, intelligent agents, meta-agents, multi-agent systems.

1 Introduction

Ontology has become a common theme in the web community. As Web 2.0 and the semantic web are developed and expanded, the notion of ontology became a core component providing a controlled vocabulary of concepts, i.e., an explicitly defined and processable semantics and syntax. Currently, there are a large number of ontologies on the web, well over 13000 of them, and the number is continuously

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 261–270, 2008.

increasing. Querying languages have been designed and reasoning systems are under way [3; 11; 15]. For example, the overall structure of the Web 2.0, described by Hendler [9], shows a hierarchy of layers, which includes using a reasoning layer over the ontology layer. The concept of reasoning over ontology is at the heart of semantic searching and is the extension of keyword based searching allowing higher precision in search, e.g., increasing relevancy of the search results.

Systems with heterogeneous ontologies arise in both searching and agent communication. A proposed solution is mapping, or aligning, terms between the heterogeneous ontologies and the efforts to study approaches for mapping and merging ontologies, are well advanced [5; 1] There approaches tend to be based on lexical and/or structural mapping methods [10]. The approaches show some success for increasing cognitive support for effective matching in small examples of searching with ontologies. But these approaches are designed for a one to one matching, which is too slow and cumbersome to deal with the vast number of ontologies on the web.

The approach taken in this paper is based on several observations. For instance, a search is unlikely to use all the contents of ontology. Moreover, a single ontology is unlikely to contain all the relevant terms or relations for a search and the use of several ontologies are not guaranteed to be logically consistent with each other. The conclusion is that joining terms and relations from ontologies can form a new composite ontology that is a domain model in response to the search terms. Further, the ontologies do not need to be completely mapped, just portions that are related to the terms of the search. The composite ontology is not a static entity, but is constructed in response to a specific search, and can be extended as the search progresses. The combined ontology must be constructed to provide consistent information. The inconsistencies between ontologies can lead to constructing multiple composite ontologies, each of which is internally consistent. The benefit is that the use of multiple composite ontologies allows larger sections of the web to be searched.

In this paper, we present a multi-agent system to construct composite ontologies. The system uses software agents to search ontologies and meta-agents to form a set of consistent ontologies. The logical reasoning process, used for constructing the composite ontologies, is based on a mapping algorithm that is initially developed for agent communication and on an application of transitive reasoning. The use of meta-agents and agents enables parallel exploration of the ontologies. The system described here is limited to a construction of the composite search domain ontologies. The use of agents also support many interesting reasoning and searching approaches based on the ontologies. One example is the extension of the initial search by handling user questions and refinements to the search.

2 Related Work

Agents have been used for ontology-based personalized searching and browsing the web [6]. The information navigation is adapted by using the user profile structure as a weighted concept hierarchy. The user profile can be automatically created from a reference ontology, which is built up from the users' browsing. By using a local browsing agent, the user can check the actual web page classified into the concept for the chosen web site. Then, the concepts in the reference ontology are matched to concepts in personal ontology [6].

In our work, however, we use software agents to query the ontologies found in web pages. The agents' collect the information from the ontologies that match the query and the results are sent to and held by a meta-level agent. The mapping we use is to compare the query results to construct composite ontologies, from the different ontology pages. Also the results from the queries are mapped against each other to determine which ontologies are relevant according to the goals of the search.

3 Ontology

Ontology originates in philosophy and has its roots all the way back to Aristotle [17]. The common definition in philosophy is "the definition of what is" (Quine's criterion). One example of a large ontology is the work of Lenat and his students called CYC [13]. There are multiple uses of ontology and, hence, multiple approaches to the contents of an ontology [4]. Moreover, multiple authors create the ontologies on the web and, unfortunately, in Web 2.0 there is no clear guidance to ontology design. Additionally, at present there is no peer review or authority on the ontologies. These all lead to serious concerns when using the existing ontologies.

Given the observations above, the problem we project with multiple ontologies, is consistency. Especially, since the two of the three most common tasks for ontology are communicating and searching, i.e., taxonomy classification and definition of language. In either case, an inconsistent ontology will cause problems. In the case of communication, the consistency between a pair of ontology is a limit to the level of communication that is achievable. In the searching case, one could use the ontologies separately but this can limit the effectiveness. The limitation is especially evident when it is necessary to combine the results to produce more complete answers to queries.

An example illustrating construction to produce complete answers is adding a geographic ontology to company ontology. A geographic ontology can provide additional defined location information that the ontology may not contained in the first place, i.e., the additional information was not defined in the primary ontology. For example, if the search was to locate information about a company, it is important to find the possible extensions of ontology. Suppose that the primary search located the company with an ontology had terms for locations, corporate offices and products to name a few items. Then we might extend the ontology, surrounding the company, with a geographic ontology that could identify distances, or countries, the cities for the company. The countries could lead to identifying governments and, therefore, business regulation. The distances and cites might lead to ontologies that cover transportation systems. Products might lead to information on product quality or recalls. Now, extend this information by noting that initial references to the company will likely occur in a number on pages with separate ontologies and it is easy to envision this process extending into many dimensions.

4 Consistency Detection

Consistency detection is the key issue to find ontologies that have similar content and, thus, can be aligned. The authors Laera *et al.*, [12] describe a solid algorithm for consistency detection in ontologies. Their approach is used for communication between agents with heterogeneous ontologies. The approach does an alignment check between two ontologies and determines the consistency and inconsistencies between these ontologies.

Anyone implementing this approach will need a more complete treatment than is provided by the authors. The algorithm detects consistency between the ontologies but the algorithm requires an $O(N^2)$ number of applications to partition a set of N ontologies into consistent sets. Our work extends the algorithm in two ways; allow it to work on partial ontologies and make it execute in parallel.

The basis of the algorithm is argumentation frameworks (AF) [12], which are applied to ontology mappings. The alignment and the arguments, to justify the alignment, are formalized as follows [12]:

An alignment is represented as a tuple: <e, e', n, R>. In the tuple, *e* and *e'* are entities from two ontologies that are being aligned with each other, e is the original and e' is aligned. *R* is the relationship that holds between the two entities, which can be simple equivalence or a complex, and more complicated, relationship. The tuple, *n*, is the degree of confidence that the relationship, R, holds.

An argument, on the other hand, has the form of a triple <G, m, σ>. In this triple, *m* is an alignment tuple as above, *G* is the ground to justify whether the mapping does hold or not. G will be a logical formula defining the justification for the alignment, e.g., the parent of e and e' are mapped. The term σ has the values {+, -}, which designate the mapping holds when + and the mapping does not hold when -. Both G and σ are used to measure if the mapping holds or not because G can be stated as positive form even though the mapping does not hold. This simplifies comparison of the grounds between arguments as they will be in the same form and logical negation does not need to be computed for the formula.

An argumentation framework is the collection of arguments as defined above. The argumentation framework organizes the arguments to support multiple levels, or kinds, of alignment. The framework is a pair <AR, A>, were *AR* is a set of arguments, with the tuple <G, m, σ>. The set *A* is a set of attacks. An attack is a pair of element from AR, <x, y> such that x attacks y. The attack relationship defines that if x is true then y cannot be. The attack is a pair of argument triples that differ only in the sign [12].

The work in Laera *et al.*, [12] further extends the argumentation framework into a Value Based Argumentation Framework. This is useful, if the definition of multiple levels of alignment is allowed. For the present multi-agent system, this is more capability than required. For our agents, we only utilize argumentation frameworks.

The alignment is defined in terms of the definitions found in Pepijn *et al.*, (1998) [14]:

Semantic: the sets of models of the ontologies compared
Internal Structural: Two entities share internal structure – attributes values, or cardinality
External Structural: The set of relationships with other entities aligned
Terminological: The entity names share lexical features

Alignment of terms can be denied if any of the above relationships fail to hold, that is, taking a very strict view of alignment. For each of these, a formula can support or deny the alignment of terms. Terminological is viewed as an equivalence relationship between the terms. The other terms are based on structural relations that hold between terms, which are aligned. For example, if a relation holds between terms that are assumed to align and, yet a relation in the two ontologies produces terms, which cannot align, then an inconsistency has been detected.

5 Multi-agent Systems

The construction of composite ontologies is implemented in a multi-agent system [18]. The multi-agent system used has two types of agents, meta-agents and ground-level software agents, also referred to as software agents or query agents.

The software agents are used to locate and explore ontologies. The agents are not intelligent in the sense that they learn from the environment but they have some of characteristics that some of the intelligent agents possess. For instance, the software agents in our work can search and collect all the web sites corresponding to the task assigned to the agent without being any user-control while moving around the environment.

In the system, we use several software agents for conducting the queries from the user. The queries are complex structures used to collect the requested information from the web pages. Therefore, the complex query structures are separated into manageable pieces that can be used to explore the ontologies.

The query agents work in a complicated environment. The environment is fully observable, which means that the agent has access to the complete state of the environment [16]. The agent can detect all the ontologies that are relevant to the query. However, the ontologies found by the agents may be redundant and not directly of interests. The reason is that the semantics can be missing from the data extracted from the query. For instance, if green is extracted from the query structure, there is no exact interpretation for the word. The word can have several meanings as green color, green party, green choice for the environment and much more. The query agent only has the data “green” and will collect all the ontologies that contain that word regardless the semantic meaning.

To be able to handle the complex query string, the task is to divide the query structure into smaller and more manageable queries, i.e., atomic episodes, before the agents start working. An atomic episode implies that each episode consists of a set of software agent performing a single action [15]. Usually, the choice of action in each episode depends on the actions taken in previous episode but here the query agents are unaware of the other parts of the query.

5.1 Meta-agents

In the multi-agent system, the meta-agent can be a superior to a software agent [7] or another meta-agent. Meta-agents can be used to improve the decision-making required to perform a task in an environment. These agents can observe the software agents at an object-level in the environment and prescribe scheduled actions.

Moreover, the meta-agents can maintain information on other agents and classify conflicts [7].

The meta-agent is capable of reasoning with software agents and in doing so the meta-agent needs to have a method to evaluate them. The meta-agents have two functions: First they drive the query work and coordinate the work of the query agents. Second, they align the ontologies into consistent sets using the algorithm of argumentation framework to carry out the reasoning. This reasoning is extended to handle multiple ontologies. A meta-agent has a set of ground-level query agents that it communicates with for query operations. As a result, the meta-agent contains a composite ontology that is formed by the alignment of the ontologies of query agents. As inconsistencies are discovered, new meta-agents will be formed as children of the first meta-agent and the query agents are partitioned into two sets, one for each new meta-agent.

The result of querying matching ontologies will form a tree of several different agents. At root, the meta-agent holds the goal of the computation, or search. At the bottom of the tree, the meta-agent communicates with the agents [8]. The ultimate goal of the process is to create a tree with the meta-agents in the interior nodes of the tree and query agents as the leaves. Each meta-agent will contain one ontology that has been constructed by the partial alignment of the ontologies of the query agents, which are found in the tree below the meta-agent. Thus, the meta-agents themselves are members of a tree where their children are either meta-agents or query agents. This also implies that the ontology in a child meta-agent is an extension of the parent's ontology. The ontologies of two sibling meta-agents will differ from each other in at least one alignment, such that the two ontologes are inconsistent with each other. That means the two sibling ontologies cannot be aligned.

Figure 1 shows an example of the agent structure in a running system. The top meta-agent is the initial agent that started the query process meta-agent. The two below were formed when an inconsistency in the ontologies was found. The agents at the bottom of the figure are query agents, used to access the ontologies, which are located in web pages.

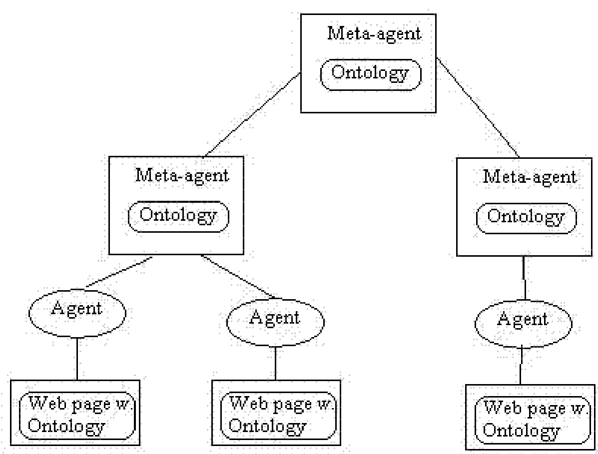


Fig. 1. An example of a tree structure of agents

The composite ontologies in the meta-agents are related in the following ways. The ontology in the parent meta-agent is a subset of the ontologies of the two siblings. The sibling ontologies will differ from each other by a set on inconsistencies that produce a set of differences between the ontologies.

5.2 The Search Process

While the use of the ontology alignment process is to enable effective search, this paper limits the discussion to ontology alignment as a first step in the process. Hence, the search corresponds to the initiation of the alignment process. At the beginning of the search, the user issues a query that becomes a search query, initiating the creation of a single meta-agent. This single agent is the first meta-agent, which will hold the complete query. The first step of the process is to locate an initial set of ontologies that have terms from the search query. A ground-level agent is created to act as the interface to each ontology. The agent is shared by any meta-agent that uses that ontology. The search, driven by the meta-agent, will generate new queries by applying transitive reasoning. As the terms are identified, the newly discovered terms are queried to identify new relationships and new terms. The query agent returns a partial view of the contents of the ontology, since only related terms are exposed by the query in the form of triples of a relation and the two terms joined by the relation. The results are used to generate a composite ontology in the meta-agent. The composite ontology is constructed by the mapping, or aligning, terms between multiple ontologies. As reasoning and querying progresses, a domain ontology for the search is constructed.

The meta-agent sends successive queries to the query agents. The query agents search for matching terms in the ontology and as the query results are returned, the ontology attached to the terms in the ontologies will be exposed. Portions of the ontology that are not considered by the query agents and left unexplored, can be assumed irrelevant and ignored.

As the alignment process continues, contradictions among the ontologies will also be encountered. Contradiction, in this case, means the portions of the ontologies will not align. At this point in the process, the new meta-agents are created and become children of a parent meta-agent and, thereby, inherit the composite ontology of the parent. The inheritance occurs when query agents of the parent are divided into consistent partitions. The partitions are defined by the sets of ontologies that align. Note, in some cases the same ontology will align in multiple partitions. The ontologies, in the newly created meta-agents, will have their ontologies aligned by the latest results from the query agents.

The other process is to add new ontologies to the search. The search for new ontologies can be initiated when a term is weakly connected to the composite ontology. That is when a term is connected by a single relation after the ontology has been searched. This is used to expand the set of source ontologies by searching for new ontologies containing the term.

5.3 The Extended Algorithm

In the system, a composite ontology is constructed by the meta-agents. The meta-agent uses several data structures. The composite ontology is one structure; the other data structures support construction using the algorithm. The meta-agent has an argument framework containing terms of the form: <G= {logical formula}, m=<e, e', n, R>, σ= {+, -}>. The terms in the composite ontology will have a set of labels that is the union of the terms' labels aligned from the ontologies being queried. Likewise, the relationships have a set of labels, and these are union of labels on the relationships that are aligned from the ontologies. The source ontology of each label is recorded to allow subsequent queries to be generated with the correct labels. For efficiency, a hashed index is employed to relate the labels and source ontologies for the entries in the argument framework that contains the label. As a set of query results is returned, the results with the labels are aligned with the current composite ontology. The alignment is incremental, driven by the queries, rather than the full alignment in the original algorithm. Also, the algorithm is able to handle results from several ontologies, instead of just two as in the original algorithm.

The first step in the alignment algorithm is to apply a similarity measure. Our multi-agent system utilizes the similarity measure on both terms and relationship labels. The index system is applied to select the entries in the argument framework that matches on the either terms or relationship labels. The logical formula in the G terms, the grounds, are examined for contradiction. In doing this computation, the σ term is critical. The use of σ allows the formula to be written in positive terms, so we can avoid dealing with logical negation in the formula. If no contradictions are found, the mappings hold, and the grounds are added to the framework of existing mappings and any new mappings are added, as well. At the same time, the composite ontology is extended with the new mappings and the relationship tuples are returned by the query.

The other extension to the algorithm is the handling of contradictions to build a pair of child ontologies in meta-agents children. For example, suppose a result is returned in a query that showed a relation: <e' is_a y>, and we already have <e is_a y> and that e and e' are not mapped. This is a contradiction. The process will leave <e is_a y> in one composite and create a second one w <e' is_a y> in place of <e is_a y> . Now, the second step in this process is that the agent, that produced the result <e' is_a y>, is connected to the second meta-agent and disconnected from the first. Likewise, agents that produced or supported <e is_a y> will be removed from the second meta-agent. The next step requires that the set of labels for terms and relations must be adjusted to remove the unused; the labels that came from the software agents no longer attached to the meta-agent. Also any grounds obtained from the removed agents will also be removed from the meta-agents data structures. This may also affect the parts of the composite ontology to be removed, which is enabled by the design of the indexing.

It is found that the extended argument framework has several advantages over simple alignment algorithms. First of all, the alignment can proceed incrementally. Secondly, since the reasoning for the alignment is preserved in the argument framework, this framework is more efficient than if the composite ontology is completely re-examined for each incremental addition of information. Third, the ability to separate the ontologies into consistent sets extends the ability to build multiple views of the domain and provide extended results.

6 Conclusion

In this paper, we present a solution of using meta-agents on top of software agents in a multi-agent system to solve the search for ontologies and combination of multiple ontologies. The ground-level agents perform the searching of ontologies and the meta-agents the combination of these. The algorithm used is an extension of an existing approach to alignment. The main extension is to merge and partition a large set of ontologies, dynamically and efficiently. Central to the efficient alignment is to align only the parts of an ontology that are relevant to the task. This is a precursor for building systems that reason and search over a large collection of ontologies and, at the same time, avoid having a totally align ontologies in advance.

The extension to searching and reasoning is the next phase of this work. Our next step is to implement reasoning systems over the ontology and, especially, look at how reasoning effectively can leverage multiple composite ontologies.

References

[1] Aleksovski, Z., ten Kate, W., van Harmelen, F.: Exploiting the structure of Background knowledge used in ontology matching. In: Proceedings of ISWC 2006. Workshop on Ontology Matching (OM-2006) (2006), http://www.om2006.ontologymatching.org/Pictures/OM-2006-Schedule.pdf

[2] Bailey, J., Bry, F., Furche, T., Schaffert, S.: Web and Semantic Web Query Languages: A Survey. In: Eisinger, N., Małuszyński, J. (eds.) Reasoning Web. LNCS, vol. 3564, pp. 35–133. Springer, Heidelberg (2005)

[3] Bonino, D., Corno, F., Farinetti, L., Bosca, A.: Ontology Driven Semantic Search. WSEAS Transaction on Information Science and Application 1(6), 1597–1605 (2004)

[4] Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: Ontology of Tasks and Methods (Acrobat), 1998 Banff Knowledge Acquisition Workshop. The first part of the paper appears in What are ontologies and why do we need them? IEEE Intelligent Systems 14(1), 20–26 (1999)

[5] Falconer, S.M., Natalya, F.N., Storey, M.-A.: Towards understanding the needs of cognitive support for ontology mapping. In: International Semantic Web Conference, Busan, Korea (November 2007)

[6] Gauch, S., Chaffee, J., Pretschner, A.: Ontology-Based Personalized Search and Browsing. Web Intelligence and Agent System 1(3-4), 219–234 (2003)

[7] Håkansson, A., Hartung, R.: Using Meta-Agents for Multi-Agents in Networks. In: (ICAI 2007) The 2007 International Conference on Artificial Intelligence, WORLDCOMP 2007, June 25-28, 2007, Las Vegas, USA (2007)

[8] Håkansson, A., Hartung, R.: Calculating optimal decision using Meta-level agents for Multi-Agents in Networks. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 180–188. Springer, Heidelberg (2007)

[9] Hendler, J.: The Semantic Web: A network of understanding (2006), http://www.aswc2006.org/documents/hendler-aswc2006.ppt

[10] Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. The Knowledge Engineering Review 18(1), 1–31 (2003)
[11] OWL (2007-10-15), http://www.w3.org/2004/OWL/
[12] Laera, L., Tamma, V., Euzenat, J., Bench-Capon, T., Payne, T.: Arguing over ontology alignments. In: Proceedings of International Workshop on Ontology Matching (OM-2006) (2006)
[13] Lenat, D.B.: CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM 38(11), 33–38 (1995)
[14] Pepijn, R.S., Visser, D., Jones, M., Bench-Capon, T.J.M., Shave, M.J.R.: Assessing Heterogeneity by Classifying Ontology Mismatches. In: Guarino, N. (ed.) Formal Ontology in Information Systems, pp. 148–162. IOS Press, Amsterdam, The Netherlands (1998)
[15] Ramachandran, R., Movva, S., Graves, S., Tanner, S.: Ontology-Based Semantic Search Tool for Atmospheric Science. In: 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology (2005), http://ams.confex.com/ams/pdfpapers/102272.pdf
[16] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Inc., Englewood Cliffs (1995)
[17] Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations,ch. 2. Brooks/Cole Thomson Learning, Pacific Grove, CA (2000), ISBN 0 534 94965-7
[18] Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons Ltd, Chichester (2002), ISBN 0-471-49691-X

Adaptive Information Provisioning in an Agent-Based Virtual Organization—Ontologies in the System*

Michal Szymczak[1], Grzegorz Frackowiak[1], Maciej Gawinecki[1], Maria Ganzha[1], Marcin Paprzycki[1], Myon-Woong Park[2], Yo-Sub Han[2], and Y.T. Sohn[2]

[1] Systems Research Institute Polish Academy of Science, Warsaw, Poland
{Maciej.Gawinecki,Maria.Ganzha,Marcin.Paprzycki}@ibspan.waw.pl
[2] Korea Institute of Science and Technology, Seoul, Korea
myon@kist.re.kr

Abstract. In this paper we consider utilization of ontologies in an agent-based virtual organization. Specifically, in the system flexible delivery of information is to be based on matching of ontologically demarcated resource profiles, work context(s), and domain specific knowledge. In this paper we introduce basic ontologies and their applications.

1 Introduction

Let us consider a *virtual organization* (*VO*) in which workers access resources to complete their tasks [1,2,3,8]. Access to resources should be adaptive (change with tasks, and evolve as tasks/projects evolve) and personalized (workers require access to different resources depending on their roles in the project and the organization). In our earlier work [13] we have outlined processes involved when a project is introduced into an organization. Later, in [6], we have considered roles played by various entities (humans and agents) identified in [13]. Separately, in [7] we have proposed how e-learning can be introduced into the system in support of adaptability of human resources. In our work we have stated that ontologies will play a crucial role in the system. Thus, the aim of this paper is to present a top level overview of ontologies used in the system and their applications. Observe that in our work ontologies appear at least in the following contexts:

1. *domain specific knowledge*—framework for management of resources(e.g. to specify relations between projects and resources, or humans and resources),
2. *structure and interactions within and between projects*—to manage resources on the basis of needs of project and responsibilities of team members,

* This work was partially sponsored by the KIST–SRI PAS "Agent Technology for Adaptive Information Provisioning" grant.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 271–280, 2008.

3. *resource profiles* (which utilize previous two ontologies)—that specify among others: (a) place of a resource within an organization, (b) employee' *interests*, *needs* and *skills*, and (c) what to do with new/incoming resources,
4. *system data model and data model access infrastructure*—to manage demarcated data, provide rapid system extensibility and persistence.

We proceed as follows. We start with an overview of two existing approaches to ontological representation of virtual organizations. Next, we introduce the proposed system and its actual application—the Duty Trip Support. This allows us to describe the most important features of the proposed ontology of an organization as well as the way it will interact with a travel ontology.

2 Related Work

2.1 Toronto Virtual Enterprise (TOVE)

TOVE project run at the University of Toronto. Its main goal was to establish generic, reusable enterprise data model with the following characteristics [5,15]:

- to provide a shared terminology for the enterprise,
- to define the meaning of each term in a precise and unambiguous manner,
- to implement the semantics in a set of axioms, to enable *TOVE* to automatically deduce the answer to "common sense" questions about the enterprise,
- to define a set of symbols for depicting a term or concept constructed thereof in a graphical context.

According to documents found within the project WWW site, ontology developed by the project included terms such as: resource, requirement, time, state or activity, and was created in Prolog. We thought about relying on the *TOVE* project and utilizing data model constructed there. Especially, since *TOVE* was based on extensive research and considered work of an enterprise from the design and operations perspectives [9]. Unfortunately, inability to find actual ontologies (except of conference papers), a long list of important features to be added found at the project web site, and the last update of that site made on February 18, 2002, resulted in utilization of only the theoretical part of *TOVE*.

2.2 OntoWeb

OntoWeb Network is an initiative aiming at promoting the Semantic Web [10]. Within the project the *OntoWeb* ontology was developed and made available at [11]. Unfortunately the *OntoWeb* ontology has also important drawbacks:

- The *OntoWeb* ontology is created in RDF Schema, which does not have rich enough semantics [14]. Reusing the *OntoWeb* ontology as the system core ontology would lead us to restricting types of reasoning available in the system (e.g. due to the RDF Schema's inability to represent quantifiers).

– The *OntoWeb* ontology does not support resource profiles and information access restrictions, while they are necessary for the proposed system [6,13].

Summarizing, we dropped the idea of reusing the *OntoWeb* ontology due to the limited expressivity of the RDF Schema and lack of necessary concepts. Instead, we followed guidelines and results obtained within both *TOVE* and *OntoWeb* projects and developed an ontology matching our project's needs. To introduce it, let us first describe main features of our system and present an application that is being developed using the proposed approach.

3 Proposed System

In Figure 1 we represent high level view of the proposed system through its use case diagram (see also [6,13]). Let us assume that a task is proposed to the system. To handle it, a *Personal Agent* undertakes a role of a *Project Manager*(*PM*) and orders the *Analysis Manager* (*AM*), to analyze the proposal and create document(s) required to decide whether to accept the job. If the job is accepted the *PM* creates a *Project Schedule* (based on analysis of available and needed resources). We assume that every *PM* has knowledge about some resources in the *VO*. As a result, available resources are reserved (a *Resource Reservation* is created). If the project requires additional resources the *PM* contacts the *Organization Provisioning Manager* (*OPM*) and requests them. The *OPM* has a

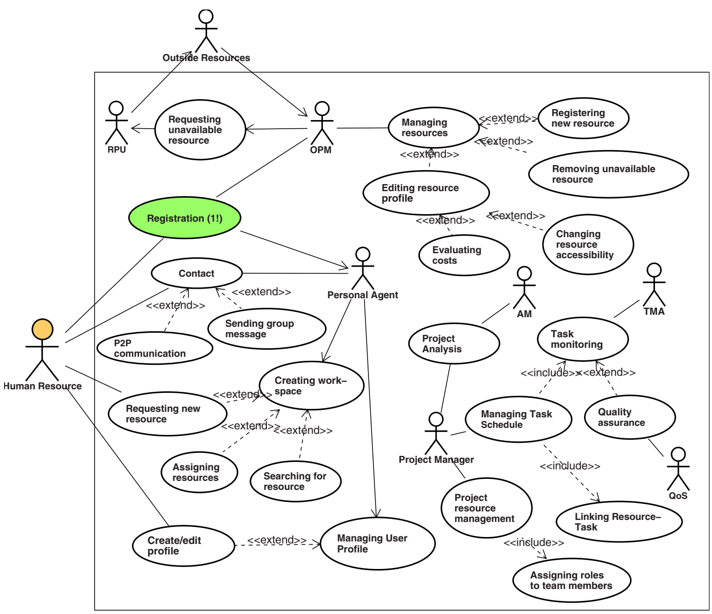


Fig. 1. Use case of the system

knowledge about all resources in the *VO* and can either find them within the organization or ask the *Resource Procurement Unit* (*RPU*) to provide them from the "outside." The *Project Schedule* is used to divide tasks among workers and assign *Task Monitoring Agents* (*TMA*s) to control their progress. Tasks completion is evaluated by a task-specific *Quality of Service* (*QoS*) module.

4 Duty Trip Support Application

Let us now discuss how the proposed system supports personalized information provisioning. The application is the *Duty Trip Support* at a Research Institute in East Asia. Here, workers use the intranet to apply for the *Duty Trip* and to submit trip report. Our aim is to provide them with additional functionalities. First, note that cost of air travel (to most destinations) is much higher—in a relative sense—than costs of a stay extended by a few days. Thus, employee traveling to a given city, may visit also near-by-located institutions (e.g. universities or companies), or persons that her institute has contacts with. Second, a recommender where to stay and eat could be of value (e.g. consider Thai researchers confronted with typical Finnish food). In addition to personalized information delivery, the system is expected to help researchers in all phases of duty trip participation; from the preparation of the initial application until filing the final report. Note that the *Trip Assistant* is actually a role played by the *OPM*, which provides the requested personalized input to the *PA* (see function *Searching for resource* in Figure 1). In Figure 2 we present the activity diagram of duty trip support (see also [6]). In this diagram we can see two moments when the *PA* communicates with the *Trip Assistant* (*OPM*), first when the application for the trip is prepared (and institutions/people to visit are sought), second, when actual details of the trip (e.g. hotels) are to be selected. Let us now utilize the general depiction of the organization (as presented in Figure 1) and the *Duty Trip Support* scenario to describe how we define (1) the *Generic Virtual Organization Ontology*—which delivers basic concepts for our system, and (2) the *Domain Ontology*—Institute of Science and Technology Ontology—which describes data model for the considered scenario and supports necessary scenario-specific reasoning operations.

5 Ontology of the Organization

Before we start let us note that delivering a comprehensive ontology for modeling an organization is beyond the *current* scope of our project. Our main aim is to deliver a framework for adaptive information provisioning. Hence, ontology requirements considered at this stage have been specified to support currently-necessary functions of the system. However, they are flexible enough to support its future extension to support comprehensive organization modeling.

5.1 Generic Top Level Ontology

We have decided to use OWL-DL as the ontology demarcation language, as it guarantees computational completeness and rich semantics[16]. The main idea

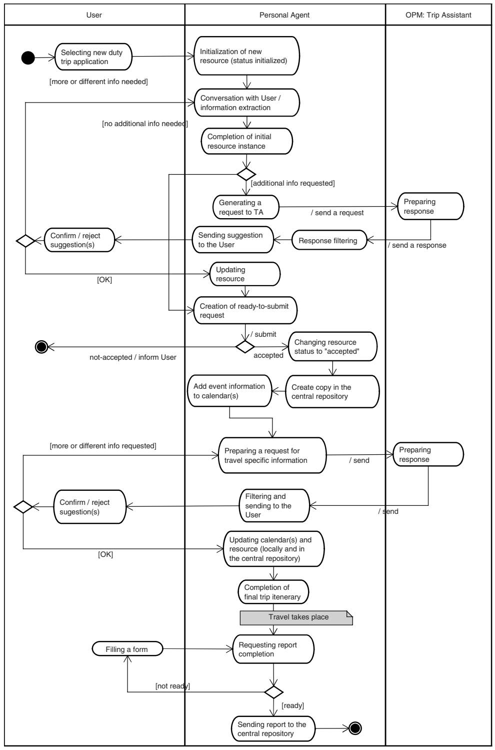


Fig. 2. Activity Diagram of the system

of our approach is to model *every* entity and person within the organization as a *resource*. Furthermore, each resource will have a *profile*. In Figure 3 we depict the generic resource and the generic profile concepts.

A *resource profile* provides detailed information about any resource (human or non-human). It is composed of a resource specific data and "opinions" about other ontology concepts or ontologically demarcated objects [17]. Classes *VOResource* and *VOProfile* are designed to be extended by any organization specific resources and their profiles (assuring that the concept is robust and flexible). For instance, this allows us to specify skills of a worker 1, as well as well as place a *Duty Trip Report* as a specific document within an organization 4 "subsystem."

Note that resource profiles may consist of private or classified information (e.g. personal data) therefore it is necessary to build an infrastructure which can restrict access to the information. This is also important since accessibility to a number of documents depends on the position within an organization (e.g. annual evaluation of a worker should be visible only to that worker and her supervisors). Note also, that each resource may be associated with multiple profiles (see the next section). A *VO Resource Profile Privilege* is a class which describes restrictions "on top" of a profile. It binds a profile with a restriction type which is applied to all resources from a particular *Organization Unit (OU)*—whenever

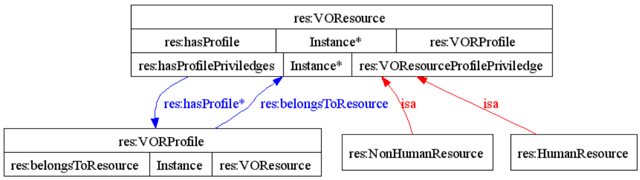


Fig. 3. Generic resource and generic profile concepts

information is requested by, or matched with, resources. The binding of the *OU* and a particular *Profile Privilege Type* is realized by the *Profile Privilege* class. The *Profile Privilege Type* is an enumerable type specifying supported access privileges: *Read*, *Write*, *Read and Write* and *Admin*. Names of the first three are self-explanatory, the fourth type represents an administrative privilege which allows to modify access restrictions of the profile. Here, for instance, the *HR Department* is expected to have *Read and Write* privileges for worker profiles, while the *PA* is going to have *Read* privileges for information provided by the *DTS* (*OPM*) (see Figure 1). The design of the Profile Privilege is depicted in Figure 4.

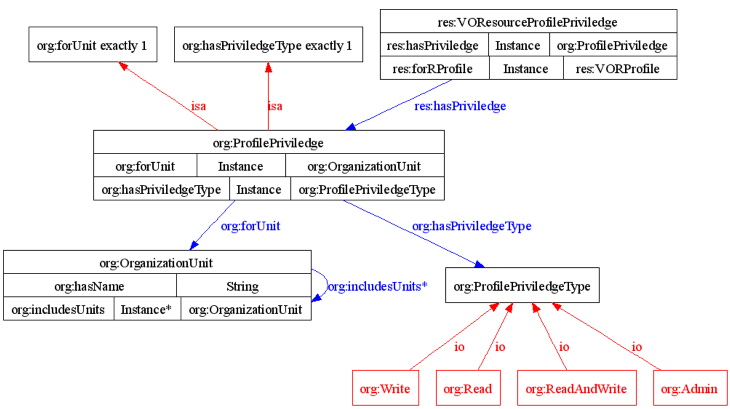


Fig. 4. Profile Privilege design

5.2 *Institute of Science and Technology IST* Ontology

To illustrate how the proposed ontology can be applied, let us discuss briefly its application to selected features of an ontological model of an Institute of Science and Technology. In the architecture of our system, the *Domain Ontology* is an extension of the *Generic Ontology* outlined in Figure 3. Here, human resources are modeled in a way that is specific to the *Institute of Science and Technology*, though similarities with common human resource descriptions can be seen. The *ISTPerson* is a class of all employees of the Institute. While human resources have (multiple) general profiles, the following restrictions on profiles have been assumed:

– at least one profile should represent experience (*IST Experience Profile*).
– at least one should represent personal information (*IST Person Profile*).
– at least one should represent "position" in the organization (*Organization Profile*).

The *IST Experience Profile* allows to describe both educational and professional experience. Additionally, multiple research field can be listed in order to describe employees competences (research fields used here are based on the South Asian RFCD [12]). It is possible to assign level of competence for each research field [18]. *Personal Profile* is a set of data typically stored by the *HR Department*. *Organization Profile* specifies, for instance, a division in which the employee works; it can be also used to establish who is the supervisor of an employee. Additionally, for each human resource, her location and dining opinions (as well as other characteristics) can be assigned [18].

5.3 Duty Trip Ontology

In addition to being able to model human resources, the same general model can be applied to non-human ones. For instance, the *Duty Trip Report* is such a resource; an abstract of a report created by an employee during her duty trip. Its attributes are grouped in the *Duty Trip Report Profile*. These attributes specify trip's period, destination, objectives, expenses, a person who substitutes the departing employee at work, attachments and contacts which were made during the trip. Additionally, user opinions concerning accommodation and dining may be associated with the report. However these opinions cannot be considered as an opinion profile of a duty trip. Instead, this association is a link between a person's opinion and a context which was set by the duty trip. This information will be used in the future for reasoning within advisory functions of the system. The general structure of the *Duty Trip Report Profile* is depicted in Figure 5 and should be viewed together with the example presented in the next section. Note that for "travel objects" we utilize ontologies introduced in [19].

5.4 Example

Let us now present an extended example of utilization of the proposed ontology to model both an organization as well as the specific *Duty Trip*. Due to the limited space we can only point to a few aspects and we hope that the reader will be able to follow the example and find more features. Here, we depict a *Duty Trip* to a conference in Annaville, Canada, where Axel Foley will stay in a Four Seasons Hotel (and visit also Stephanie Brown). We can see also, (1) how the geo-location will be demarcated (following the travel ontology proposed in [19]), and (2) the direct connection between the travel ontology ([19]) and the organization ontology as the city *geo:AnnavilleCity* is an instance of travel ontology element: *SpatialThing* and organization ontology class *City*.

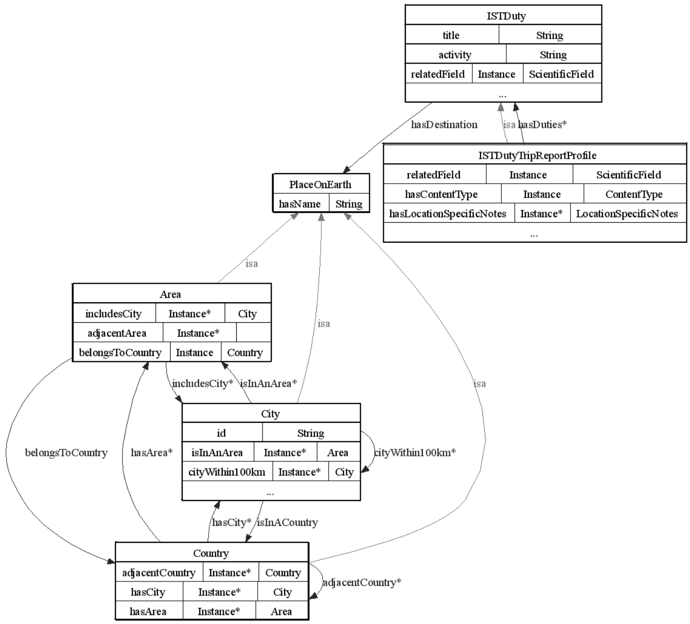


Fig. 5. Duty Trip Report Profile

```
geo:CanadaCountry a onto:Country;
    onto:name ``Canada''^^xsd:string.
geo:QuebecArea a onto:Area;
    onto:name ``Quebec''^^xsd:string;
    onto:isInCountry :CanadaCountry.
geo:AnnavilleCity a onto:City;
    onto:name ``Annaville''^^xsd:string;
    onto:long ``-72.433''^^xsd:float;
    onto:lat ``46.217''^^xsd:float;
    onto:isInCountry :CanadaCountry;
    onto:isInArea :QuebecArea.
geo:AnjouCity a onto:City;
    onto:name ``Anjou''^^xsd:string;
    onto:long ``-73.533''^^xsd:float;
    onto:lat ``45.6''^^xsd:float;
    onto:isInCountry :CanadaCountry;
    onto:isInArea :QuebecArea.
hot:AnnavilleFourSeasons a tss:Hotel;
        onto:locatedAt geo:AnnavilleCity.

:ContactPerson#1 a onto:ContactPerson;
                onto:fullname ``Stephanie Brown''^^xsd:string;
                onto:doesResearch science:Aerodynamics-30501;
                onto:worksIn geo:AnjouCity.

:HROU a onto:OrganizationUnit;
  onto:name ``HR Organization Unit''^^xsd:string.
:GOU a onto:OrganizationUnit;
  onto:name ``Default Organization Unit-for all employees''^^xsd:string.
:Employee#1 a onto:ISTPerson;
            person:fullname ``Axel Foley''^^xsd:string;
            person:gender person:Male;
            onto:hasProfile (<:PersonalProfile#1>);
```

```
              onto:hasProfilePriviledges :ResProfPriv#2.
              onto:belongsToOUs (<:GOU>).
:Employee#2 a onto:ISTPerson;
              person:fullname``John Doe''^^xsd:string;
              person:gender person:Male;
              onto:belongsToOUs (<:GOU> <:HROU>).
:DTR#1 a onto:KISTDutyTripReport;
         onto:hasProfile (<:DTRProfile#1>);
         onto:hasProfilePriviledges :ResProfPriv#1.


:DTRProfile#1 a onto:KISTDutyTripReportProfile;
   onto:destination geo:AnnavilleCity;
   onto:traveler :Employee#1;
   [a onto:Period;
   onto:from ``2008-06-07T00:00:00''^^xsd:dateTime;
   onto:to ``2008-06-19T00:00:00''^^xsd:dateTime.].
   onto:stayedAt hot:AnnavilleFourSeasons
   onto:expense ``4000''^^xsd:integer;
   onto:expenseCurrency ``USD''^^xsd:string;
   onto:visited :StephanieBrown;
   onto:purpose ``Conference''^^xsd:string.
:PersonalProfile#1 a onto:ISTPersonalProfile;
   person:birtdate ``1968-03-07''^^xsd:date;
   onto:currentSalary ``6000''^^xsd:integer;
   onto:currentSalaryCurrency ``USD''^^xsd:string;
   onto:privateAddress ``Alma Drive 0007 Nothingam''^^xsd:string.


:ProfPriv#1 a onto:ProfilePriviledge;
              onto:forUnit :HROU;
              onto:hasPriviledgeType priv:Admin.
:ProfPriv#2 a onto:ProfilePriviledge;
              onto:forUnit :GOU;
              onto:hasPriviledgeType priv:Read.
:ResProfPriv#1 a onto:VOResourceProfilePriviledge;
              onto:forRProfile :DTRProfile#1;
              onto:hasPriviledge (<:ProfPriv#2>).
:ResProfPriv#2 a onto:VOResourceProfilePriviledge;
              onto:forRProfile :DTRProfile#2;
              onto:hasPriviledge (<:ProfPriv#1> <:ProfPriv#2>).
```

Observe that defined privileges allow members of the *HR* unit (e.g. John Doe) to administer selected profiles (e.g. *:PersonalProfile#1*), while members of the *General Organization Unit* are not allowed to access it (by default all access is forbidden). On the other hand the *:DTRProfile#1* can be read by all employees of the *GOU*. In this way we assure control of access rights within the organization.

6 Concluding Remarks

Concluding, we would like to stress the need for ontology verification and testing, especially of the *VO Generic Ontology*. The first step will be creation of the *Duty Trip Support* application and a somewhat related *Grant Announcement Support*. During their implementation and testing completeness of our ontologies will be verified. Furthermore, we will devote our attention to resource matching methodologies, as one crucial functions of the proposed system. Specifically, it is necessary not only to model profiles of resources, but also to establish their "distances." This would allow, for instance, to specify who should receive a grant announcement, or which restaurant should be suggested to a given employee.

References

1. Barnatt, C.: Office Space, Cyberspace and Virtual Organization. Journal of General Management 20(4), 78–92 (1995)
2. Bleeker, S.E.: The Virtual Organization. In: Hickman, G.R. (ed.) Leading Organizations, Sage, Thousand Oaks (1998)
3. Dunbar, R.: Virtual Organizing. In: Warner, M. (ed.) International Encyclopedia of Business and Management, pp. 6709–6717. Thomson Learning, London (2001)
4. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Heidelberg (2001)
5. Fox, M.S., Gruninger, M.: Enterprise Modelling. AI Magazine, 109–121 (Fall, 1998)
6. Ganzha, M., Paprzycki, M., Gawinecki, M., Szymczak, M., Frackowiak, G., Bădică, C., Popescu, E., Park, M.-W.: Adaptive Information Provisioning in an Agent-Based Virtual Organization — Preliminary Considerations. In: Proceedings of the SYNASC Conference, IEEE CS Press, Los Alamitos (in press)
7. Ganzha, M., Paprzycki, M., Popescu, E., Badica, C., Gawinecki, M.: Agent-Based Adaptive Learning Provisioning in a Virtual Organization. In: Wegrzyn-Wolska, K.M., Szczepaniak, P. (eds.) Advances in Intelligent Web Mastering, pp. 112–117. Springer, Berlin (2007)
8. Goldman, S.L., Nagel, R.N., Preiss, K.: Agile Competitors and Virtual Organizations. Van Nostrand Reinhold, New York (1995)
9. Kim, H.M., Fox, M.S., Gruninger, M.: An ontology for quality management - enabling quality problem identification and tracing. BT Technology Journal 17(4), 131–140 (1999)
10. http://ontoweb.org/About/Deliverables/ppOntoweb.pdf
11. http://ontoweb.aifb.uni-karlsruhe.de/Ontology/index.html
12. South Asian RFCD, http://www.kosef.re.kr/english_new/index.html
13. Szymczak, M., Frackowiak, G., Ganzha, M., Gawinecki, M., Paprzycki, M., Park, M.-W.: Resource Management in an Agent-based Virtual Organization—Introducing a Task Into the System. In: Proc. of the IAT Conference, IEEE CS Press, Los Alamitos (in press, 2007)
14. Szymczak, M., Gawinecki, M., Vukmirovic, M., Paprzycki, M.: Ontological Reusability in State-of-the-art Semantic Languages. In: Olszak, C., et al. (eds.) Knowledge Management Systems, pp. 129–142. PTI Press, Katowice (2006)
15. http://www.eil.utoronto.ca/enterprise-modelling/index.html
16. http://www.w3.org/TR/2004/REC-owl-guide-20040210/#OwlVarieties
17. Greer, J., i McCalla, G.: Student modeling: the key to individualized knowledge-based instruction. NATO ASI Series F, 125 (1993)
18. Gawinecki, M.: Modelling User on the Basis of Interactions with a WWW Based System, Master Thesis, AMU (2005)
19. Gawinecki, M., Gordon, M., Nguyen, N.T., Paprzycki, M., Szymczak, M.: RDF Demarcated Resources in an Agent Based Travel Support System. In: Golinski, M., et al. (eds.) Informatics and Effectiveness of Systems, pp. 303–310. PTI Press, Katowice (2005)
20. http://www.w3.org/TR/rdf-schema/

Resource and Remembering Influences on Acquaintance Networks

Chung-Yuan Huang[1,2], Chia-Ying Cheng[3,*], and Chuen-Tsai Sun[3]

[1] Department of Computer Science and Information Engineering
[2] Research Center for Emerging Viral Infections
Chang Gung University
259 Wen-Hwa 1st Road, Taoyuan 333, Taiwan, R.O.C.
[3] Department of Computer Science
National Chiao Tung University
1001 Ta Hsueh Road, Hsinchu 300, Taiwan, R.O.C.
gis92566@cis.nctu.edu.tw

Abstract. To better reflect actual human interactions in social network models, the authors take a bottom-up, agent-based modeling and network-oriented simulation approach to analyzing acquaintance network evolution based on local interaction rules. Resources and remembering are considered in addition to common friends, meeting by chance, and leaving and arriving. Based on these factors, friendships that have been established and built up can be strengthened, weakened, or broken up. Using computer simulations, Results from a series of experiments indicate that (a) network topology statistics, especially average degree of nodes, are irrelevant to parametric distributions because they rely on average values for initial parameters; and (b) resource, remembering, and initial friendship all raise the average number of friends and lower both degree of clustering and separation. These findings indicate a strong need for a bottom-up, agent-based modeling and network-oriented simulation approach to social network research, one that stresses interactive rules and experimental simulations.

Keywords: Resource, remembering, acquaintance networks, social networks, small world, individual interaction.

1 Introduction

The primary feature of any social network is its small world characteristics of low degree of separation (showing how small the world is) and high degree of clustering (explaining why our friend's friend is often our friend also). Watts and Strogatz's research have been the basis for models that apply small world phenomena to measuring social networks. Most of these models (e.g., [9][11]) mix regular and random networks—good for explaining small world features but inadequate for explaining the evolutionary mechanisms and detailed connection structures of social

* Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 281–291, 2008.

networks due to their neglect of local rules that affect interactions and relationships between individuals. Furthermore, researchers have generally focused on the final products of network topologies—that is, they have shown a bias toward top-down approaches that facilitate theoretical analyses [9][10][11]. But actual social network evolution involves bottom-up processes entailing human interactions and daily contacts. Our focus in this study is on the local interaction rules and dynamic growth processes that affect social network formation.

Other complex network models and applications can be referenced when researching social networks. Examples include Barabási and Albert's [2] use of dynamic growth and preferential attachment mechanisms to establish a Barabási-Albert scale-free model where the distribution of node degrees (number of links per node) follows a power-law form, and Albert et al.'s [1] proposal that the Internet possesses scale-free features. Nodes and edges in BA scale-free model are added continuously to the network as time passes (growth); and edges are more likely to connect to nodes of high degree than to one of low degree. To simulate this kind of complex network more realistically, Li and Chen [7] also have introduced the concept of *local-world* connectivity, which exists in many complex physical networks. Their local-world model exhibits transitions between power-law and exponential scaling; BA scale-free model is one of several special cases.

As is the case with other complex networks, social networks exhibit the small world phenomenon but differ in terms of characteristic detail [8]. Growth models on Internet or Web, however, are quite inappropriate as models of the growth of social networks. The reasons are (a) the node degree distribution of many social networks does not appear to follow a pure power-law distribution; (b) the preferential attachment mechanisms is not an important one in social networks; and (c) social networks usually appear high clustering, but growth models of the Internet or Web show weak clustering. Hence, Jin et al. [6] first attempted at modeling the evolution of the structure of social networks. Their model creates a sharply peaked distribution that is in accordance with the observation in a lot of real social networks. They observed many of the features of real social networks from the simulation results of their models, including the formation of closely-knit communities within the social network, and the development of a high degree of network transitivity. At the similar time, Davidsen et al. [4] also proposed a model based on local rules for building *acquaintance* networks; the rules involve such factors as making acquaintances, becoming familiar with other individuals, and losing connections.

Acquaintance networks are a type of social network that represents the processes and results of people meeting people for the first time. Befriending someone in a social network is an important daily activity, one that more often than not involves two individuals and a go-between who is a friend or acquaintance of both. However, a significant number of acquaintances are made by chance, with two or more individuals coming together at a specific time in a specific place because of a shared interest or activity. As part of their acquaintance network model, Davidsen et al. [4] identified two local and interactive rules, the first involving introductions and meeting by chance and the second involving the effect of aging on acquaintance networks. Their model is considered more realistic because it acknowledges a simple observation from everyday experience that often one of our friends introduces us to one of his or her friends. However, the model is still inadequate for explaining changes that occur once acquaintances evolve into friendships—for example, weakening, strengthening, or separation that does not involve the death of one party.

In this paper we will introduce an essential rule that involves individual limited resources and remembering—the latter a memory factor through which individuals remember or forget their friends. This bottom-up, agent-based modeling and network-oriented simulation approach allows for analyses of how individual interaction factors affect the evolution and detailed connection structures of acquaintance networks. For example, researchers can adopt different friend-making resource distributions and probe their effects on entire acquaintance networks and measure correlations between statistical topology features under different circumstances. We believe that when individual limited resources and remembering factor are taken into consideration, the simulation model and many experimental results for exploring the growth of acquaintance networks will be useful to sociologist, epidemiologists, and public health professionals working on epidemic dynamics and related social issues applied in acquaintance networks.

2 The Three-Rule Model

Davidsen et al. [4] propose a two-rule model of acquaintance network evolution, with the first rule addressing how people make new friends—via introductions or meetings-by-chance. The second rule is that friendships are broken when one partner dies. The model formulates a fixed number of N nodes and undirected links between pairs of nodes representing individuals who know each other. We will introduce a *friend remembering* rule that allows for the weakening and strengthening of friendships (Fig. 1). The model repeats the three rules until the acquaintance network in question reaches a statistically stationary state.

- **Rule 1. Friend Making:** Randomly chosen persons introduce two friends to each other. If this is their first meeting, a new link is formed between them. Randomly chosen persons with less than two friends introduce themselves to one other random person. Thus, we use the term "introduce" to describe meetings by chance as well as meetings via a common friend.
- **Rule 2. Leaving and Arriving:** At a probability p, a randomly chosen individual and all associated links are removed from a network and replaced by another person. Accordingly, acquaintances can be viewed as circles of friends whose members can leave for reasons other than death and enter the circle for reasons other than birth.
- **Rule 3. Friend Remembering:** A certain number of friendships are updated, with the number depending on an update proportion b. This proportion and details about updating will be explained in the next two sections.

2.1 Friendship Selection Methods

We considered three selection methods for updating friendships. In the first, *person selection*, an individual chooses $b \times N$ persons before picking a specific friend for each person and updating their friendship. The update does not occur if the chosen person does not have any friends. In this method, b is a proportion factor for deciding how many persons are chosen and N represents the number of persons in the network. In the second method, *pair selection*, the individual chooses $b \times N$ pairs of persons and updates their friendships. Updating is canceled if the paired persons don't know each other—a frequent occurrence, since the network in question is sparse in comparison to a complete graph. In this method, b is a proportion factor for deciding how many pairs are chosen. In the last, *edge selection*, the individual has more direct choice in selecting $b \times M$ friendships for updating. In this method, b is a proportion factor for deciding how many friendships are chosen and M is the number of friendships (or edges) at a specific moment. Without lack of generality, we rejected the first two methods because in both cases, the number of chosen friendships is in proportion to N (number of nodes or persons). Since $N \times (N - 1) / 2$ (the upper boundary of the number of friendships) is directly proportional to M (the number of edges or friendships), we adopted the *edge selection* method for choosing friendships.

2.2 Friendship Update Equation

During friend remembering, the model uses the selection tactic described in the preceding section for choosing a specific number of friendships. If a selected friendship links person u with person v, their friendship is updated using Equation (1), dependent upon individual remembering, resource, and breakup threshold factors:

$$f_{u,v}^{new} = \begin{cases} q \cdot f_{u,v}^{old} + (1-q) \cdot J\left(D\left(\frac{r_u}{k_u}\right), D\left(\frac{r_v}{k_v}\right)\right), & if\ f_{u,v}^{old} \geq \theta, \\ 0, \quad if\ f_{u,v}^{old} < \theta \end{cases} \tag{1}$$

where $f_{u,v}^{new}$ represents the new friendship between u and v, $f_{u,v}^{old}$ the original friendship, q the old friend remembering, θ the breakup threshold, r_u person u's friend-making resources, k_u his or her number of friends, and r_v and k_v person v's resources and friend numbers, respectively. J is a joint function and D a distribution function. For convenience, the friend remembering q, resource r, and breakup threshold θ parameters are normalized between 0 and 1. Simplification without loss of generality is behind our decision to use $D(x) = x$ as the distribution function and $J(a, b) = (a + b) / 2$ as the joint function. The updated equation is written as

$$f_{u,v}^{new} = \begin{cases} q \cdot f_{u,v}^{old} + (1-q) \cdot \left(\frac{r_u}{k_u} + \frac{r_v}{k_v}\right) \cdot \frac{1}{2}, & if\ f_{u,v}^{new} \geq \theta. \\ 0, \quad if\ f_{u,v}^{new} < \theta \end{cases} \tag{2}$$

The equation is divided into two parts by the breakup threshold, θ. The first part consists of the terms q (representing the effect of old friendships) and $(1 - q)$ (representing the effect of limited resources). The newly updated friendship may be weakening or strengthening. It may also theoretically equal zero if the new friendship is below the breakup threshold, as shown in the second part of the equation.

2.3 Expected Effects of Local Rules

Acting locally, the three rules influence several aspects of an acquaintance network: (a) the friend-making rule adds links, thereby increasing the average number of friends; (b) the leaving and arriving and friend-remembering rules both remove links, thereby reducing the average number of friends; (c) increases in average number of friends $<k>$ lead to decreases in the average shortest path length L; and (d) the direction of the clustering coefficient C and average shortest path length L will reverse. As opposed to the large number of factors associated with the friend-remembering rule, the leaving and arriving rule has a single parameter (probability p). The factor q denotes a person's ability to remember friends, thus increasing that person's number of friends. The resource factor r determines an individual's resources for making friends, thereby setting an upper limit. The breakup threshold θ determines the difficulty of cutting off a friendship—a negative influence. The initial friendship factor f_0 is a reflection of how much attention a person is paying when making a new acquaintance—a positive contribution to friend-making. We expect that parameters q, r, and f_0 will exert positive (increasing) influences on $<k>$ and that parameters p and θ will exert negative influences on $<k>$.

2.4 Fitting a Normal Distribution

For sensitivity analyses of skewness and critical parameters affecting distribution, a feasible probability-distribution function (pdf) must be applied. In most situations a normal distribution is considered the best choice, but it does not fit our purposes in this study. Since critical parameters such as initial friendship, old friends remembering, resources, and breakup thresholds have ranges of 0 to 1, we chose a beta distribution—a two-parameter family of continuous probability distributions defined according to the interval [0, 1] with a probability density function of

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{3}$$

where B is the beta function and α and β must be greater than zero. We used a beta distribution subset called *beta14* that satisfies $\alpha + \beta = 14$ and has $\mu = \alpha / (\alpha + \beta)$ as its average. Figure 2 presents pdf curves for *beta14* distributions with averages of 0.1, 0.5, and 0.9. Figure 3 presents pdf curves for comparing *beta14* and normal distributions. Once a simulation reaches a statistically stationary level, then clustering coefficient C, average path length L, average degree of nodes $<k>$, average square degree of nodes $<k^2>$, and node degree distribution statistics can be collected. Degree distributions in our simulations involved some random rippling, especially for smaller populations. However, since large populations consume dramatically greater amounts of simulation time, we applied Bruce's [3] ensemble average as follows:

$$\overline{p}(k) = \frac{1}{M}\sum_{v=1}^{M} p_v(k), \tag{4}$$

where M is the number of curves to be averaged and $p(k)$ a curve that represents.

3 Experiment

A simulation of our model begins with parameter initialization and ends once the acquaintance network reaches a statistically stationary state. Initialized parameters included the number of persons *N*, leaving and arriving probability *p*, updated friendship proportion *b*, old friend remembering *q*, breakup threshold θ, distribution of friend-making resources *r*, and distribution of initial friendship f_0. Statistically stationary states were determined by observing average degree of nodes $<k>$, average square degree of nodes $<k^2>$, clustering coefficient *C*, and average path length *L*. Each of these four statistics eventually converged to values with slight ripples. A statistically stationary state of parameter initialization at $N = 1{,}000$, $p = 0$, $b = 0.001$, $q = 0.9$, $\theta = 0.1$, *r* with a fixed value of 0.5, and a *beta14* f_0 ($\mu = 0.9$) is shown in Figure 4. Blue solid lines indicate the acquaintance network and the green dashed lines in Figures 4c and d indicate the ER random model at the same average degree of nodes as the acquaintance model.

3.1 Effects of Leaving and Arriving

For comparison, we reproduced Davidsen et al.'s [4] simulations using their original parameters of $N = 7{,}000$ and *p* at 0.04, 0.01, and 0.0025. We then changed *N* to 1,000 and tested a broader *p* range. As noted in an earlier section, the leaving and arriving probability *p* is the only parameter in rule 2. In addition to using various degree distribution diagrams, we gathered $<k>$, *C*, and *L* varying in *p* and analyzed their correlations to determine the effects of *p* on the acquaintance network.

The degree distribution $P(k)$ from the two-rule model is shown in Figure 5. All $<k>$, *C*, and *L* values with parameter initializations for various probability *p* values are shown in Figure 6. The solid lines in Figure 6 reflect the application of Davidsen et al.'s two-rule model; the dashed lines reflect the application of the ER model at the same average degree of nodes. Contrasts between the two lines in Figures 6b and 6c indicate that the acquaintance network has the small world characteristic. Figure 6a shows that the number of friends increases as the lifespan of an individual lengthens. According to Figure 6, the clustering coefficient closely follows average degree of nodes not but average path length.

A larger *p* indicates a higher death rate and a lower *p* a longer life span. Thus, parameter *p* acts as an aging factor. Relative to other species, humans require more time to make friends; Davidsen et al. therefore only focused on the $p << 0.1$ regime. To satisfy the needs of integrity theory, we also explored the $p >> 0.1$ regime and found that average degree of nodes $<k>$ decreased for *p* values between 0 and 0.5. The decrease slowed once $p > 0.1$.

Note that in our model, a leaving and arriving probability of 0 means that rule 2 is inactive, and a friendship update proportion of 0 means that rule 3 is inactive. Once rule 3 becomes inactive, our three-rule model becomes the equivalent of Davidsen et al.'s two-rule model. In all of the experiments described in the following sections, *N* was initialized at 1,000 and *b* at 0.001.

3.2 Effects of Breakup Threshold

To determine the effects of the breakup threshold, experiments were performed with parameters initialized at different levels of the friendship-breakup threshold θ. Other initialized parameters were $q = 0.6$ and the constants $r = 0.5$ and $f_0 = 0.5$. The solid lines in Figure 7 represent <*k*>, *C*, and *L* statistics without rule 2 ($p = 0$) and the dashed lines represent the same statistics with rule 2 included ($p = 0.0025$). The data indicate that rule 2—which acts as an aging factor on acquaintances in the network—reduced both average degree of nodes <*k*> and clustering coefficient *C* and increased average path length *L*. According to the data presented in Figure 7, the breakup threshold θ lowers the average degree of nodes <*k*> and raises both the clustering coefficient *C* and average path length *L*. The threshold reflects the ease with which a friendship is broken. As expected, a higher θ results in a smaller number of "average friends" and greater separation between individuals.

3.3 Effects of Resources

To determine the effects of resources and memory factors on acquaintance networks, we ran a series of experiments using parameters initialized with different friend-making resource *r* and friend-remembering *q* values. Initialized parameters also included $p = 0$, $\theta = 0.1$, and a fixed f_0 value of 1. According to our results, a larger *r* raised the average degree of nodes <*k*> but lowered the clustering coefficient *C* and average path length *L* (Fig. 8). While it is not obvious that statistical characteristics are influenced by different resource distributions, they are clearly influenced by different resource averages. In other words, <*k*>, *C*, and *L* are affected by different resource averages but not by different resource distributions. The Figure 8 data also show that an increase in *q* raised <*k*> and lowered both *C* and *L*.

3.4 Effects of Initial Friendship

Experiments were run using parameters initialized at different initial-friendship f_0 and friend-remembering *q* values for the purpose of determining the effects of those factors on acquaintance networks. Other initialized parameters were $p = 0$, $\theta = 0.1$, and a fixed *r* value of 0.5. Our results show that a larger f_0 raised the average degree of nodes <*k*> but lowered both the clustering coefficient *C* and average path length *L* (Fig. 9). That different distributions of initial friendship influenced the statistical characteristics was not obvious, but different averages of initial friendship clearly did. In other words, <*k*>, *C*, and *L* were affected by different initial friendship averages but not by different initial friendship distributions. The Figure 9 data also show that the friend remembering *q* factor raised <*k*> and lowered both *C* and *L*.

We analyzed the effects of different parameters on our proposed model by relationally cross-classifying all experiments; our results are shown in Tables 3 and 4. The plus/minus signs in Table 3 denote positive/negative relations between parameters and statistics. In Table 4 the plus or minus signs denote the strength and direction of correlations. As Table 3 indicates, in addition to the effects of rule 1, q, r, and f_0 had positive correlations with average degree of nodes <*k*> and p and θ had negative correlations with <*k*>. Furthermore, each average degree of nodes had a negative relationship with its corresponding average path length. All of the rule 3 parameters affected the clustering coefficient C and average length L in a positive manner, while rules 1 and 2 affected C and L negatively. Note that friendships are initialized in rule 1 and updated in rule 3.

3.5 Distribution of Co-Directors

The three main properties of social networks are (a) the small-world phenomenon, (b) the high-clustering characteristic, and (c) skewed node degree distribution. In this chapter, we focus on the third property. Networks of board and director interlocks reveal a remarkable node degree distribution that is different by far from either scale-free or ER random networks [5]. For example, the nearly 8000 directors on the board of Fortune 1000 companies in 1999 are connected, and the corresponding degree distribution has a strongly peak and a fast approximately exponential decay in the tail, much faster than a power-law distribution but slower than a Poisson or normal distribution. Figure 10 shows the node degree distribution comparison between one of our acquaintance networks after at a statistically stationary state (blue solid curve) and co-directors for Davis' boards-of-directors data (green dashed curve). Selected acquaintance network parameters were initialized at N = 1,000, p = 0, b = 0.001, q = 0.4, θ = 0.1, r = 0.5 (fixed) and f_0 = 0.5 (fixed). Davis' data is about the nearly 8000 directors on the board of Fortune 1000 companies in 1999 [7]. Both curves exhibit similar peaks and then a long tail that doesn't appear to decay smoothly.

Table 1. Effective directions of the parameters on <*k*>, *C*, *L*

Statistics	Rule1	Rule2	Rule 3			
		P	*q*	θ	*r*	f_0
<*k*>	+*	−	+	−	+	+
C	+*	−	−	+	−	−
L	−*	+	−	+	−	−

Table 2. Correlations between <*k*>, *C*, *L* from experiments

Experiments	Variational Parameters	*C*-<*k*>	*L*-<*k*>	*C*-*L*
4.1	*P*	+++	—	——
4.2	θ, *p*	—	——	+++
4.3	*q*, *r*	——	—	+++
4.4	*q*, f_0	——	—	+++

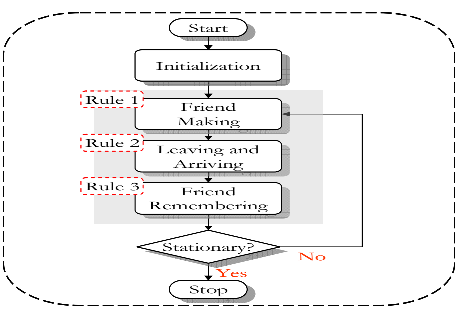


Fig. 1. Three-rule model flow diagram

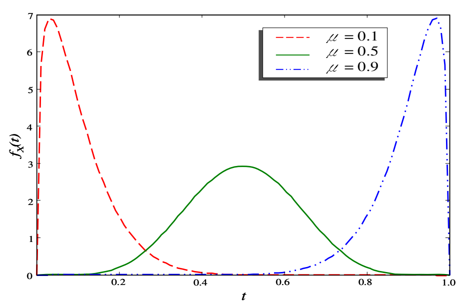


Fig. 2. *Beta 14* pdf curves at different average

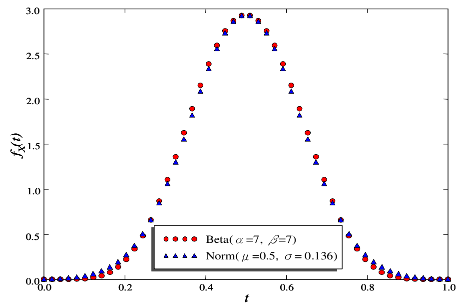


Fig. 3. Comparison of beta and normal distributions

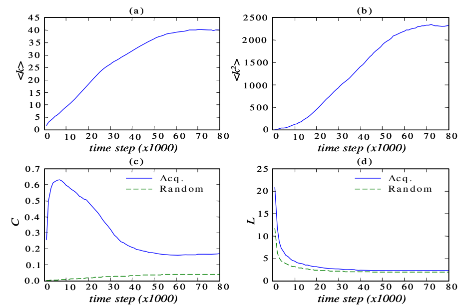


Fig. 4. Example of a statistically stationary state using the proposed model

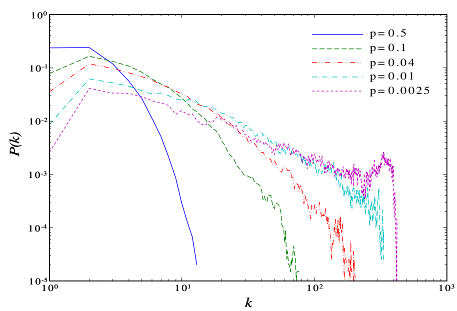


Fig. 5. Two-rule model degree distribution

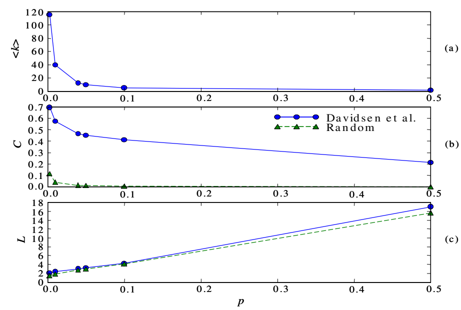


Fig. 6. <*k*>, *C* and *L* varying in leaving and arriving probability *p*

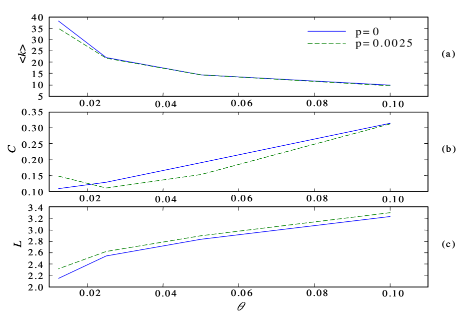


Fig. 7. <*k*>, *C* and *L* varying in breakup threshold θ with different leaving and arriving probability *p*

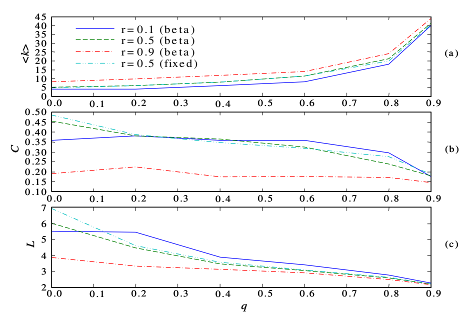


Fig. 8. <*k*>, *C* and *L* varying in friend-remembering *q* value with different distributions of friend-making resource *r*

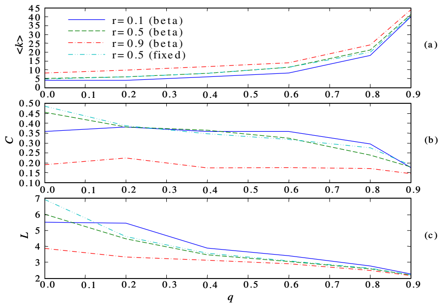


Fig. 9. <k>, C and L varying in friend-remembering q value with different distributions of initial friendship f_0

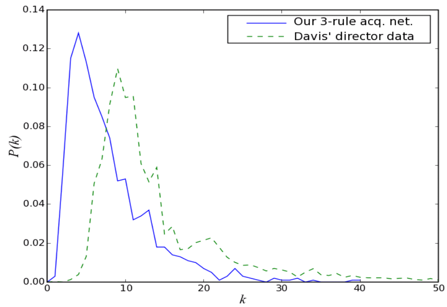


Fig. 10. Node degree distribution comparison between one instance of our acquaintance networks and Davis' board of directors data

4 Conclusion

Experimental simulations are a necessary aspect of social network research, not only because of the expenses and other difficulties involved with fieldwork, but also because widely used sampling approaches cannot capture real social network distributions, since distributions for higher sampling rates differ from those for lower sampling rates. Taking a bottom-up, agent-based modeling and network-oriented simulation approach to modeling is a reflection of the evolution mechanism of real social networks. Building on insights from previous studies, we applied local and interactive rules to acquaintance network evolution. This approach produced new findings that can be used to explore human activity in specific social networks—for example, rumor propagation and disease outbreaks.

Acknowledgement

This work was supported in part by National Science Council, Taiwan, Republic of China under grant NSC 96-2221-E-182-041.

References

1. Davidsen, J., Ebel, H., Bornholdt, S.: Emergence of a small world from local interactions: Modeling acquaintance networks. Physical Review Letter 88(12), 128701 (2002)
2. Jin, E.M., Girvan, M., Newman, M.E.J.: The structure of growing social networks. Physical Review Letters E 64, 046132 (2001)
3. Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. Physical Review E 68(3), 036122 (2003)
4. Davis, G., Yoo, M., Baker, W.: The small world of the American corporate elite, 1982-2001. Strategic Organization 1(3), 301–326 (2003)
5. Li, X., Chen, G.: A local world evolving network model. Physical A 328, 274–286 (2003)

6. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. Nature 393(6684), 100–442 (1998)
7. Zanette, D.H.: Dynamics of rumor propagation on small-world networks. Physical Review E 65(4), 41908 (2002)
8. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
9. Watts, D.J.: Networks, dynamics, and the small-world phenomenon. American Journal of Sociology 105(2), 493–527 (1999)
10. Newman, M.E.J., Watts, D.J.: Renormalization group analysis of the small-world network model. Physics Letters A 263(4-6), 346–347 (1999)
11. Watts, D.J., Dodds, P.S., Newman, M.E.J.: Identity and search in social networks. Science 296(5571), 1302–1305 (2002)

A Formalization for Distributed Cooperative Sensor Resource Allocation

Toshihiro Matsui and Hiroshi Matsuo

Nagoya Institute of Technology
{matsui.t, matsuo}@nitech.ac.jp

Abstract. Distributed sensor resource allocation problem is an important research area of multi agent systems. In this paper we propose a model of distributed resource allocation problem for distributed sensor networks. Several models based on constraint network and another model based on concept of agency, are compared. Then, constraint network formalization which are similar to resource allocation problem of agency model, is shown.

1 Introduction

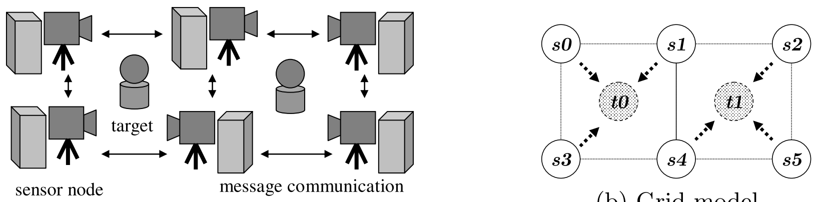
Distributed sensor network is studied as an application domain of multi-agent system. In this paper, we focus on a distributed observation system shown in Figure 1(a). In the observation system, sensor nodes have its sensor and processor that are connected with message communication link. Intrusion detection, target tracking and scheduled observation are included in tasks of the system. The observation tasks are performed as distributed cooperative processing using sensor node's processors and communication link.

In the observation systems, resource allocation, which allocates sensors to targets, is an important problem. In generally, the resource allocation problem contains optimization problems. Therefore, consideration of applying distributed optimization algorithm is useful to understand the problems and to design the cooperative protocols. Distributed constraint satisfaction/optimization problems, which is a fundamental formalism for multi-agent cooperation, have been studied [1], [2], [3], [4], [5], [6] . Formalizations which represents distributed sensor networks as distributed constraint networks have been proposed [7],[8].

On the other hand, a distributed cooperative observation system using agency model has been developed [9],[10]. The agency model has been applied to practical environment consists on autonomous camera sensor nodes, which own pan-tilt-zoom controlled cameras, computers and local area network. In this model, total processing, including camera input and pan-tilt-zoom output, are integrated as a hierarchical distributed processing.

In this paper, a DCOP formalization is applied to a cooperative sensor resource allocation problem. Formalizations using DCOP and agency model are compared. Furthermore a cooperative formalization is proposed intend to integrate DCOP approach into agency model.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 292–301, 2008.



(a) Sensor nodes and targets (b) Grid model

Fig. 1. A model of sensor network

2 Background: Modeling for Sensor Allocation

2.1 Sensor Allocation Problem

In this paper, we focused on a resource allocation problem such that observation resources of sensor nodes are allocated to targets. For the sake of simplicity, the sensor resource allocation problem is represented using grid model shown in Figure 1. The grid models are used in related works [8]. In Figure 1(b), s_i and t_j represent a sensor node and a target respectively.

The goal of the problem is an optimal allocation of sensor resources for all targets. The allocation have to satisfy conditions for as follows.

- Sensors have limited observation area. For example, each sensor observes targets which is inside of neighbor grids. Other targets are invisible.
- Sensors have limited observation resources. For example, each sensor simultaneously observes one target in observation area.
- There are requirements of sensor resources for observation of targets. For example, three sensors are required for a target to estimate coordinate of the target in enough accuracy.

In Figure 1(b), an optimal solution is shown as arrows.

Sensor nodes solve the problem using distributed cooperative processing. In this paper, two approaches for the cooperation are focused. In 2.2, 2.3 and 2.4, formalization using distributed constraint satisfaction/optimization problem [7],[8] is shown. In 2.5 another model using concept of agency [9],[10] is shown.

2.2 DCOP

Distributed constraint optimization problem (DCOP) [1], [2], [5], [6], is a fundamental formalism for multi-agent cooperation. In the DCOP, multi-agent systems are represented as variables and constraints. Definition of DCOP is as follows.

- A DCOP is defined by a set A of agents, a set X of variables, a set C of binary constraint and a set F of binary functions.
- Agent i has its own variable x_i. x_i takes a value from discrete finite domain D_i. The value of x_i is controlled by agent i.

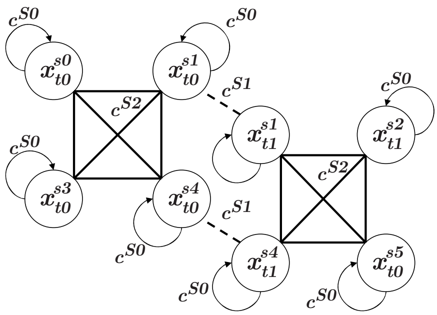


Fig. 2. STAV

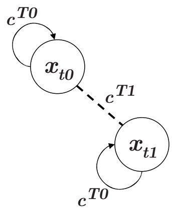


Fig. 3. TAV

- Relation of an assignment $\{(x_i, d_i), (x_j, d_j)\}$ is defined by a binary constraint $c_{i,j}$.
- Cost for $c_{i,j}$ is defined by a binary function $f_{i,j}(d_i, d_j) : D_i \times D_j \rightarrow \mathbb{N}$.
- The goal is to find global optimal solution $\mathcal{A}$ such that it minimizes the global cost function: $\sum_{f_{i,j} \in F,\ \{(x_i,d_i),(x_j,d_j)\} \subseteq \mathcal{A}} f_{i,j}(d_i, d_j)$.

Agents cooperatively search the optimal solution using distributed constraint optimization algorithms. In recent years, a number of algorithms are proposed for DCOP [1] [2], [3], [4], [5], [6] . These algorithms are categorized into exact solution methods and inexact solution methods. In sensor network, inexact methods will be useful because of its scalability. However, the methods frequently obtains local optimal solution.

2.3 STAV (Sensor Target As Variable)

In this section, a DCOP formalism, in which variables are defined for sensors and targets, is shown. The example of sensor network shown in Figure 1 is formalized as a STAV shown in Figure 2.

Variable $x^{s_i}_{t_j}$ is defined for sensor s_i and target t_j which is inside of observation area of sensor s_i. The variable $x^{s_i}_{t_j}$ takes a value which represents an allocation of sensors for target t_j. For example, if target t_j is inside of observation area of sensors $s_0, \cdots, s_n$, The variable $x^{s_i}_{t_j}$ takes a value which represents $\{\phi, \{s_0\}, \cdots, \{s_n\}, \{s_0, s_1\}, \cdots, \{s_0, \cdots, s_n\}\}$.

Constraints which are defined for each variables are as follows.

- $c^{S0}(x^{s_i}_{t_j})$: A unary constraint which represents requirement of sensor resource for target t_j.
- $c^{S1}(x^{s_i}_{t_j}, x^{s_i}_{t_{j'}})$: A binary constraint which represents limitation of sensor resource. It is disallowed that a sensor s_i is multiply allocated to different targets $t_j and t_{j'}$.
- $c^{S2}(x^{s_i}_{t_j}, x^{s_{i'}}_{t_j})$: A binary constraint which represents consistency of decision of sensors. Two sensor allocations, which are decided in s_i and $s_{i'}$ for a target t_j, must be equal.

In this example, it is disallowed that one sensor is multiply allocated to different targets. Therefore, the constraint c^{S1} is formalized as a binary constraint. If one sensor is multiply allocated to targets, the constraint is formalized as a n-ary constraint.

Moreover, constraint c^{S2} is defined for each two variables which are related to t_j. As another formalization, these constraints are integrated into a n-ary constraint.

In the STAV, variables are distributed into sensor nodes, and agreement of sensor nodes is explicitly formalized. However, this detailed formalization increases number of variables and constraints.

2.4 TAV

In this section, a DCOP formalism, in which variables are defined for targets, is shown. The example of sensor network shown in Figure 1 is formalized as a TAV shown in Figure 3.

Variable x_{t_j} is defined for t_j. The variable x_{t_j} takes a value which represents an allocation of sensors for target t_j. The domain of the variable are same as STAV.

Constraints which are defined for each variables are as follows.

- $c^{T0}(x_{t_j})$: A unary constraint which represents requirement of sensor resource for target t_j.
- $c^{T1}(x_{t_j}, x_{t_{j'}})$: A binary constraint which represent consistency of sensor allocation for $t_j, t_{j'}$. This represents each sensor is allocated to at most one target.

In the TAV, agreement of sensor nodes is not considered. STAV is translated to TAV as follows.

1. Constraints of c^{S0} for same target t_j are integrated into a constraint c^{T0}.
2. Constraints of c^{S2} for same target t_j are removed.
3. Constraints of c^{S1} for same pair of targets (t_i, t_j) are integrated into a constraint c^{T1}.

In the TAV, number of variables and constraints are less than one in SAV. However, in practical problem, variables can not be processed in targets. In this paper, we assume that available resources for computation is only contained in sensor nodes. Therefore, solving another problem, that distributes the variables and constraints on sensor nodes, is necessary.

2.5 Agency Based Cooperative Model

As another approach different form DCOP formalization, distributed cooperative observation system using agency model has been proposed [9],[10]. The agency model has been applied to practical environment consists on autonomous camera

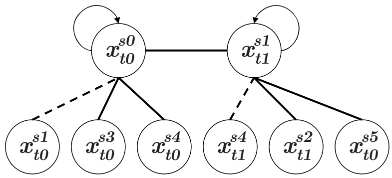


Fig. 4. A constraint network similar to agency model

sensor nodes, which own pan-tilt-zoom controlled cameras, computers and local area network. In this model, total processing, including camera input and pan-tilt-zoom output, are integrated as a hierarchical distributed processing. The outline of the system is as follows.

- The observation system is consist on sensor nodes which are called AVA. The hardware of each AVA consists on pan-tilt-zoom controlled camera, computer and local area network interface.
- When each AVA discovers targets, AVAs form a group (agency) for each target.
- One of AVAs in a agency performs as a manager of the agency. Other AVAs follow decision of the manager. The managers negotiate to share sensor resources (AVAs) among agencies.

In this paper, our discussion is focused on sensor resource allocation of the agency model. Important points as the sensor resource allocation problem is as follows.

- Similarly to TAV, information is gathered for each target.
- The gathered information is allocated to a manager node.
- Other nodes which are related to same target follow its manager node's decision.

From these points of view, the agency model is considered as an integrated model which partially contains STAV and TAV. A constraint based formalization, which is similar to the agency model, is shown in Figure 4. In this formalization, variables are defined same as STAV. However, for target t_0, a variable which is contained in sensor s_0 is prioritized. Similarly, for target t_1, a variable which is contained in sensor s_1 is prioritized. Other values of variables follow the variable which is prioritized for the target.

In some cases, variables which, are related to same sensor node, are follow different variables. For example, in figure 4, x^{s4}_{t0}, x^{s4}_{t1} are follow different variables. Similarly, $x^{s1}_{t0} and x^{s1}_{t1}$ are follow different variables. In this case x^{s1}_{t1} follows itself.

The agency model has been successfully demonstrated in practical environment. On the other hand, sensor resource allocation problem and its solver is implicitly contained. The solver is basically considered as hill-climb based method.

3 Cooperative Model Using Constraint Network

In TAV, it is necessary that variables are allocated to sensor nodes in processing. On the other hand, in the agency model, this variable allocation and sensor resource allocation using the variables are integrated. For practical model, this integration is useful. However, representation of sensor allocation problem and its solver is implicitly contained. Our purpose is to represent these problems as constraint formalization. In this section, a formalization which hierarchically integrates two problems as follows.

– Variables of STAV are prioritized for each target. The prioritization is considered as an allocation of computation resources.
– Then the most prioritized variables are used to solve for sensor resource allocation which is based on TAV.

3.1 Allocation of Computation Resources

Allocation of computation resources is as follows. STAV is translated to TAV. Then variables of TAV are allocated to sensor nodes (computation resources). In actually, this is done by solving a leader election problem which satisfy constraints as follows.

– At least one STAV variable for each target must be the manager of the target.
– At most one STAV variable for each sensor must be the manager. This constraint represents limitation of computation resources. We assume that only one target is processed in one sensor node. This limitation may generalized to handle multiple targets.

Allocation of computation resources is a basis of allocation of sensor resources. Unless allocation of computation resources is not solved, allocation of sensor resources is not solved. Therefore these constraints must be immediately satisfied to solve problem. In this paper, we assume that the allocation of computation resources is always solved.

3.2 Allocation of Sensor Resources

After computation resources are allocated, allocation of sensor resources is solved as TAV using most prioritized variables. Other variables must take same value as the corresponding most prioritized variable.

3.3 TAV+SAV (Sensor As Variable) — Gathering of Decision Making

Constraint c^{T1} for TAV variables is related to multiple variables. Therefore, communication between variables is necessary to evaluate the constraint. For example, when variable x_{t_i} takes a value from its domain $\{\phi, \{s_0\}, \{s_1\}, \{s_0, s_1\}\}$, verifying exclusive assignment is necessary for each other variable which is related to sensor s_0 or s_1.

To reduce this communication, domains of variables are modified. The modified domain consists on tuple of assignments of sensors. The assignment of each sensor represents a target which is allocated to the sensor. For example, domain of x_{t_i} consists on tuple of assignment of s_0 and s_1. The assignment for $s0$ represents a target in $\{\phi, t_k, \cdots, t_{k'}\}$ which are inside observation area of $s0$. Here ϕ represents that no target is allocated. In this translation, resource constraint of sensor is considered.

As a result of the translation, information about targets and sensors is gathered into sensor nodes which own most prioritized variables. This is considered as gathering of decision making into agency managers.

When a sensor is related to multiple most prioritized variables (agency managers). Assignment for the sensor is decided by one agency manager using tie-break. A cooperative method using the similar concept of multiple agency is shown in [10].

3.4 Formalization

A cooperative model shown in above subsections is formalized as follows. Each agent knows information as follows.

- T_s : A set of target which is inside observation area of sensor s.
- N_s : A set of neighbor sensors of sensor s. Information including T_s and resource constraint of sensor s is shared with other sensors in N_s.

According to these information, each sensor generates variables and constraints. Variables are defined as follows.

- $x^T_{s,t}$: A variables which represents manager or member of agency. $x^T_{s,t}$ takes a value form its domain $D^T_{s,t} = \{0, 1\}$. $x^T_{s,t} = 1$ represents that sensor node s is manager of agency for target t. Otherwise s is member of the agency.
- x^S_s : A variable which represents allocation of sensor s for targets. X^S_s takes a value from its domain D^S_s. D^S_s represents a subset of tuple of T_s or nothing to allocate. In our example problem definition, sensor s is allocated to only one target in T_s. Therefore D^S_s=$T_s \cup \{\phi\}$.
  Each sensor node s has variables of related sensor nodes in N_s. The copy of other neighbor node's variables are necessary when s performs as a manager.

Constraints are defined as follows.

- $c^{A0}_{s,t}$: A constraint which prohibits confliction of multiple agency manager between two sensor nodes. The constraint is defined as follows. Number of variable such that $x^T_{s,t} = 1$ for all $s \in N_s$, must be 1. Only one sensor node performs as an agency manager for a target.
- $c^{A1}_{s,t}$: A constraint which prohibits confliction of multiple agency manager for own targets. The constraint is defined as $\neg(x^T_{s,t} = 1 \land x^T_{s,t} = 1)$ for all $t \neq t'$. In our example problem definition, a sensor node performs as a agency manager for only one target.

– $c^{A2}_{s,s',t}$: A constraint which represents membership of agency. If $x^T_{s',t} = 1$ then $x^S_s = x^S_{s'}$ must be satisfied. If $x^T_{s',t} = 1$ in multiple sensors s' then one of those is prioritized using tie-break of sensor identifier.
– $c^{A3}_{s,t}$: A constraint which represents requirement of sensor resources for target t. This constraint is relaxed using cost function if it is necessary. The cost is evaluated using number of variables such that $x^S_s = t$. Cost value of constraints is defined as follows.

$$f^{A3}_{s,t}(n+1) \ll f^{A3}_{s,t}(n), \quad \text{where } n \text{ denotes number of variables such that } x^S_s = t. \tag{1}$$

Additionally, we also use tie-break using priority of variables identifier to avoid confliction of decisions between nodes.

3.5 Problem Solving

In 3.4, a cooperative model, which is similar to agency model, is formalized as DCOP. However, solving this problem ignoring hierarchy structure contained in that, is obviously inefficient. Therefore search processing is pruned using the hierarchy.

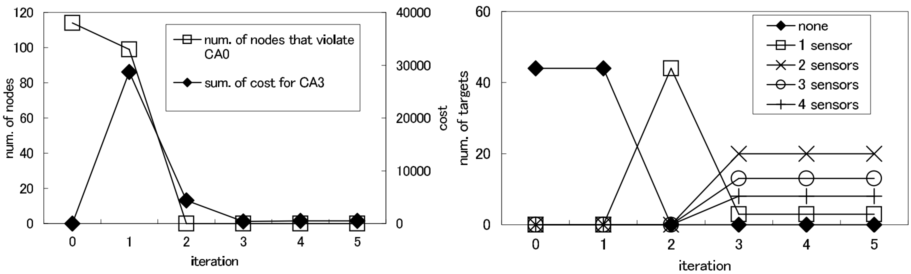
Priority of constraints
A priority relation between constraints is defined from hierarchy structure of problem. The priority relation is shown as follows: $c^{A0} \succ c^{A1} \succ c^{A2} \succ c^{A3}$. Here $c \succ c'$ denotes that constraint c is prior to c'. Constraints are satisfied according to the priority.

c^{A0} and c^{A1} are the constraints for leader (manager) election problem. These constraints are related to set of variables $x^T_{s,t}$. Therefore, in first step of search processing, partial solution for the set of variables. Then other constraints related to x^S_s are satisfied. This prioritization is done using a variable ordering such that $x^T_{s,t}$ is prior to $x^S_{s,t}$.

Gathering decision
In member (non-manager) sensor node, search processing to satisfy c^{A2} and c^{A3} is redundant. Therefore, the member nodes only receive the assignment of variables. Search processing of the assignment is not performed in the member nodes.

4 Experiment

As first experiment, proposed model is applied to a grid sensor network problem. The problem is generated using parameters w, h, c. w and h determines width and height of grids. c determines a degree of constraint network. Each target is added to grids such that, number of targets inside 8 neighbor of the grid, is more than 1 and less than c. The cost of c^{A3} is set to $\{0, 1, 10, 100, 1000\}$ for number $\{4, 3, 2, 1, 0\}$ of allocated sensor nodes respectively. In the experiment



(a) Constraint and cost (b) Allocation

Fig. 5. An example of execution

Table 1. Error rates to optimal cost

w, h	c	num. of targets	error
3	3	7.6	2.74
3	4	8.5	0.94
5	1	11.7	1.63
5	2	15.5	2.57
10	1	43.1	1.42

basic hill-climb based method using variable ordering is applied. The outline of the processing is as follows. Each node sends its assignments to neighbor nodes. Messages are exchanged by simulator. In this experiment, processing of nodes is synchronized to global message cycles. When each node receives neighbor node's assignment, the node modifies its assignment. These processing is iteratively repeated until assignment is converged. An example of execution is shown in Figure 5. In fast step of the execution, constraints c^{A0} and c^{A1} are solved. Then, sensor allocation, represented as c^{A2} and c^{A3}, are solved. In these problems, assignments are converged within 5 iterations. However, most assignments are converged into local optima. Error rates to optimal cost are shown in Table 1. It is considered that the sensor allocation problem is rather complicated problem. Therefore applying DCOP algorithms is necessary to improve solution.

5 Conclusion

In this paper, DCOP formalization is applied to a cooperative sensor resource allocation problem. Formalizations using DCOP and agency model are compared. And a cooperative formalization is proposed intend to integrate DCOP approach into agency model.

Integration of distributed algorithms including construction of constraint network, applying other solver for DCOPs and extension for practical observation problems will be included in our future work.

References

1. Liu, J., Sycara, K.: Exploiting problem structure for distributed constraint optimization. In: Proc. 1st Int. Conf. on Multiagent Systems, pp. 246–253 (1995)
2. Modi, P.J., Shen, W., Tambe, M., Yokoo, M.: Adopt: asynchronous distributed constraint optimization with quality guarantees. Artificial Intelligence 161(1–2), 149–180 (2005)
3. Maheswaran, R.T., Tambe, M., Bowring, E., Pearce, J.P., Varakantham, P.: Taking DCOP to the Real World: Efficient Complete Solutions for Distributed Multi-Event Scheduling. In: Proc. 3rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems, pp. 310–317 (2004)
4. Ali, S.M., Koenig, S., Tambe, M.: Preprocessing techniques for accelerating the DCOP algorithm ADOPT. In: Proc. 4rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems, pp. 1041–1048 (2005)
5. Petcu, A., Faltings, B.: A Scalable Method for Multiagent Constraint Optimization. In: Proc. 9th Int. Joint Conf. on Artificial Intelligence, pp. 266–271 (2005)
6. Zhang, W., Wang, G., Wittenburg, L.: Distributed stochastic search and distributed breakout: Properties, comparison and applications to constraint optimization problems in sensor networks. Artificial Intelligence 161(1–2), 55–87 (2005)
7. Béjar, R., Krishnamachari, B., Gomes, C., Selman, B.: Distributed Constraint Satisfaction in a Wireless Sensor Tracking System. In: Proc. Workshop on Distributed Constraint of IJCAI 2001 (2001)
8. Béjar, R., Domshlak, C., Fernández, C., Gomes, C., Krishnamachari, B., Selman, B., Valls, M.: Sensor networks and distributed CSP: communication, computation and complexity. Artificial Intelligence 161, 117–147 (2005)
9. Ukita, N., Matsuyama, T.: Real-time Cooperative Multi-target Tracking by Communicating Active Vision Agents. Computer Vision and Image Understanding 97(2), 137–179 (2005)
10. Ukita, N.: Real-time Cooperative Multi-target Tracking by Dense Communication among Active Vision Agents. Web Intelligence and Agent Systems: An International Journal 5(1), 15–30 (2007)

Ant Colony Programming with the Candidate List

Mariusz Boryczka

Institute of Computer Science, University of Silesia,
Bedzińska 39,41-200 Sosnowiec, Poland
mariusz.boryczka@us.edu.pl

Abstract. This work develops an idea of the ant colony programming in which an ant colony algorithm is applied to search for the computer program. We shows, that the candidate list (introduced in ant colony system) allows to reduce the algorithm's execution time and improves the quality of results.

Keywords: ant colony system, symbolic regression, ant colony programming, candidate list.

1 Introduction

Given a problem one usually builds an appropriate computer program to solve the problem. Automatic programming makes possible to avoid a tedious task of creating such a program. In automatic programming the program is obtained by specifying first the goals which are to be realized by the program. Then, based on this specification, the program is constructed automatically.

This work develops an idea of the ant colony programming (ACP) in which an ant colony algorithm is applied to search for the program. This approach in many cases may outperform other ones (see. e.g. [3]). We propose the use of the candidate list introduced in ant colony system. We shows, that this list allows to reduce the algorithm's execution time and improves the quality of results.

This work consists of six sections. Section 2 characterizes ant colony systems, section 3 describes the idea of ant colony programming applied to approximation problems. In section 4 the idea of the candidate list is discussed. Section 5 contains the test results, and section 6. concludes the work.

2 Ant Colony Systems

An ant colony system (ACS) has been proposed by Dorigo, Maniezzo and Colorni in 1991 as an ant system for solving the traveling salesman problem [8] and it derives from research on systems inspired by the behavior of real ants. In the artificial ant colony systems the following assumptions are made:

1. Each artificial ant in the ACS has a memory, called the tabu list, in which a set of visited cities for the traveling salesman problem is stored.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 302–311, 2008.

2. Artificial ants are not completely blind. They move according to some probability function determining the next move. Like in the colonies of real ants it depends on the parameters corresponding to the distance of an ant colony's nest from a source of food, and the amount of (artificial) pheromone deposited on the route.
3. Artificial ants live in an environment in which time is discrete.

The aim of a single ant in the traveling salesman problem is to find a salesman tour in the graph, whose nodes are the cities, and the edges connecting the cities have been initialized with some amount of pheromone trail, τ_0. Each ant located at time t at city i makes the decision regarding the next city on its tour using a transition rule. For this goal it generates a random number q, $0 \leq q \leq 1$. If $q \leq q_0$, where q_0 is a parameter of the algorithm, then the "best" available edge is chosen, otherwise the edge is chosen in a random fashion. The rule is:

$$j = \begin{cases} \arg\max\{\tau_{ij}(t)[\eta_{ij}]^{\beta}\} & \text{if } q \leq q_0 \text{ (exploitation)}, \\ S & \text{otherwise (exploration)}, \end{cases}$$

where $\tau_{ij}(t)$ is the amount of pheromone trail on edge $(i,\, j)$ at time t, η_{ij} is the visibility of city j from city i and equals $1/d_{ij}$, where d_{ij} is the distance between cities i and j, β is a parameter which controls the relative weight of the pheromone trail and visibility, and S is a city drawn by using the probabilities:

$$p_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}(t)\cdot[\eta_{ij}]^{\beta}}{\sum\limits_{r \in J_i^k} [\tau_{ir}(t)]\cdot[\eta_{ir}]^{\beta}} & \text{if } j \in J_i^k, \\ 0 & \text{otherwise}, \end{cases}$$

where J_i^k is the set of the cities to which ant k can move when being located in city i (i.e. the set of unvisited cities).

After having found a salesman tour, an ant deposits pheromone information on the edges through which it went. It constitutes a local update of the pheromone trail, which also comprises partial evaporation of the trail. The local update proceeds according to the formula:

$$\tau_{ij}(t+1) = (1-\rho)\cdot\tau_{ij}(t) + \rho\cdot\tau_0$$

where $1-\rho$, $\rho \in (0,\, 1]$, is the pheromone decay coefficient, and τ_0 is the initial amount of pheromone on edge $(i,\, j)$.

After all ants have completed their tours, a global update of pheromone trail takes place. The level of pheromone is then changed as follows:

$$\tau_{ij}(t+n) = (1-\rho)\cdot\tau_{ij}(t) + \rho\cdot\frac{1}{L}$$

where edges $(i,\, j)$ belong to the shortest tour found so far, and L is the length of this tour.

On the updated graph the consecutive cycles of the ant colony algorithm are carried out. The number of cycles is the parameter of the algorithm. The output of the algorithm is the shortest salesman tour found by the ants during the whole experiment.

Ant colony algorithms are one of the most successful examples of swarm intelligent systems and have been applied to many types of problems, including: the traveling salesman problem [8], the problem of task allocation [5], the problems of discrete optimization [7], the vehicle routing problem [11], the graph coloring problem [6], and the graph partitioning problem [13].

3 Ant Colony Programming

Ant colony programming (ACP) is the idea of the ACS used for problem of symbolic regression (in the meaning of Koza [12]). The ACP was described in several works, e.g. [1,2,3,4] and was also considered in [9]. There are two approaches refereed to as expression approach and program approach. In the first one ants create arithmetic expressions in the prefix (Polish) notation while in the second, expressions are built in the form of a sequence of simple assignment instructions.

3.1 Expression Approach

In this approach the ACP is applied for generating arithmetic expressions which are represented in the prefix notation. The ACS as the basis of the ACP is modified in the way given below:

1. The components of graph $G = (N, E)$ have the following meaning: N is the set of nodes, where each node represents either a terminal symbol (including variables), or a function of an arithmetic expression; E is the set of edges representing branches between the parts of an arithmetic expression given in the form of the tree.
2. The probability of moving ant k located in node r to node s in time t equals:

$$p_{rs}^k(t) = \frac{\tau_{rs}(t) \cdot [\gamma_s]^\beta}{\sum\limits_{i \in J_r^k} [\tau_{ri}(t)] \cdot [\gamma_i]^\beta}.$$

   Here $\gamma_s = (1/(2 + \pi_s))^d$, where π_s is the power of symbol s which can be either a terminal symbol or a function, and d is the current length of the arithmetic expression. The definition of γ_s causes that in the process of generating an expression we prefer symbols of large power π_s when the expression is long, i.e. of large d. This guarantees that the generated expression eventually closes and is short.

There are two versions of this approach: with or without the tabu list. If the tabu list is used it holds the information about the path pursued in the graph. It prevents visiting the nodes of the graph several times. If the list is absent, the multiple occurrences of a node in the expression are not prohibited. Thus, in first

Table 1. Power of terminal symbols and functions

Terminal symbol or function	No of arg.	Power
constant, variable	0	−1
NOT, N, 1-arg. functions	1	0
$+, -, *, /$, AND, OR, XOR	2	1

version of the ant colony programming N is, in fact, a multiset as a terminal symbol or a function may (must) be represented in N several times, because these symbols may occur in a generated expression several times as well.

The expression created by a single ant is initiated with symbol F which is the starting node and whose power equals 1. When the expression is expanded by a terminal symbol or function, the power of the expression is increased by the power of this terminal symbol or function (Tab. 1). The expression becomes closed when its power equals 0. The process of determining the power of the expression is illustrated in Fig. 1 for a sample expression $+x + 11$.

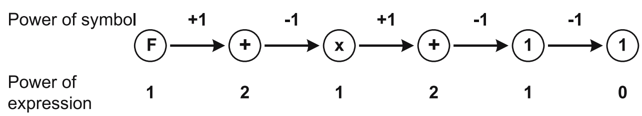


Fig. 1. Determining the power of the expression $+x\,+\,1\,1$

3.2 Instruction Approach

In the program approach the ACP system creates an approximating function in the form of a sequence of assignment instructions. The elements of the system are as follows:

1. The nodes of set N of graph $G = (N, E)$ represent the assignment instructions out of which the desired program is built. The instructions comprise the terminal symbols, i.e. constants, input and output variables, temporary variables and functions. The edges of the graph, $(i, j) \in E$, indicate the instructions (represented by nodes j) which can expand the program, provided its last instruction is the one represented by node i.
2. The probability of moving ant k located in node r to node s in time t equals:

$$p^k_{rs}(t) = \frac{\tau_{rs}(t) \cdot [\psi_s]^\beta}{\sum\limits_{i \in J^k_r} [\tau_{ri}(t)] \cdot [\psi_i]^\beta}.$$

   Here $\psi_s = 1/e$, where e is an approximation error given by the program while expanded by the instruction represented by node s. The approximation error is computed based on the raw fitness function $r_i = \sum_{j=1}^{n} |W_{i,j} - C_j|$, where $W_{i,j}$ is the value returned by the i-th program for the j-th test, C_j is the correct answer for test j, and n is a number of tests.

Table 2. Set of instructions

1	$a = x + 1$	8	$y = a \ / \ b$	15	$y = a * b$
2	$b = x + 1$	9	$y = b + c$	16	$a = x * c$
3	$c = a * b$	10	$a = 10 * x$	17	$y = c + 2$
4	$a = b + 1$	11	$b = a * x$	18	$c = b + 2$
5	$b = 10 * a$	12	$c = 2 * x$	19	$a = c + 2$
6	$c = a + 2$	13	$y = b \ / \ c$	20	$b = c \ / \ a$
7	$y = a * b$	14	$a = b + c$		

Table 3. a) Program generated by the ACP, b) Program after optimization

a)

1	$c = 2 * x$	12	$a = b + 1$	23	$y = a * b$	34	$c = 2 * x$
2	$c = a * b$	13	$a = x * c$	24	$y = a * b$	35	$c = a + 2$
3	$c = b + 2$	14	$b = x * x$	25	$y = a * b$	36	$c = 2 * x$
4	$a = b + 1$	15	$y = b + c$	26	$y = b + c$	37	$a = c + 2$
5	$y = a \ / \ b$	16	$y = c + 2$	27	$y = c + 2$	38	$b = x * x$
6	$y = c + 2$	17	$b = x + 1$	28	$y = a * b$	39	$a = x + 1$
7	$b = x * x$	18	$a = 10 * x$	29	$a = b + c$	40	$c = b + 2$
8	$y = a * b$	19	$a = x + 1$	30	$a = 10 * x$	41	$y = b \ / \ c$
9	$y = a * b$	20	$y = a \ / \ b$	31	$b = a * x$	42	$b = 10 * a$
10	$a = b + c$	21	$y = a \ / \ b$	32	$b = a * x$	43	$b = 10 * a$
11	$a = b + 1$	22	$y = b + c$	33	$b = a * x$	44	$y = b \ / \ c$

b)

1	$b = x * x$
2	$a = x + 1$
3	$c = b + 2$
4	$b = 10 * a$
5	$y = b \ / \ c$

As in the expression approach, the tabu list may be present or absent. If it is used, set N of instructions may (must) include several nodes representing the same instruction. The process of program generation consists in expanding the program by consecutive instructions taken from set N. When the tabu list is exploited, not all instructions from set N need to be used. If the current program gives the approximation error equals 0,the generation process completes. If all the instructions from set N have been exhausted and no exact solution has been found, one may accept a current program, or repeat the generation process with the modified values of parameters, including, perhaps, the set N of instructions. When the tabu list is absent, a program is generated as long as its quality (the approximation error) is not satisfied or the assumed length of the program is achieved.

A sample set of instructions used for program generation is shown in Tab. 2 (as mentioned before each instruction can appear in multiset N more than once; e.g. if every instruction was appeared 5 times, our sample set N would contain 100 instructions).

The instructions consist of constants, input variable, x, the temporary variables, a, b and c, the output variable, y, and functions. A sample program generated by using set N is shown in Tab. 3a.

It is easy to see that the program contains many redundant instructions which can be removed. After elimination of redundant instructions the program takes the form shown in Tab. 3b. The program evaluates the approximating function

$$y = 10(x+1)/(x^2+2)$$

4 The Candidate List in ACP

In the ACS for the traveling salesman problem the set of cities that each ant choses from may become excessively large. This causes the ACS to devote much execution time to evaluate the transition rule (which city to add to the solution). Computational time can be significantly reduced by using a candidate list [10]. It contains for ant k at node i a set of promising nodes (of size cl — the algorithm's parameter) and nodes in the candidate list are evaluated first (Fig. 2). Only when this list is empty the other nodes are evaluated. The ACP with the tabu list may exploit the similar idea. We will present it for the program approach, but the candidate list can be successfully used in the expression approach.

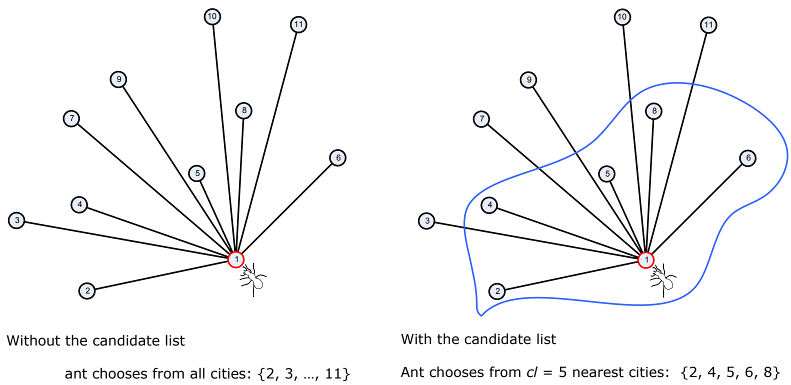


Fig. 2. ACS with the candidate list

If, in the program approach, the tabu list is present, each instruction must be represented in set N several times, because it may occur in a generated program several times as well. An ant completing a program does not have to distinguish elements within this set of identical instructions. Therefore the ant may calculate the transition rule for set of unique instructions only (Fig. 3 and Tab. 4). This idea is similar to the candidate list mentioned before. Such a list includes only unique instructions (promising nodes) so computational time can be reduced as in the ACS with the candidate list.

Table 4. Set of instructions with the candidate list

1	a=x+1	7 (4')	a=b+1
2	b=x+1	8	y=b / c
3 (2')	b=x+1	9 (5')	b=10*a
4	a=b+1	10 (6')	c=a+2
5	b=10*a	11 (8')	y=b / c
6	c=a+2		

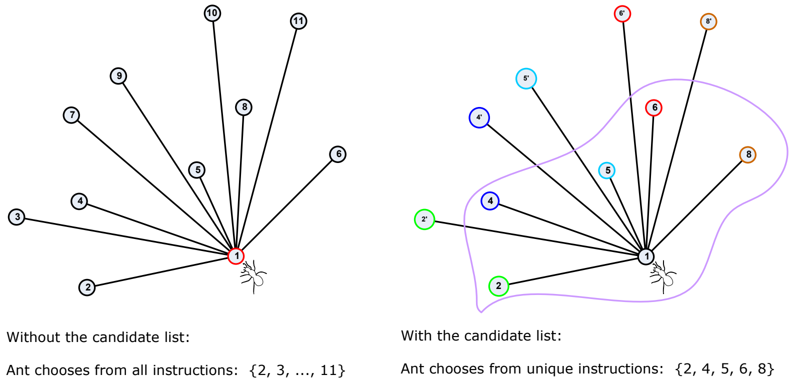


Fig. 3. ACP with the candidate list

5 Experiments

To examine how the candidate list influences the execution time and the quality of solutions generated by the ACP system with the program approach, series of experiments with over 50 functions have been performed. The examined functions were of different form and different number of variables (1, 2 and 3 variables). Among them were:

- $F_1 : y = x^3 - 3x^2 - 4x + 3$,
- $F_2 : y = (1 + \sqrt{u} + \frac{1}{v} + \frac{1}{\sqrt{w^3}})^2$.

The goals of our experiments were:

- comparison of execution time for two versions of ACP: without and with the candidate list; we compared the number of decisions the ants took during the whole experiment, and the cycle number in which each ant found the best solution;
- the quality of solutions generated by ants in both versions of ACP.

Additionally, the parameter determining the number of identical instructions in set N was 2, 3, 4 and 5. Tab. 5–8 show the datasets for our experiments. The number of ants were equal to the number of instructions, and the parameters of ACP (determined by the initial experiments) were: $\tau_0 = 0.00001$, $\beta = 7$, $\rho = 0.1$, $q_0 = 0.9$.

For each function 10 experiments were performed. The results obtained for the number of identical instruction equal to 3 (for other values the results were of similar meaning) are summarized in Tab. 9. It contains execution time (the number of tests) for mentioned versions of ACP. Tab. 10 compares the quality of solutions by use of the APE error defined as follows:

$$\text{APE}_i = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{W_{ij} - C_j}{C_j} \right| \cdot 100\%.$$

Table 5. Values of function F_1

No.	x	y	No.	x	y	No.	x	y	No.	x	y	No.	x	y
1	-2.00	-9.00	7	-0.50	4.13	13	1.00	-3.00	19	2.50	-10.13	25	4.00	3.00
2	-1.75	-4.55	8	-0.25	3.80	14	1.25	-4.73	20	2.75	-9.89	26	4.25	8.58
3	-1.50	-1.13	9	0.00	3.00	15	1.50	-6.38	21	3.00	-9.00	27	4.50	15.38
4	-1.25	1.36	10	0.25	1.83	16	1.75	-7.83	22	3.25	-7.36	28	4.75	23.48
5	-1.00	3.00	11	0.50	0.38	17	2.00	-9.00	23	3.50	-4.88	29	5.00	33.00
6	-0.75	3.89	12	0.75	-1.27	18	2.25	-9.80	24	3.75	-1.45	30	5.25	44.02

Table 6. Instructions for function F_1

1	$a = x + 1$	10	$b = b - a$	19	$a = e\hat{}x$	28	$a = 3 * x$	37	$b = a * b$
2	$b = x - 2$	11	$a = a + 3$	20	$a = a\ /\ 2$	29	$b = b - a$	38	$a = 3 * x$
3	$b = b * b$	12	$a = x + 1$	21	$b = b - a$	30	$a = e\hat{}x$	39	$a = a\ /\ 2$
4	$b = a * b$	13	$b = x - 2$	22	$a = a + 3$	31	$a = a\ /\ 2$	40	$a = a + 3$
5	$a = 3 + x$	14	$b = a * b$	23	$a = x + 1$	32	$b = b - a$	41	$a = e\hat{}x$
6	$b = \sin(b)$	15	$b = a * b$	24	$b = x - 2$	33	$a = a + 3$	42	$a = a\ /\ 2$
7	$a = \cos(a)$	16	$a = 3 * x$	25	$b = a * b$	34	$a = x + 1$	43	$b = b - a$
8	$a = e\hat{}x$	17	$a = x + 1$	26	$b = a * b$	35	$b = x - 2$	44	$a = a + 3$
9	$a = a\ /\ 2$	18	$b = x - 2$	27	$a = 3 * x$	36	$b = a * b$		

Table 7. Values of function F_2

No.	w	x	y	z	No.	w	x	y	z	No.	w	x	y	z
1	0.112	1.165	2.016	6.47	9	0.121	2.064	0.682	13.00	17	0.130	1.280	1.200	8.42
2	0.113	0.607	0.879	17.62	10	0.122	0.893	0.390	43.31	18	0.131	1.086	1.569	7.79
3	0.114	1.300	7.012	4.68	11	0.123	0.245	1.277	37.47	19	0.132	1.036	1.662	7.81
4	0.116	0.336	0.366	77.94	12	0.124	1.235	2.335	5.97	20	0.133	4.288	4.479	2.90
5	0.117	0.081	1.914	197.32	13	0.125	0.321	0.466	57.93	21	0.134	0.641	11.038	8.73
6	0.118	0.281	4.028	25.20	14	0.127	3.741	0.708	10.89	22	0.135	4.754	72.027	2.50
7	0.119	1.223	1.698	6.83	15	0.128	1.160	0.487	26.67					
8	0.120	0.315	1.407	26.19	16	0.129	2.312	0.188	197.36					

Table 8. Instructions for function F_2

1	$a = a + b$	11	$a = y$	21	$a = a * b$	31	$a = y\hat{}2$	41	$a = 1\ /\ x$
2	$b = b + a$	12	$a = a\hat{}2$	22	$b = a - b$	32	$b = a - b$	42	$b = \mathrm{sqrt}(w)$
3	$b = b * 1$	13	$a = a + b$	23	$b = w$	33	$a = a * b$	43	$a = y\hat{}2$
4	$a = 1\ /\ x$	14	$b = b + a$	24	$a = x$	34	$b = w$	44	$b = a - b$
5	$b = \mathrm{sqrt}(w)$	15	$b = b + 1$	25	$a = y$	35	$a = x$	45	$a = a * b$
6	$b = 2$	16	$a = 1\ /\ x$	26	$a = a + b$	36	$a = y$	46	$b = w$
7	$b = a - b$	17	$b = \mathrm{sqrt}(w)$	27	$b = b + a$	37	$a = a + b$	47	$a = x$
8	$a = a * b$	18	$a = y\hat{}2$	28	$b = b + 1$	38	$b = b + a$	48	$a = y$
9	$b = w$	19	$b = a - b$	29	$a = 1\ /\ x$	39	$b = b + 1$		
10	$a = x$	20	$a = 1\ /\ x$	30	$b = \mathrm{sqrt}(w)$	40	$b = a - b$		

Table 9. The average execution time

Function	Candidate list	Cycle number	Number of decisions (experiment)	Average number of decisions (experiment)	Number of decisions (cycle)	Average number of decisions (cycle)
F_1	absent	18	764086	17366	31539	717
	present	19	240151	5458	9355	212
F_2	absent	46	2326610	52878	41558	945
	present	55	896517	20375	11780	268

Table 10. APE error

Function	Candidate list	Cycle number	Number of decisions (experiment)	Average number of decisions (experiment)	Number of decisions (cycle)	Average number of decisions (cycle)
F_1	absent	18	0	0	764086	31539
	present	19	0	0	240151	9355
F_2	absent	46	11.5	15.4	2326610	41558
	present	55	7.9	14.0	896517	11780

The results of experiments are:

– For function F_1, for both versions of ACP (with and without the candidate list) we obtained the same, exact solutions (with the APE error equal to 0). The number of cycles leading to obtain solutions was about 7% greater for the candidate list, but the execution time was 3.18 times shorter when we used the candidate list (the number of tests performed by ants was substantially smaller).
– For function F_2, while using the candidate list we obtain better solutions (about 30%) and shorter execution time (2.6 times). Here, the number of cycles leading to obtain solutions was about 20% greater, but the global execution time was shorter.

6 Conclusions

We proposed the improvement of the ACP system with instruction approach by use of the candidate list introduced in ACS. This improvement may be applied in the ACP with the tabu list. The experiments proved that the candidate list is very useful. It allowed to reduce the time to calculate solutions, and it may help to generate better solutions.

Acknowledgments. This work was carried out under UE Grant: "Decision Support — New Generation Systems".

References

1. Boryczka, M.: Ant colony programming for approximation problems. In: Proceedings of the Eleventh International Symposium on Intelligent Information Systems, Sopot, Poland, June 3-6, 2002, pp. 147–156 (2002)
2. Boryczka, M.: Eliminating introns in ant colony programming. Fundamenta Informaticae 68(1, 2), 1–19 (2005)
3. Boryczka, M.: The ant colony programming in the process of approximation (in Polish). University of Silesia, Katowice (2006)
4. Boryczka, M., Czech, Z.J.: Solving approximation problems by ant colony programming. In: Cantú-Paz, E. (ed.) Late Breaking Papers at the Genetic and Evolutionary Computation Conference (GECCO–2002), New York, July 9–13, 2002, pp. 39–46 (2002)
5. Colorni, A., Dorigo, M., Maniezzo, V., Trubian, M.: Ant system for job–shop scheduling. Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL) 34, 39–53 (1994)
6. Costa, D., Hertz, A.: Ants can colour graphs. Journal of the Operational Research Society 48, 295–305 (1997)
7. Dorigo, M., DiCaro, G.: The ant colony optimization meta–heuristic. In: Corne, D., Dorigo, M., Glover, F. (eds.) New Ideas in Optimization, McGraw–Hill, London (1999)
8. Dorigo, M., Maniezzo, V., Colorni, A.: Positive feedback as a search strategy. Technical Report 91–016, Politechnico di Milano, Milano (1991)
9. Engelbrecht, A.P.: Fundamentals of computational swarm intelligence. Wiley, Chichester (2005)
10. Gambardella, L.M., Dorigo, M.: Solving symmetric and asymmetric TSPs by ant colonies. In: Proceedings of the IEEE Conference on Evolutionary Computation, ICEC 1996, Nagoya, pp. 622–627. IEEE Press, Los Alamitos (1996)
11. Gambardella, L.M., Taillard, E., Agazzi, G.: ACS–VRPTW: A Multiple Ant Colony System for Vehicle Routing Problems with Time Windows. In: Corne, D., Dorigo, M., Glover, F. (eds.) New ideas in optimization, pp. 63–77. McGraw-Hill, New York (1999)
12. Koza, J.R.: Genetic Programming: On the Programming of Computers by Natural Selection. MIT Press, Cambridge (1992)
13. Kuntz, P., Snyers, D.: Emergent colonization and graph partitioning. In: Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats 3, MIT Press, Cambridge (1994)

A Rough Set Approach on Supply Chain Dynamic Performance Measurement

Pei Zheng[1] and Kin Keung Lai[1,2]

[1] College of Business Administration, Hunan University, Changsha, 410082, China
zphnu@yahoo.com.cn
[2] Department of Management Science, City University of Hong Kong, Hong Kong
mskklai@cityu.edu.hk

Abstract. Most of the times, traditional supply chain performance measurement is a static method. However, in the real world, the supply chain is a dynamic system, which needs dynamic performance measurement methods. For the sake of integrative performance measurement of agile virtual enterprise, the traditional Balanced Scorecard is extended into 5 dimensions. According to it, incorporated with the Rough Set theory, the decision table of dynamic performance measurement is constructed. The decision rule set of performance measurement prediction is obtained by attribute reduct and value reduct of decision table. Finally, a calculation example of performance measurement is provided, which shows that the suggested evaluation method is feasible and efficient for dynamic performance measurement and forecasts. Thus, it supplies reasonable analysis and policy making tools for supply chain management.

Keywords: Agile Virtual Enterprise (AVE), Performance measurement, Dynamic Balanced Scorecard, Rough Set (RS), Attribute Reduct.

1 Introduction

Competition among individual enterprises has grown into competition among aggregated supply chains in the world. Performance measurement, an important component of strategic management of a supply chain, is an evaluation system with a set of quantitative indices and/or subjective estimations, by which enterprises or organizations, including whole supply chains, may assess the performance of their daily operations[1]. Every node enterprise of a supply chain is dynamically adjustable, and the corresponding operational time is stochastic. Now there are such performance evaluation systems as ROI (Return on Investment), SCOR (Supply Chain Operations Reference model) and BSC (Balanced Scorecard) in the academe and the industrial world. The BSC performance evaluation system, put forward by Kaplan and Norton [2-4] in 1992, offers a comparative advantage over others, theoretically, and in practice. It requires that performance measurement and energization of an Agile Virtual Enterprise (AVE) should be obtained from measures such as management of finance, customers, internal operations and the ability to learn and grow. What's more, it demands that the management standpoint should be shifted from focusing on short-term goals to long-term strategic goals, and from only results to analyzing and supervising reasons.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 312–322, 2008.

However, traditional BSC theory neglects the factor of supplier while emphasizing the balance between inner and outer layers of the supply chain system. Especially, as far as performance measurement of supply chain of an AVE is concerned, one usually needs to analyze the performance of the 'supplier value chain', to help the enterprise to implement strategic plans and to improve innovation of 'supplier value chain', which decreases production and purchasing cost of the enterprise. With regard to the drawbacks of the traditional BSC, in this paper, the traditional BSC is extended into 5 dimensions, namely, satisfaction degree of the lower members of AVE, inner business process of the supply chain of AVE, satisfaction degree of the supplier, economic benefit of supply chain and innovation and development capability of the supply chain. Based on this, the performance of the supply chain (all past periods) is measured by the Rough Set method. Finally, an example of calculation of Rough Set performance measurement is illustrated, which shows that the evaluation method is scientific and feasible. Thus, the evaluation result provides the basis of supply chain analysis and decision-making.

2 Performance Measurement Index Framework of Supply Chain of AVE

2.1 5D dynamic BSC Model

Brewer and Speh [6] first explored the application of BSC in supply chain performance measurement. They presented a new tool for performance measurement of supply chain, i. e. supply chain BSC, which builds a bridge between SCM and BSC. Based on the supply chain operation process, with focus upon enterprise strategic goals, the supply chain BSC provides the connection between performance indices and enterprise strategies, which can comprehensively evaluate enterprise business achievement and cultivate kernel competitive capability [7].

However, the factor of supplier is neglected in the traditional BSC theory; it emphasizes only the balance between inner and outer layers of the supply chain system. A supplier, as an important component of the value chain, plays a key role in enterprise business. Only if a supplier can make available the desired goods/services in time and in the right quantities, can an enterprise operate regularly and efficiently. Especially, as far as performance measurement of supply chain of an AVE is concerned, analysis of performance of the supplier is necessary to help an enterprise to carry out its strategic plans and to improve innovation of the supplier, which decreases production and purchasing costs. In addition, performance measurement of an AVE is different from that of individual enterprises, in that performance measurement indices of an AVE not only include those of the node enterprise, but also its influence on upstream enterprises, or the AVE. Therefore, such dimensions as satisfaction degree of lower members of AVE, inner business process of supply chain of AVE, satisfaction degree of supplier, economic benefit of supply chain and innovation and development capability of the supply chain must be synthesized while setting up the performance measurement framework. That is to say, the traditional BSC is extended into 5 dimensions, so that the supplier dimension is considered; see Fig. 1.

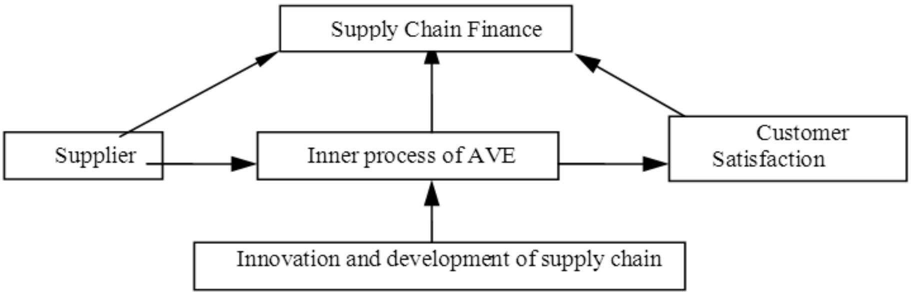


Fig. 1. Dynamic 5D-BSC model

2.2 Selection and Measurement of Supply Chain Performance Indices

Selection of a supply chain performance index is currently in focus as it is difficult to measure performance of an entire supply chain. Opinions of researchers on index framework differ[8, 9]. We think an index should be balanced, in order to implement a proper performance measurement system. A proper performance measurement system not only includes indices that balance short-term and long-term goals, but also those that balance advanced and lagging indices, and quantitative and qualitative indices. In previous sections of the paper, we have presented the 5D-BSC model. So supply chain performance indices must be selected from these five dimensions. For example, indices of the upstream supplier of an AVE cover punctuality of delivery rates and product flexibility. Similarly, indices for the other four dimensions can be inferred and selected from the classical measurement indices, which were deduced in the SCOR model, or as presented by Kaplan [5] and Beamon [10], including capital to turnover ratio, response time of supply chain and customer satisfaction rate, etc.

On the other hand, Kaplan and Norton [5] have argued that the total number of indices in a BSC should be between 15 and 25. Otherwise, too many performance indices may cause information overload; key performance indices may get neglected and the performance measurement system may become too complicated and disorderly. Accordingly, 15 classical indices are selected in our performance measurement framework, i. e. profitability F1, capital to turnover ratio F2, capital turnaround time F3, customer satisfaction degree C1, market share C2, market growth rate C3, response time P1, merchandise turnover ratio P2, reject rate P3, technique advantage P4, profit growth rate L1, employee satisfaction degree L2, research and development cycle of new products L3, punctual delivery rate S1 and flexibility S2. These indices' measurements in an evaluation period are illustrated as follows; see Table 1.

Table 1. Performance indices and measurement of supply chain BSC

Dimension	KPIs	Description	Measurement	Value
Finance	F1	Profit of supply chain	Retained profit/total income ()	quan*
	F2	Capital flow of supply chain	Total sale value/total equity ()	quan
	F3	Cash flow speed of supply chain	Stock supply duration + receivable duration - accounts payable duration	quan
Customer	C1	Customer assessment of product and service	Fuzzy assessment	qual**
	C2	Product market share	Sales of the product / Total market demand for the product (%)	quan
	C3	Market share growth rate	(Market share of this period - market share of last period) / Market share of last period	quan
Inner process	P1	Flexible response of requirement variable	Response time of casual requirement	quan
	P2	Stock capital share	Sale cost / average merchandise share	quan
	P3	Product quality	Number of inferior units / Total yield	quan
	P4	Competitiveness of product technology	Fuzzy assessment	qual
Learning and development	L1	Development potential of enterprise	(Profit of this period - profit of last period)/Profit of last period	quan
	L2	Employee satisfaction	Fuzzy assessment	qual
	L3	Research and development cycle of new product	Statistical mean value	quan
Supplier	S1	Punctual delivery ability of supplier	Punctual delivery times / Total delivery times ()	quan
	S2	Product flexibility	Fuzzy assessment	qual

quan- quantitative; **: qual- qualitative

3 Dynamic Performance Measurement Model of Supply Chain Based on Rough Set[11-14]

3.1 Knowledge Discovery based on Rough Set

A rough set is a set of theory and techniques about representation, learning and induction of uncertain and imperfect data and knowledge. It presents an integrated method to find the rules from a small sample set. According to it, the least forecast set of system description can be obtained which may be utilized for intelligent information processing and knowledge discovery. The knowledge discovery technique based on Rough set usually involves sequential steps such as data preprocessing, decision attribute reduct, value reduct, rule generation and dependence relation extract.

3.2 Performance Evaluation Model Incorporating BSC and Rough Set

It is in the section above that 15 classical indices are selected for performance measurement system. Suppose the performance measurement results of supply chain are classified into four levels - G1, G2, G3 and G4 - according to historical experience or guild regulations. These levels correspond with the performance measurement results Excellent, Good, Satisfactory and Poor, respectively. The performance of AVE supply chain dynamically changes with time. When the performance result is G1 at some point of time, then the performance result at the next observation time is probably G2, G3, G4, or G1 itself. The change depends upon 15 classical evaluation indices. From the perspective of a performance decision table, these 15 indices make up the condition attributes set of the decision table. What's more, these indices have different weights on performance measurement of AVE. On the other hand, {G1、G2、G3、G4} make up the decision attribute set of the decision table. The idea of a performance evaluation model incorporating BSC and Rough set is to diminish the dimension of input space and extracting input features, in order to simplify calculations by the Rough set. The decision rules set of performance forecast is deduced on the basis of attribute reduct and value reduct of decision table. The dependability of decision rules is related with the number of the past performance data. The more the number is, the higher the dependability is. At the next observation time, the performance result can be obtained according to the decision rules, from reduct on the basis of the sample indices at this point of time; see fig 2.

3.3 Calculation Illustration

3.3.1 Form the Decision Table and Preprocess Data

The 5D-BSC of supply chain of AVE consists of 15 key performance indices in five dimensions, including finance, customer, inner process, provider, learn and development. These 15 key performance indices form the condition attributes set of performance decision table of supply chain *C*, *C*={F1, F2, F3, C1, C2, C3, P1, P2, P3, P4, L1, L2, L3, S1, S2}. The attribute value is discrete or continual. The discretization process of all attributes is essential before future information reduct. As for the index with qualitative values, for example, the customer assessment of product and service C1, it can be classified into 4 levels: 1-excellent, 2-good, 3-satisfactory, 4-poor. As for the index with quantitative values, for example, the profit of supply chain, the discretization process is a bit troublesome. We can first get the value interval of the index, according to the historical datum. Assuming that the profit of supply chain in the past has always been between -10% and 20% (the negative value means loss), the discretization process can be implemented in steps of 5% each. Then

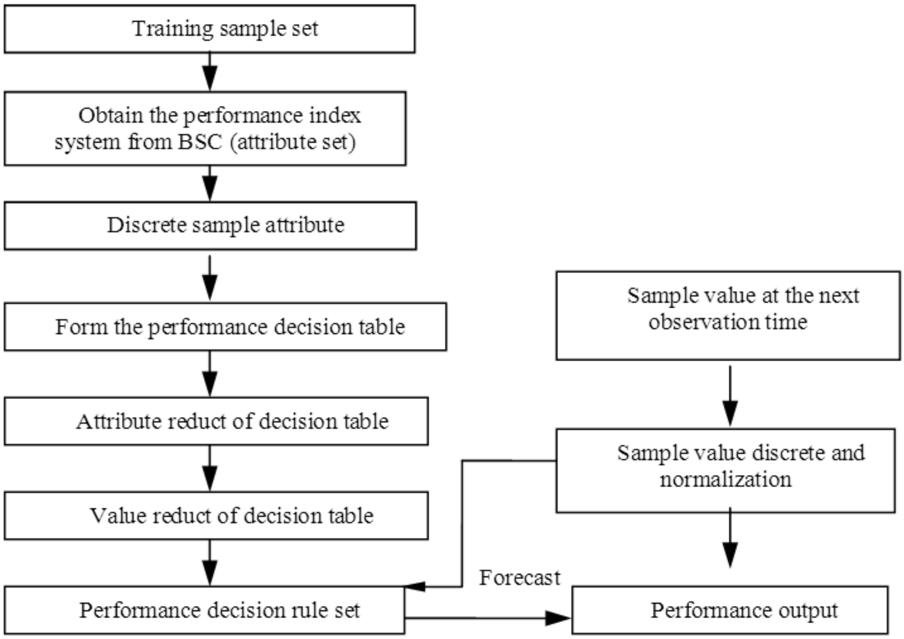


Fig. 2. Performance evaluation model incorporating BSC and Rough set

Table 2. The discretization of the condition attribute of some performance decision table

Dimension	KPIs	Levels	Explication
Finance	F1	1 6	1 [-10%-5%]2 [-5%0]3 [05%]4 [5%10%]5 [10%15%]6 [15%20%]
	F2	1 3	1 [05%]2 [5%10]3 [1015%]
	F3	1 5	1 less than one week2 between one and two weeks3 between two and three weeks4 between three and four weeks5 between four and five weeks
Customer	C1	1 4	1-excellent, 2-good, 3-satisfactory, 4-poor
	C2	1 5	1 [05%]2 [5%10]3 [1015%]4 [1520%]5 [2025%]
	C3	1 3	1 [05%]2 [5%10]3 [1015%]
Inner process	P1	1 5	1 less than one day2 between two and three days3 between three and five days4 between five and seven days5 more than one week
	P2	1 3	1 [05%]2 [5%10]3 [1015%]
	P3	1 5	1 [01%]2 [1%2]3 [23%]4 [35%]5 5
	P4	1 3	1 advanced2 moderate3 poor
Learning and development	L1	1 3	1 [05%]2 [5%10]3 [1015%]
	L2	1 4	1-excellent, 2-good, 3-satisfactory, 4-poor
	L3	1 5	1 less than one month2 between one month and three months3 between three and six months4 between six and twelve months5 more than twelve months
Supplier	S1	1 4	1 [95100%]2 [90%95]3 [8590%]4 85
	S2	1 3	1 good2 satisfactory3 poor

F1 has six types of values, from 1 to 6. These levels correspond with the value intervals [-10%-5%][-5%0][05%][5%10%][10%15%] and [15%20%]". Similarly, the other attributes are also processed into discrete values; see Table 2.

Assuming that the discrete values of the condition attributes of a performance decision table are obtained, the performance evaluation result is presented, based on historical datum; see Table.3.

Table 3. The performance decision table of AVE supply chain

U	The condition attribute set C															The decision attribute d
	F1	**F2**	**F3**	**C1**	**C2**	**C3**	**P1**	**P2**	**P3**	**P4**	**L1**	**L2**	**L3**	**S1**	**S2**	
T1	5	3	1	2	4	3	2	2	1	1	3	2	2	1	2	**G1**
T2	3	3	1	2	4	3	2	2	5	2	1	2	2	1	2	**G3**
T3	4	3	1	3	4	3	2	2	5	2	3	2	2	1	2	**G2**
T4	1	3	1	3	4	3	2	2	5	2	1	2	5	1	3	**G4**
T5	3	3	1	3	3	3	2	2	5	2	1	2	5	2	3	**G3**
T6	4	3	1	3	3	2	2	2	5	2	1	2	2	2	2	**G2**
T7	3	3	1	2	3	3	1	2	4	2	1	2	5	2	2	**G3**
T8	3	3	3	2	3	3	2	2	4	2	1	2	5	2	3	**G3**
T9	5	3	1	1	4	3	2	2	1	1	3	2	2	1	2	**G1**
T10	3	2	3	2	3	3	1	2	5	2	1	3	2	2	2	**G3**

Table 3 shows the performance evaluation results of AVE supply chain in the past 10 observations. For example, at observation time T2, the value of the performance measurement index set {F1,F2,F3,C1,C2,C3,P1,P2,P3,P4,L1,L2,L3,S1,S2} is {3,3,4,2,2,1,5,1,5,2,1,2,4,3,2}, and the performance evaluation result is G3. It means that when the profitability of the supply chain is between 0 and 5%, capital turnover ratio between 10% and 15%, cash flow speed less than one week, customer assessment of product and service good, market share between 15% and 20%, market growth rate between 10% and 15%, response time between two and three days, stock capital share between 5% and 10%, reject rate more than 5%, competitiveness of product technology satisfactory, profit growth rate less than 5%, employee satisfaction good, research and development cycle of new product between one and three months, punctual delivery ability of provider more than 95% and product flexibility satisfactory, the performance of supply chain is regarded as satisfactory.

3.3.2 Attribute Reduct and Value Reduct of Performance Decision Table

As far as the performance decision table in Table 3 is concerned, redundant condition attributes of the decision table can be eliminated by the attribute reduct technique. We first calculate the indiscernible matrix C_D (symmetry matrix). The value of element of the matrix is:

$(C_D)_{1,2}$={F1, P3, L1 };

$(C_D)_{1,3}$={F1, C1, P3, P4 };

$(C_D)_{1,4}$={F1, C1, P3, P4, L1, L3, S2};

$(C_D)_{1,5}$={F1, C1, C2, P3, P4, L1, L3, S1, S2};

$(C_D)_{1,6}$={F1, C1, C2, C3, P3, P4, L1, S1};
$(C_D)_{1,7}$={F1, C2, P1, P3, P4, L1, L3, S1 };
$(C_D)_{1,8}$={F1, F3, C2, P3, P4, L1, L3, S1, S2};
$(C_D)_{1,10}$={F1, F2, F3, C2, P1, P3, P4, L1, L2, S1};
$(C_D)_{2,3}$={F1, C1, L1 };
$(C_D)_{2,4}$={F1, C1, L3, S2};
$(C_D)_{2,6}$={F1, C1, C2, C3};
$(C_D)_{2,9}$={F1, C1, P3, P4, L1};
$(C_D)_{3,4}$={F1, L1, L3, S2};
$(C_D)_{3,5}$={F1, C2, L1, L3, S1, S2};
$(C_D)_{3,6}$={C2, C3, L1, S1};
$(C_D)_{3,7}$={F1, C1, C2, P1, P3, L1, L3, S1};
$(C_D)_{3,8}$={F1, F3, C1, C2, P3, L1, L3, S1, S2};
$(C_D)_{3,9}$={F1, C1, P3, P4};
$(C_D)_{3,10}$={F1, F2, F3, C1, C2, P1, L1, L2, S1};
$(C_D)_{4,5}$={F1, C2, S1};
$(C_D)_{4,6}$={F1, C2, C3, L3, S1, S2};
$(C_D)_{4,7}$={F1, C1, C2, P1, P3, S1, S2};
$(C_D)_{4,8}$={F1, F3, C1, C2, P3, S1, S2};
$(C_D)_{4,9}$={F1, C1, P3, P4, L1, L3, S2};
$(C_D)_{4,10}$={F1, F2, F3, C1, C2, P1, L2, L3, S1, S2};
$(C_D)_{5,6}$={F1, C2, L3, S2};
$(C_D)_{5,9}$={F1, C1, C2, P3, P4, L1, L3, S1, S2};
$(C_D)_{6,7}$={F1, C1, C3, P1, P3, L3};
$(C_D)_{6,8}$={F1, F3, C1, C3, P3, L3, S2};
$(C_D)_{6,9}$={F1, C1, C2, C3, P3, P4, L1, S1};
$(C_D)_{6,10}$={F1, F2, F3, C1, C3, P1, L2};
$(C_D)_{7,9}$={F1, C1, C2, P1, P3, P4, L1, L3, S1};
$(C_D)_{8,9}$={F1, F3, C1, C2, P3, P4, L1, L3, S1, S2};
$(C_D)_{9,10}$={F1, F2, F3, C1, C2, P1, P3, P4, L1, L2, S1};

The values of other elements in the matrix C_D are zero. Followed by the attribute reduct algorithm, based on the indiscernible matrix and logic arithmetic, all reducts of the performance decision table in Table 3 are presented as follows:

{F1, C2}{F1, C3}{F1, L1}{F1, S1}{P3, C1, L3, C3}{P3, C1, S2, C3}{P3, C1, C2}{P3, C1, S1}{P3, C1, L1}{P4, L1, C1}{P3, L1, L3, C2}{P3, L1, S2, C2}{P4, L1, L3, C2}{P4, L1, S2, C2}{P3, L1, L3, C3}{P3, L1, S2, C3}{P4, L1, S2, C3}{P4, L1, L3, C3, S1}{P4, L1, L3, C3, P1, F3}

Moreover, the value reduct is made by the induce value reduct algorithm and the decision rule sets of performance forecast are obtained. Since the decision attribute value is classified into four levels, the information system has four concepts; the most general rules of these four concepts are listed as follows:

G1: $(F1=5) \wedge (F2=3) \wedge (C1=2) \wedge (C2=4) \wedge (P3=1) \wedge (P4=1) \wedge (L1=3) \wedge (L2=2) \wedge (S1=1) \xrightarrow{1} G1$

This rule indicates that if a certain AVE's profitability is 10%-15%, capital turnover rate is 10%-15%, customer satisfaction is high, market share exceeds 15%, defective rate is below 1%, it has advanced technology, profit increase ratio exceeds 10%, employee satisfaction is high and on-time-delivery rate is 95%, then we can conclude that the comprehensive performance is Excellent; the confidence degree of the rule is 1.

G2: $(F1=4) \wedge (F2=3) \wedge (F3=1) \wedge (C1=3) \wedge (P1=2) \wedge (P3=5) \wedge (L2=2) \wedge (L3=2) \wedge (S2=2) \xrightarrow{1} G2$

This rule indicates that if a certain AVE's profitability is 5%-10%, capital turnover rate is 10%-15%, cash turnover is within one week, customer satisfaction is common, response time is 2 or 3 days, defective rate exceeds 5%, employee satisfaction is also high, new product development cycle is one to three months, and flexibility of supplier is common, then we can conclude that the comprehensive performance is Good; the confidence degree of the rule is 1.

G3: $(F1=3) \wedge (C3=3) \wedge (P2=2) \wedge (P4=2) \wedge (L1=1) \xrightarrow{1} G3$

This rule indicates that if a certain AVE's profitability is below 5%, market share growth rate is 10%-15%, stock capital share is 5%-10%, competitiveness of product technology is normal and development potential of enterprise is below 5%, then we can conclude that the comprehensive performance is Satisfactory; the confidence degree of rule is 1.

G4: $(F1=1) \wedge (F3=1) \wedge (C1=3) \wedge (C2=4) \wedge (P3=5) \wedge (P4=2) \wedge (L1=1) \wedge (L2=2) \wedge (L3=5) \wedge (S2=3) \xrightarrow{1} G4$

This rule indicates that if a certain AVE's profitability is -5% (loss is 5%), cash flow speed of supply chain is less than one week, customer assessment of product and service is normal, market share is 15%-20%, defective rate exceeds 5%, competitiveness of product technology is normal, profit increase rate below 5%, employee satisfaction is high, research and development cycle of new product is longer than one year and suppliers' flexibility is bad, then we can conclude that the comprehensive performance is Poor; the confidence degree of rule is 1.

According to the above rule of attribute reduction of a decision table, we can evaluate the dynamic comprehensive performance of AVE in a given period. When the condition attributes set doesn't completely comply with the antecedents of the rule, we can choose the secondary general if-then rule (or the if-then rule with fair good trust) for performance evaluation. In the real world, owing to inadequacy of historical data, or if the values of some attributes are lost, there may be some conflict in the deduced decision rules. Under these circumstances, inconsistent reasoning of rules is necessary. The corresponding reasoning techniques include weighted synthesis, rule pruning, high priority with high trust, etc.

4 Conclusion

This supply chain performance evaluation system uses a quantified indicator system to measure and evaluate the performance of AVE; it plays a key role in supply chain's daily operational and management activities. However, the traditional SC-BSC neglects the factor of supplier while emphasizing the balance between inner and outer layers of the supply chain system. Meanwhile, it neglects the real time and dynamics of AVE's performance evaluation. Therefore, the traditional BSC is extended into 5 dimensions in this paper, based on the theory of Rough Set, to work out the future evaluation results of a dynamic supply chain and its trends; it also provides the basis of supply chain analysis and decision-making. Illustration provided implies that the evaluation method is scientific and feasible. Future research should include, but not be limited to: Selection and measurement of supply chain dynamic performance indices, quantification and pretreatment of evaluation indices, evaluations of decision rules rapid reduction and inconsistency inference, etc.

References

1. Van Hoek, R.I.: Measuring the Unmeasurable-Measuring and Improving Performance in the Supply Chain [J]. Supply Chain Management 3, 187–192 (1998)
2. Kaplan, R.S., Norton, D.P.: The balanced scorecard-measures that drive performance, pp. 71–79. Harvard Business Review (January-February 1992)
3. Kaplan, R.S., Norton, D.P.: Putting the Balanced Scorecard to Work. Harvard Business Review, Boston (9/10) (1996)
4. Kaplan, R.S., Norton, D.P.: Using the Balanced Scorecard as a Strategic Management System. Harvard Business Review, Boston (3/4) (1996)
5. Kaplan, R.S., Norton, D.P.: The Strategy Focused Organization: How Balanced Scorecard Companies Thrive in the New Competitive Environment, p. 2. Harvard Business School Press, Boston (2001)
6. Brewer, P.C., Speh, T.W.: Using the Balanced Scorecard to Measure Supply Chain Performance. Journal of Business Logistics 21(1) (2000)
7. Ma, S.H., Li, H.Y., Lin, Y.: Study on application of Balanced Scorecard to performance measurement of supply chain. Industry Engineering and Management (4), 5–10 (2002)
8. Lohman, C., Fortuin, L., Wouters, M.: Designing a performance measurement system: A case study [J]. European Journal of Operational Research 156, 267–286 (2004)
9. Gijerdrum, J., Shah, N.: A Combined Optimization and Agent2based Approach to Supply Chain Modeling and Performance Assessment [J]. Production Planning and Control 12, 81–88 (2001)
10. Beamon, B.M.: Measuring Supply Chain Performance [J]. International Journal of Operations & Production Management 19, 275–292 (1999)
11. Pawlak, Z.: Rough set-theoretical aspects of reasoning about data [M]. Kluwer Academic Publishers, Boston, MA (1991)

12. Skowron, Rauszer, C.: The discernibility matrices and functions in information systems [A]. In: Slowinski R. Intelligence decision support-handbook of application and advances of the rough sets theory [C], pp. 331–362. Kluwer Academic Publishers, Dordrecht (1992)
13. Rosetta [EB/OL]. Knowledge Systems Group, Dept. of Computer and Info. Science, Norwegian University of Science and Technology, Trondheim, Norway and Group of Logic, Inst. of Mathematics, University of Warsaw, Poland http://rosetta.lcb.uu.se/general/
14. Rough Analysis [EB/OL]. Enrique Alvarez (August 1998), http://www.lsi.upc.es/~ealvarez/rough.html

Analyzing Knowledge Exchanges in Hybrid MAS GIS Decision Support Systems, Toward a New DSS Architecture

D. Urbani and M. Delhom

SPE LAB UMR CNRS 6134 - University of Corsica
Quartier Grossetti BP52 - 20250 Corté – France
durbani@laposte.net, delhom@univ-corse.fr

Abstract. In this paper we present the strong points and the weakness of GIS used as decision support systems. We expose the advantages of adding MAS to GIS. We present the common critters used to characterize the MAS-GIS links, and introduce a new point of view about these relations studying the nature of the knowledge used to help decision makers. We explicit the nature and the purpose of these knowledge, and we study the knowledge exchanges between MAS, GIS, experts and decision makers. Then, we propose a new architecture, SMAG, for hybrid MAS-GIS decision support systems development platforms. We implement this architecture and build a decision support system dedicated to the fresh water problem. Thus we illustrate the relevance of our operational decision making solution that is of one the first to make possible to take into account together individual and social behaviors, spatial component and specificities of target sites.

Keywords: multi-agent systems, geographic information systems, decision support systems, knowledge management, model, water, corsica.

1 Introduction

In this paper we present the decision support systems focusing on those based on geographic information systems (GIS). We highlight the strengths and the weak points of this approach and we propose solutions to reduce its main insufficiencies. To overcome the limits due to the intrinsic static property of GIS representations, we study the opportunity to couple a GIS with an individual based model of the system built following the multi-agents systems (MAS) paradigm.

First, we present our work about coupling MAS and GIS, and then we introduce and explain together the nature, the purposes, and the management of the knowledge required to support decision makers. Using this analysis, we propose a framework for hybrid MAS-GIS decision support systems. Following our ideas about knowledge and decision making process, we present our SMAG platform architecture, dedicated to the development of hybrid MAS-GIS decision support systems (DSS). Finally we present CORMGIS, the first SMAG compliant platform dedicated to DSS development, and its first application: HydroValleA, a decision support system used to solve the fresh water problem in Mediterranean islands.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 323–332, 2008.

2 DSS Based on GIS

Due to their capabilities GIS are used by decision makers in geo marketing, navigation, agriculture, advertising, infrastructure maintenance and many other fields [1]. Use GIS is a fruitful approach for choice's problems with strong spatial component. But this approach is limited by the intrinsic static nature of GIS representations witch makes impossible to take into account the systems ' dynamics [2]. To reduce this weakness, GIS are commonly coupled with generic dynamic models of the systems. The advantages of using such a way are obvious to most users: integrate the spatial dimension and take into account the system's process. In the aim to take into account the individual properties and behaviors, linking GIS with Multi-Agents paradigm is one of the major topics in decision support systems researches.

3 Advantages of a Hybrid MAS-GIS Approach

The MAS-GIS approach to build decision support systems provides many advantages:

- This solution makes available the considerable amount of data and information produced and used by the all the major stakeholders who commonly employ GIS.
- The MAS paradigm, used to take into account the process, underlies a bottom-up approach of the system's dynamics where the global behavior emerges from the sum of individuals' behaviors and interactions [3]. This approach, based on generic models of the stakeholders, makes possible to consider in the decision making process the heterogeneity of systems where protagonists have diverging interest [4].
- This approach enables the models' builders to validate the dynamic models comparing huge simulation data with field observations. Thus a combined use of GIS and MAS improves the reliability of the DDS [5].
- Moreover people feel more implicated during the decision making process when they can view their own territory through a GIS. Thus more information can be collected about problems. This solution makes easier the cooperation between stakeholders and decision makers and thus improves the results of decision process.

Our hybrid approach fills the lacks of DSS based only on GIS taking into account the dynamic of the systems using agent based models. The GIS data are used to fill the too genericity of the elements obtained using DSS based only on MAS. Finally, the decision makers dispose of elements fitted to the each targeted site. Thus they can improve and explain their choices [6].

4 Different Approaches to Link GIS and MAS

According to Mandl [7] and Koch [8] we use two criterions to characterize the MAS - GIS links:

- The links can be "dynamic" or "static",
- The interaction "tight" or "loose".

The first criterion is related to the access to the geographic data during the simulations made by the MAS. When the MAS only access to the GIS data to initialize the software and instance the agents, the link is "static". When the MAS also access to the GIS during the simulations to instance new agents, modify the environment, and store the data obtained during computations, the link is "dynamic" [9].

The second criterion is related to the ways of interactions and communications between the MAS and GIS softwares. When the softwares are linked sharing code or call foreign functions, the link is said "tight". This way requires to access to two the softwares' codes. Moreover it implies parallel evolutions of the two softwares. On the other hand, when the GIS and the MAS softwares only use data to interact and to communicate, it becomes possible to change the GIS or the MAS softwares to others, and makes easier to develop aside the MAS platform and the GIS softwares. This kind of link is said "loose".

5 Analyzing Knowledge in Decision Support Systems

Beyond the previous criterions based on a software developers' point of view, we follow a new approach analyzing the links between MAS and GIS. We consider the nature and the management of the knowledge used during the decision making process. Compared to other approaches, we are the first to focus on the nature of the knowledge engaged in MAS-GIS systems to build DSS. According to Heijst [10], we differentiate the domain and the solving knowledge.

Domain and solving knowledge

The domain knowledge embraces all the data about the system: state of the system, rules governing its dynamic, features of the stakeholders, and constraints of the problem. This knowledge is used to build models that makes possible to the decision makers to valuate the different options. The solving knowledge holds the procedures to exploit the domain knowledge, the know how to solve problems, and finally to provide help to decision makers. The DSS use computerized translations of problematic and apply solving algorithms and procedures to test the different alternatives and find the best solutions. When the knowledge can be translated into a computerized language, decision support systems can easily be built. Unfortunately when these cognitions belong to experts and are issued from their background, they cannot always be translated into a computerized language and the decision support systems must include ways to interact with experts and use these heuristic knowledge.

GIS knowledge management

Considering the intrinsic capabilities of GIS softwares we explicit the main GIS functions in hybrid MAS-GIS decision support systems:

- Administer and filing the database;
- Consult, control and store the geographical data;
- Build spatial representations;
- Shape and show the systems;
- Define and manage scenarios;
- Store and analyze experimentation's results.

The GIS software manages all the attributes and metadata describing all the stakeholders of both problematic and decision making process. The GIS supervises domain knowledge and forms a static and factual representation of the problematic built with a geographer point of view aiming to help the decision makers. The spatial and statistical analyze GIS native functions make it depositary of a part of the computerized solving knowledge useful to help the decision makers. The expert's heuristic knowledge can be take into account thank to the GIS interfaces.

MAS multilevel knowledge management

Following the domain/solving knowledge classification, we explicit the nature and the purpose of the knowledge supervised by the MAS in hybrid MAS-GIS decision support systems. Analyzing the knowledge engaged from the decision maker to the core of the MAS, we organized it in a hierarchical multilevel structure, mirroring its mobilization during the steps of the decision making process [11].

Four knowledge levels:

In order to take into account the complexity of decision support process we defined four levels of knowledge:

1) The top level includes the global solving knowledge about systems and problems. It's used to explain the problems' nature, to search ways to find solutions and help the decision makers. To use a decision support system, this corpus is essential to the decision makers, enabling them to describe, to compare and to valuate the various options. This knowledge is necessary to plan experimentations. When the solving knowledge is fully translated in a computer language, the software can exhaustively explore all the alternatives. On the other hand, when the time of computation becomes too long due to a combinatorial explosion, or if this knowledge is not fully explainable in a programming language, the decision makers can appeal to the experts' heuristic knowledge and know how through graphic user interfaces.

2) The second level contains the domain knowledge about systems and problems. This corpus is useful to describe the global systems' dynamics. It's used to model the systems' architectures following the multi-agents paradigm, identifying the agents and the interactions between the agents "stakeholder". These agent based model are built to provide help to decision makers and must be in adequacy with the problem. This level crystallizes the academic knowledge about the systems considered under the systemic and global point of view.

3) The domain knowledge used to describe and to model the agent's behaviors, the attributes and the dynamic of the environment, represents the third level of knowledge. In agent based models, the global behavior of the system emerges from the sum of individuals' behaviors and interactions. This level aggregates the knowledge about each stakeholder: their qualifying attributes, their possible actions, how they update their beliefs, how they perceive and modify their environment, how they communicate and exchange information, how they decide what to do to satisfy their desires. This level of knowledge about the agent's behaviors and dynamic models is closely related with the static model managed by the GIS: the géodatabase is used to initialize the DSS and to store the data obtained during computations.

4) The forth level contains the agents' beliefs built dynamically during the experimentations. In MAS, each agent builds its own system of representation about itself, the others agents and its environment. The agents use them to decide what to do to reach their goals. This level is built dynamically by the agents during the simulations from their percepts (captors, messages) and background without access to the géodatabase.

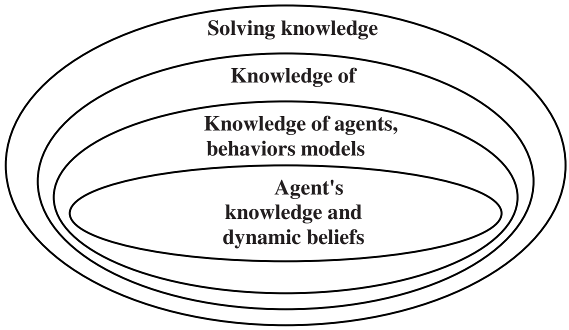


Fig. 1. Layered knowledge and MAS in DSS

Figure1 summarizes the multilayered structure of knowledge engaged in the MAS in hybrid MAS GIS decision support systems.

Systems' representations in MAS and GIS

In MAS-GIS based DSS two system's representations are used together. According to Oussalah [12] who considers modeling as an epimorphism between the real system and its model, we use 6 morphisms ($\Phi 1$, $\Phi 2$, $\psi 1$, $\psi 2$, $\varphi 1$, $\varphi 2$) to formalize the relations between the real, the GIS representation, and the MAS representation.

The constraints between the GIS and MAS models are due to their respective functions in the decision support systems. Taking into account that the first function of the GIS is the initialization of the MAS model, all the elements and stakeholders must compulsory be included in the GIS. Thus the $\varphi 1$ morphism is onto. Moreover in order to insert the data generated by the MAS in the GIS database, the $\varphi 2$ morphism should be injective

Knowledge exchanges

Knowledge exchanges occur during the decision making process using hybrid MAS-GIS decision support system between MAS, GIS, experts and decisions makers. During computations, MAS and GIS exchange domain knowledge about the stakeholders whose behaviors are modeled in MAS, while their attributes are recorded in the GIS database. These exchanges concern only the agents' static attributes and the data generated during simulations. Moreover the GIS feeds the MAS with collections of events constituting the scenarios and the options that are considered by the decision makers. In the other way, the MAS feeds the GIS with all the events occurring during the experimentations coming from scenarios, decision makers' actions, agents' initiatives, social networks' evolutions, and environment natural evolutions.

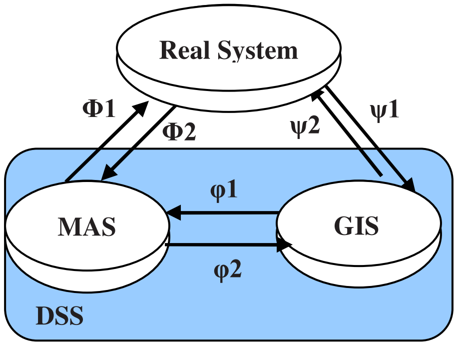


Fig. 2. Morphisms in hybrid MAS-GIS DSS

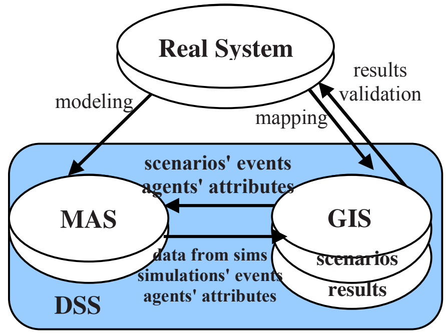


Fig. 3. Data exchanges between MAS &GIS

Figure 3 shows the data exchanges between MAS, GIS and the real system. At the end of the computations, experts compare the data obtained during the decision making process with the real data to validate the results and the information given by the DSS. Thus the decision makers can trust the software indications.

Besides, experts are useful to feed the géodatabase and model the stakeholders behaviors using the data obtained from the field's observations. They transfer their fully computerized domain knowledge to the MAS and GIS. Moreover, experts transfer their heuristics domain knowledge during the decision making process and simulations through dedicated DSS software interfaces.

During decision making process, experts and decision makers exchange solving knowledge with the decision support system. Together they can fix the experiment plans or define algorithms to obtain best elements to make the right decision. When the solving knowledge cannot be completely transferred to the software, experts help the decision makers to drive the experimentations through graphic user interfaces. Thus the decision support system can take into account their heuristic knowledge, know how and backgrounds.

6 SMAG Architecture

Using the previous considerations we propose a new architecture, SMAG, to build platform dedicated to the development of hybrid MAS-GIS decision support systems. We justify our linking choices, explicit the MAS and GIS functions, and specify data and knowledge exchanges.

To take advantage of the GIS capabilities to analyze data and its embedded solving knowledge, we decide to use a dynamic link between MAS and GIS. Thus the géodatabase is used beyond the initialization; the data (attributes of new agents, environmental and social networks changes) generated during the experimentations are stored in GIS database to be analyzed.

Taking into account that the decision makers need géodatabase from several institutions and companies that use different GIS softwares, we decide to not link MAS and GIS mixing their source code. Thus, during the implementation of the SMAG architecture, developers can build new GIS or MAS softwares, or reuse (commercial/open-source) existing softwares already employed by the different stakeholders. This "loose" linking choice makes easer the evolutions and updates of the MAS and GIS.

The modular SMAG functional architecture showed Figure 4 takes into account our analysis about the knowledge exchanges between the MAS, GIS, decision makers and experts. Our modular architecture can be implemented without mix the MAS and GIS softwares' codes. The SMAG architecture is a global solution to support decision makers: from the problematic and system's modeling to the scenario's exploration.

The SMAG's "GIS unit" is standard and so can directly be implemented by any GIS software, thus all the stakeholders can use the software they usually employ. The "GIS Unit" is first used to acquire and to manage geo data obtained from the ground watching. It is also used to define the events and the scenarios. Its capabilities make it possible to analyze the data generated during the experimentations in the aim to support the decision makers.

The SMAG's modular "MAS unit" is designed to take into account the experts' and decision makers' knowledge. It's useful to manage the experimentations.

The "*New model entry*" module is dedicated to model the system and the stakeholders following the multi-agents paradigm, using the MAS unit modeling directives.

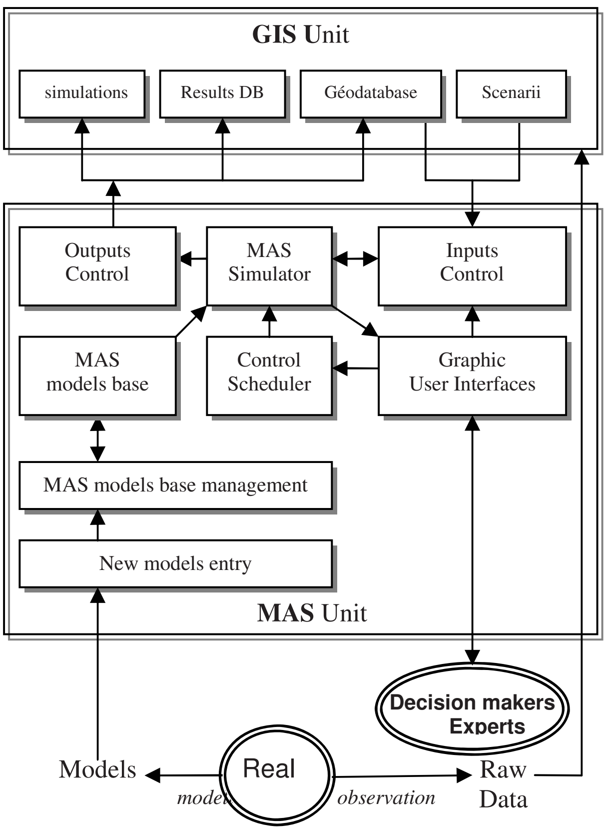


Fig. 4. SMAG functional architecture

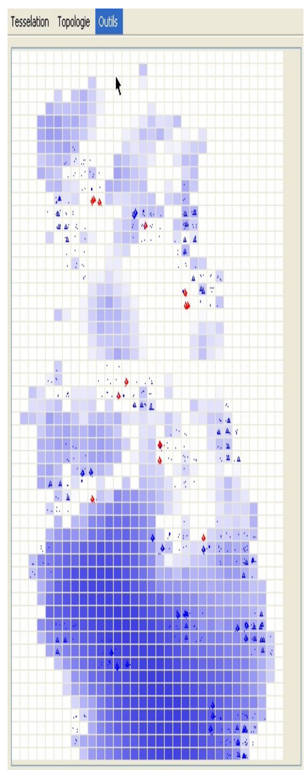


Fig. 5. Agents in north Corsica

The "*MAS models management*" module, dedicated to the models' management. It's essential to link experiments, simulations, and hypothesis about the stakeholders' behaviors and the systems' dynamics.

The "*Inputs Control*" module supervises the data inputs to the core simulator. The first function of this component is to question the géodatabase to follow the current scenario, to read the new agents' attributes and environmental changes. Its second function is to take into account the event and directives coming from the DSS users through the graphic interfaces.

The "*Control Scheduler*" module controls the time step and the simulations' advance. It's also run the planed experiences and the solving algorithms developed to find the best solution.

The "*Outputs Control*" module supervises the data recorded in the géodatabase. The core simulator feed it with the events (from scenarios, users, and agents themselves) and the data produced during the experimentations. Afterwards these records can be analyzed to provide support to decision makers.

The "*Graphic Users Interface*" is dedicated to the software communication with humans being during the experimentations. It provides a time step control and is helpful to plan experiences. Moreover it makes possible to control agents and environment. Through the "*Inputs control*" module, the users' actions are perceived as events by the core simulator like scenario's events. Experts and decision makers can use this module to add their heuristic knowledge to the decision making process.

7 Application

Our theoretical works about the knowledge in hybrid MAS GIS decision support systems and the SMAG architecture are keeping alive in our CORMGIS platform. This first implementation of the SMAG architecture is build customizing and augmenting the open source CORMAS [13] multi-agent platform developed by the CIRAD [14], and the ArcGIS software of the ESRI Company [15].

We used CORMGIS to develop the HydroValleA decision support system dedicated to the fresh water problem in the Mediterranean islands. This software takes into account the characteristics of the climate, the consumers' behaviors, the water supply network, a watershed hydrologic model, the resource manager and the public management policy. This DSS makes possible for the first time to highlight the influence of individuals' and socials' behaviors linked to the stakeholders' spatial distribution. HydroValleA enables us to study a 32x16 km target area situated in the north of the Corsica Island, represented figure 5 by a 64x32grid. In this zone, hydraulically isolated from the rest of the island, the peak demand occurs during the dry period. Moreover the water demand is only from the permanent inhabitants and the tourism activity; industry and agriculture consumptions can be neglected. In this area, our géodatabase records 33 drillings, 12 tanks, 534 homes and hotels. Our experimentations first focus on the importance of propagation of consumers' individual behaviors [16] to cope with dry periods, and to limit the water shortages. Moreover we run experimentations to find the best policy to control the use of water during lack of rain: the best efficiency with the lowest frequency of police's controls [17]. Developing CORMGIS we illustrate the relevance and capabilities of our knowledge based of hybrid MAS-GIS decision support systems. Up to now the DSS dedicated to fresh water problem were taken into account the specificities of the target sites without the systems' dynamics employing GIS; or they were taken into account the dynamic neglecting the specificities of the target site when they were based on generic models. HydroValleA makes now possible to take into account together the systems' dynamics and the specificities of the target sites [11]. It confirms that CORMGIS can be used to develop DSS that take into account the heterogeneity of stakeholders' social and individual behaviors, and the spatial dimension of support decision makers.

8 Conclusion and Perspectives

In this paper we have presented the weakness of using GIS as decision support systems due to their intrinsic static nature. We have exposed the advantages of adding MAS to take into account the systems' process, and the current state of the art about linking MAS and GIS. Overtaking the static/dynamic and tight/loose criterions commonly used to describe the links between MAS and GIS, we introduced a new point of view focusing on knowledge used and exchanged in such hybrid decision support systems. Building the SMAG architecture, we concretized this approach. The CORMGIS implementation and the HydroValleA decision support system used in Corsica confirm the relevance and capabilities of our approach. CORMGIS stands out from most of the DSS making possible to take into account the stakeholders' individual and social behaviors, the spatial components and the specificities of each targeted site. Making a distinction between the decision makers solving knowledge and the domain knowledge about the target site, we introduced a novelty in the decision support system architectures and development. Following this new approach we propose to break the DSS building process into two steps. First collect the domain knowledge to build a multi-agents model and map the target site. Then collect the solving knowledge among decision makers and managers. We plan to use CORMGIS to build new decision supports system dedicated to other domains: fire forest, fishing, sylviculture... Furthermore our future works will concentrate on the theoretical aspects of the interactions and exchanges between DSS and their users.

References

1. Shahriari, N., Tao, C.: GIS applications using agent technology. In: Proceedings of the Symposium on Geospatial Theory, Processing and Applications (ISPRS Commission IV), July 9-12, 2002, Ottawa, Canada (2002)
2. Gonçavels, A.S., Rodrigues, A., Correia, L.: Multi-Agent Simulation within Geographic Information Systems. In: Coelho, H., Espinasse, B. (eds.) Proceedings of 5th Workshop on Agent-Based Simulation (2004)
3. Axtell, R.: Why Agents? On the Varied Motivations for Agent Computing in the Social Sciences. In: Working Paper 17, Center on Social and Economic Dynamics, Brookings Institution (2000)
4. Wooldridge, M.: An introduction to multi-agent systems. John Wiley & Sons, Chichester (2001)

5. Sargent, R.G.: Verification and validation: some approaches and paradigms for verifying and validating simulation models. In: Proceedings of the 33rd conference on winter simulation, pp. 106–114. IEEE Computer Society, Los Alamitos (2001)
6. Gimblett, R.: Integrating Geographic Information Systems and Agent-Based Modeling Techniques for Simulating Social and Ecological Processes. Oxford University Press, Oxford (2002)
7. Mandl, P.: Fuzzy-System-Umgebungen als regelgesteuerte Simulationsmaschinen für Geographische Informationssysteme (1996), http://www.uni-klu.ac.at/groups/geo/gismosim/paper/mandl/mandl.htm
8. Koch, A.: Linking multi agent systems and GIS, modeling and simulating spatial interactions. In: 29th International Geographical Congress, Seoul, South Korea (2000)
9. Westervelt, D.: Geographic Information Systems and Agent-Based Modeling. In: Gimblett, H.R. (ed.) Integrating Geographic Information Systems and Agent-Based Modeling Techniques, Oxford University Press, New York (2002)
10. Heijst, G., Schreiber, A., Wielinga, B.: Using Explicit Ontologies in KBS Development. International Journal of Human-Computer Studies/Knowledge Acquisition (Fall, 1996)
11. Urbani, D.: Development of a new MAS GIS hybrid approach for Decision Support Systems; application to water management, Ph. D. These, University of Corsica (2006)
12. Oussalah, C.: Modèles Hiérarchisés Multi Vues pour le Support de Raisonnement dans les Domaines Techniques, Rapport technique, Université d'Aix-Marseille (1988)
13. Bousquet, F., Bakam, I., Proton, H., Le Page, C.: Cormas: common-pool resources and multi-agent Systems. In: Mira, J., Moonis, A., de Pobil, A.P. (eds.) IEA/AIE 1998. LNCS, vol. 1416, pp. 826–838. Springer, Heidelberg (1998)
14. CIRAD, French agricultural research centre working for international development, description http://www.cirad.fr/indexeng.htm
15. ArcGIS description, http://www.esri.com/
16. Urbani, D., Delhom, M.: Water Management Using a New Hybrid Multi-Agents System – Geographic Information System Decision Support System Framework. In: IEEE International Symposium on Environment Identities and Mediterranean area (July 2006)
17. Urbani, D., Delhom, M.: Water Management Policy Selection Using a Decision Support System Based on a Multi-agent System. In: Proceedings 9th Congress of the Italian, Association for Artificial Intelligence, Milan, Italy, September 21-23 (2005), Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)

Multi-agent System for Custom Relationship Management with SVMs Tool

Yanshan Xiao[1], Bo Liu[2, 3], Dan Luo[2], and Longbing Cao[2]

[1] Guangzhou Asian Games Organizing Committee, Guangzhou 510623, P.R. China
[2] Faculty of Information Technology, University of technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
[3] College of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, P.R. China
xiaoyanshan@gmail.com; csbliu@gmail.com; dluo@it.uts.edu.au; lbcao@it.uts.edu.au

Abstract. Distributed data mining in the CRM is to learn available knowledge from the customer relationship so as to instruct the strategic behavior. In order to resolve the CRM in distributed data mining, this paper proposes the architecture of distributed data mining for CRM, and then utilizes the support vector machine tool to separate the customs into several classes and manage them. In the end, the practical experiments about one Chinese company are conducted to show the good performance of the proposed approach.

1 Introduction

Distributed Data Mining (DDM) aims at extracting useful information from distributed heterogeneous data bases. [1] A lot of modern applications fall into the category of systems that need DDM supporting distributed decision making. Applications can be of different natures and from different scopes [2]. Customer relationship management (CRM) has gained its importance role in the marketing decisions strategies [3], which aiming at learning available knowledge from the customer relationship by machine learning or statistical method so as to instruct the strategic behavior.

In the methods for CRM, machine learning is a power tool for this problem [4], and was widely used in the corresponding fields. Network [5] as a popular method in the machine learning which based on the experiential risk minimizations (ERM) is a popular algorithm for CRM. However, some drawbacks exist in the algorithm, as the size of training set increase, training time become sharply long [6]. In addition, because network is based on ERM, the performance of the approach is poor [7]. In the recent years, support vector machines (SVMs) have been introduced for solving pattern recognition problems because of their superior performance [8]. The SVMs are developed based on the idea of structural risk minimizations (SRM), which guarantee the good general performance of the method. In the approach, one need map the original data into the feature space and one constructs an optimal separating hyperplane with maximal margin in this space.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 333–340, 2008.

In order to resolve the CRM in distributed data mining for CRM, this paper proposes the architecture of distributed data mining for CRM and utilizes the support vector machine tool to separate the customs into several classes. In the end, the practical experiments are conducted to verity the performance of the approach. The rest of the paper is organized as follows. Section 2 gives the architecture of distributed data mining for CRM, the support vector machine tool is reviewed in Section 3, Section 4 presents the practical experiments and the conclusions are given in Section 5.

2 Architecture of Distributed Data Mining for CRM

We utilize the multi-agent system (MAS) architectures for the CRM in DDM and distributed classification (DC) systems, which is demonstrated in Fig. 1. The main functions of them are as follows:

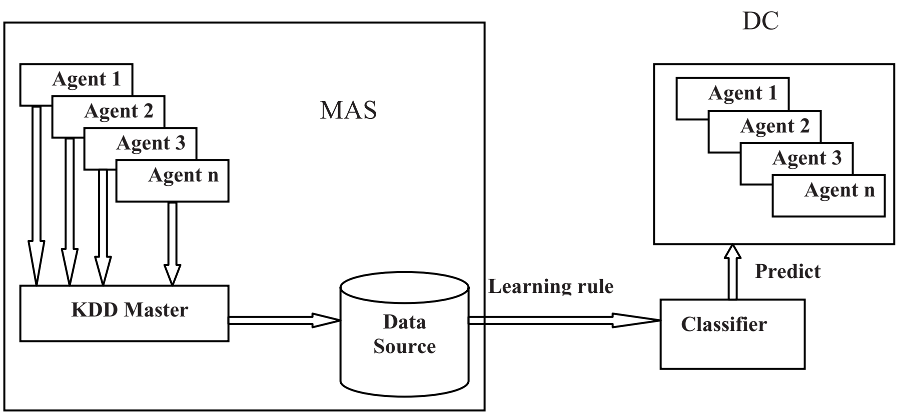


Fig. 1. Architecture of Distributed Data Mining for CRM

1 KDD Master: collect the samples from the agents, and determine what kind the examples should be stored and filtrate the noise samples.

2 Data source (1): Participates in the distributed design of the consistent shared by DDM and DC MAS components of the application ontology; (2): Collaborates with machine learning tools in the training and testing procedures; (3): Provides gateway to databases through performing transformation of queries from the language used inontology into SQL language.

3 Learning rule: through the machine learning tools to obtain the rules of the examples and form the classifier for future predict.

4 Classifier: this the core of the DC, which is used to predict the examples from the agent and give the corresponding labels for decision making.

In this paper, we will focus on how to determine the classifier by utilizing the machine learning tools. After analyze the data in the data source, we can split them into the training dataset and testing dataset. The train dataset is used to extract the rules and form the classifier, the testing dataset is utilized to verity the performance of the classifier.

Because support vector machines (SVMs) have been introduced for solving pattern recognition problems because of their superior performance, which is developed based on the idea of structural risk minimizations (SRM), we will adopt this technology to form the classifier.

3 Support Vector Machine Tool

Let $S=\{(\mathbf{x}_1, y_1),(\mathbf{x}_2, y_2),\dots,(\mathbf{x}_l, y_l)\}$ be a training set, where $\mathbf{x}_i$ are m-dimensional attribute vectors, $y_i \in \{+1,-1\}$. The SVMs classifier is defined as follows:

$$D(\mathbf{x}) = w^T \Phi(\mathbf{x}) + b = 0. \tag{1}$$

where $\Phi(\mathbf{x})$ is a mapping function, w^T is an m-dimensional vector and b is a scalar.

To separate the data linearly separable in the feature space, the decision function satisfies the following condition:

$$y_i(w^T \Phi(\mathbf{x}_i) + b) \geq 1 \qquad for \quad i = 1,\cdots,l. \tag{2}$$

Among all the separating hyperplanes, the optimal separating hyperplane which has the maximal margin between two classes can be formed as follows:

$$\min_{w,b} J(w,b) = \frac{1}{2} w^T w \tag{3}$$

subject to (2). If the training data are nonlinearly separable, slack variables ξ_i is introduced into (2) to relax the hard margin constraints as follows:

$$y_i(w^T \Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \qquad for \quad i = 1,\cdots,l, \tag{4}$$

$$\xi_i \geq 0 \qquad for \quad i = 1,\cdots,l. \tag{5}$$

This technique allows the possibility of having examples that violate (2). In order to obtain the optimal separating hyperplane, we should minimize

$$\min_{w,b,\xi} J(w,b,\xi_i) = \frac{1}{2} w^T w + C \frac{1}{2} \sum_{i=1}^{l} \xi_i \tag{6}$$

subject to (4) and (5), where C is a parameter which determines the tradeoff between the maximum margin and the minimum classification error.

The optimization problem of (6) is a convex quadratic program which can be solved using Lagrange multiplier method. By introducing Lagrange multipliers α_i and $\beta_i\,(i=1,2,\cdots,l)$, one can construct the Lagrangian function as follows:

$$L(w,b,\xi_i,\alpha_i,\beta_i)=J(w,b,\xi_i)-\sum_{i=1}^{l}\alpha_i\{y_i[w^T\Phi(\mathbf{x}_i)+b]-1+\xi_i\} \\ -\sum_{i=1}^{l}\beta_i\xi_i. \tag{7}$$

According to the Kuhn-Tucker theorem, the solution of the optimization problem is given by the saddle point of the Lagrangian function and can be shown to have an expansion:

$$w=\sum_{i=1}^{l}\alpha_i y_i\Phi(\mathbf{x}_i), \tag{8}$$

the training examples $(\mathbf{x}_i, y_i)$ with nonzero coefficients α_i are called support vectors. The coefficients α_i can be found by solving the following problem:

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j (\Phi(\mathbf{x}_i)^T\cdot\Phi(\mathbf{x}_j))\alpha_i\alpha_j+\sum_{i=1}^{l}\alpha_i, \tag{9}$$

subject to

$$\sum_{i=1}^{l}\alpha_i y_i=0, \qquad i=1,2,\cdots,l. \tag{10}$$

$$0\le\alpha_i\le C, \qquad i=1,2,\cdots,l. \tag{11}$$

By substituting (8) into (1), the classifier can be obtained. Given a new input $\mathbf{x}$, $f(\mathbf{x})$ can be estimated using (12). If $f(\mathbf{x})>0$, the sample belongs to Class 1; on the contrary, it belongs to Class 2.

$$f(\mathbf{x})=\mathrm{sgn}\{\sum_{i=1}^{l}\alpha_i y_i\cdot(\Phi(\mathbf{x}_i)^T\cdot\Phi(\mathbf{x})+b\}. \tag{12}$$

where

$$\mathrm{sgn}(x)=\begin{cases}1 & x>0\\ 0 & x\le 0\end{cases}. \tag{13}$$

In the (12), the pairwise inner product in the feature space can be computed from the original data items using a kernel function, and the kernel function can be denoted by

$$K(\mathbf{x},\mathbf{x}_i)=\Phi(\mathbf{x})^T\cdot\Phi(\mathbf{x}_i) \tag{14}$$

In this way, $f(\mathbf{x})$ can be rewritten as follows:

$$f(\mathbf{x})=\mathrm{sgn}\{\sum_{i=1}^{l}\alpha_i y_i\cdot K(\mathbf{x}_i,\mathbf{x})+b\}. \tag{15}$$

As for multi-class classification, they are usually converted into binary ones, One-against-All and One-against-One schemes [8, 9] are preferred methods.

4 Practical Experiments about Chinese Company

For the one company in china, we collect the data samples from all its branches all over the china, which can be considered as the agents of the system, and then the samples are stored in the data source. All the dataset consists of 26 data points with 4 features and 2 classes, which are listed in the appendix. The attributes of the dataset is as follows: the age and education of the employee, the production level and the distribution area. In order to obtain the classifier to predict in the future, we select randomly 16 samples to form the training set and the remaining as the testing dataset.

The experiments are run on a PC with a 2.8GHz Pentium IV processor and a maximum of 512MB memory. The program is written in C++, using Microsoft's Visual C++ 6.0 compiler. The support vector machine tool is used to build the classifier and an RBF kernel function is employed as follows:

$$K(\mathbf{x},\mathbf{x}_i)=\exp(-\|\mathbf{x}-\mathbf{x}_i\|_2^2/2\sigma^2) \tag{16}$$

For the hyperparameters, we estimate the generalized accuracy using different kernel parameters σ and cost parameters C : $\sigma=[2^{-3},2^{-3},2^{-2},\cdots,2^{6}]$ and $C=[2^{2},2^{3},2^{4},\cdots,2^{11}]$. We have to try 100 combinations to select the better hyper-parameters. The results of computation are listed in the Table1, which involves ten figures. Each figure describes the training and testing accuracies of the given problem when fixing the parameter C and let the parameter σ vary in the range of $[2^{-3},2^{-3},2^{-2},\cdots,2^{6}]$.

In general, the performance of the algorithm with SVMs is sensitive to the hyper-parameter, and the testing accuracy represents the predictive ability. Similar to the common method in SVMs, we report the highest testing accuracy and the hyper-parameters in the Table 2.

Table 1. Results of experiments

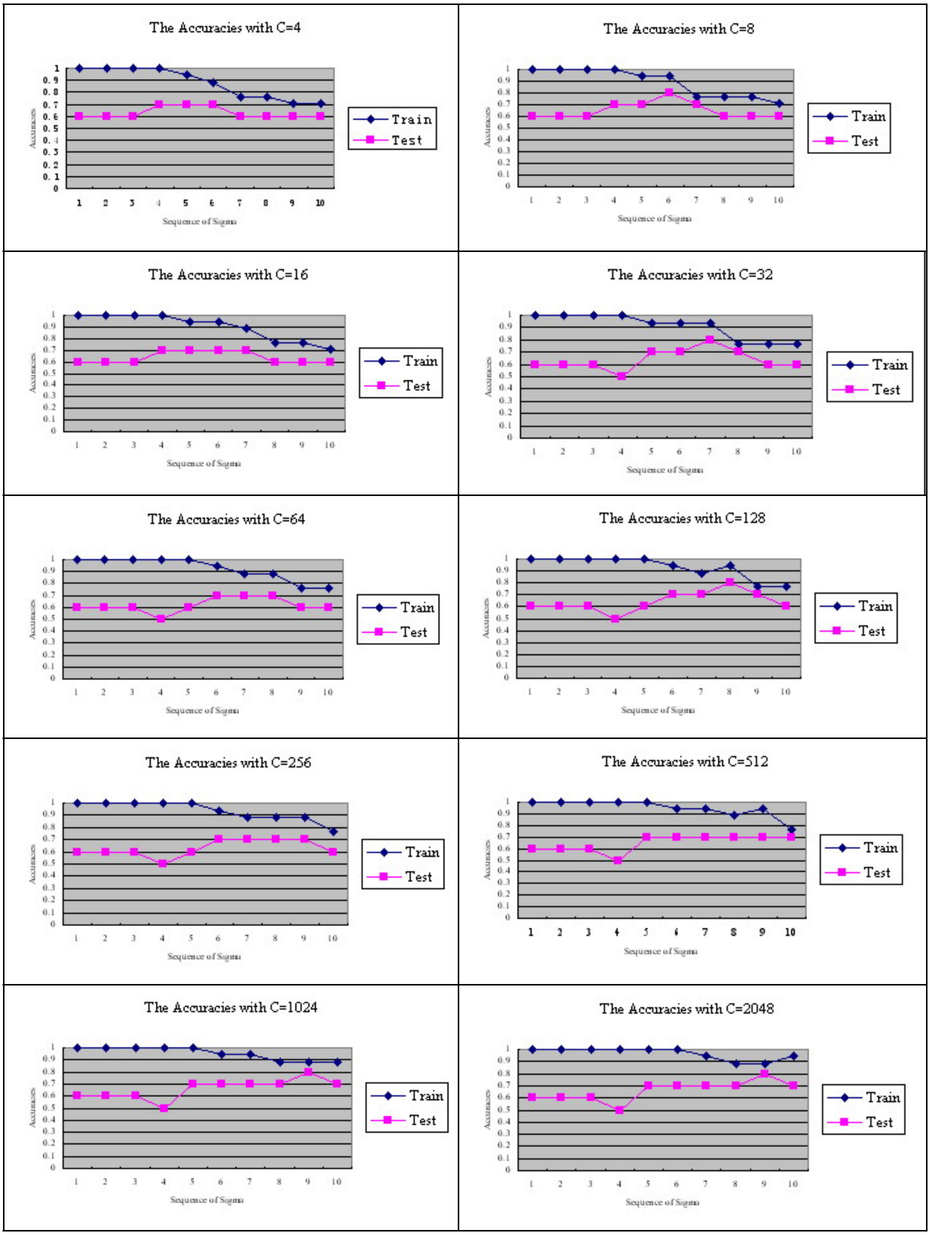


Table 2. The highest testing accuracy and corresponding hyperparameters

Parameters (C,σ)	Training Accuracy (%)	Testing Accuracy (%)
(128,16)	93.75	80
(8,4)	93.75	80
(32,8)	93.75	80
(1024,32)	93.75	80
(2048,32)	93.75	80

From the results above, we can easily conclude that the training and testing accuracies of the classifier with SVMs is very high and SVMs could resolve the customer relationship management. Because SVMs is aiming at the machine learning of little sample, it could be successfully used in many practical problems.

5 Conclusions

Distributed data mining in the CRM is to learn available knowledge from the customer relationship so as to instruct the strategic behavior and obtain the most profit. In order to resolve the CRM in distributed data mining for CRM, this paper presents the architecture of distributed data mining, and then utilizes the SVMs tool to separate the customs into several classes and manage them. In the end, the practical experiments about one Chinese company are conducted to show the good performance of the proposed approach.

References

[1] Prodromidis, A., Chan, P., Stolfo, S.: Meta-learning in distributed data mining systems: Issues and approaches. Advance in Distributed Data Mining. AAAI Press, Menlo Park (1999)
[2] Provost, F., Hennessy, D.: Scaling up: Distributed machine learning with cooperation. Working Notes of IMLM-96, pp. 107–112 (1996)
[3] Blattberg, R.C., Deighton, J.: Manage marketing by the customer equity test, pp. 136–144. Harvard Business Review (July–August 1996)
[4] He, Z.Y., et al.: Mining class outliers: concepts, algorithms and applications in CRM. Expert Systems with Application 27, 681–697 (2004)
[5] Baesens, B., et al.: Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operational Research 156, 508–523 (2004)
[6] Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI), San Jose, pp. 223–228. AAAI Press, Menlo Park (1992)
[7] Heckerman, D.: A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research (1995)
[8] Vapnik, V.N.: Statistical learning theory. John Wiley & Sons, Chichester (1998)

[9] KreBel, U.H.G.: Pairwise classification and support vector machines. In: Schölkopf, B., Burges, C.J., Smola, A.J. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 255–268. The MIT Press, Cambridge (1999)

Appendix

Samples	Age	Education	Production Level	Distribution Area	Classification
1	<=30	High	Low	A	Bad
2	<=30	High	High	A	Good
3	<=30	High	Medium	B	Bad
4	<=30	High	High	B	Good
5	<=30	Low	High	A	Good
6	<=30	Low	Low	A	Good
7	<=30	Low	Low	B	Good
8	<=30	Medium	High	A	Good
9	<=30	Medium	Medium	A	Good
10	<=30	Medium	Medium	B	Good
11	<=30	Medium	Low	A	Good
12	31-50	Medium	Medium	A	Good
13	31-50	Medium	Medium	B	Good
14	31-50	Medium	Low	A	Bad
15	31-50	High	High	A	Good
16	31-50	High	Medium	A	Good
17	31-50	High	Low	A	Good
18	31-50	High	High	B	Bad
19	31-50	High	Low	B	Bad
20	31-50	Low	High	A	Good
21	31-50	Low	Low	A	Good
22	31-50	Medium	High	B	Bad
23	31-50	Medium	High	A	Good
24	>=50	Medium	High	A	Bad
25	>=50	Medium	High	B	Bad
26	>=50	Medium	Medium	A	Good

A Model for Fuzzy Grounding of Modal Conjunctions in Artificial Cognitive Agents

Radosław Piotr Katarzyniak[1], Ngoc Thanh Nguyen[1], and Lakhmi C. Jain[2]

[1] Institute of Information Science and Engineering, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-350 Wrocław, Poland
{Radosław.Katarzyniak, Thanh}@pwr.wroc.pl
[2] School of Electrical and Information Engineering, University of South Australia
Lakhmi.Jain@unisa.edu.au

Abstract. An original approach to fuzzification of grounding sets is introduced. This model is considered for the case of grounding knowledge, belief and possibility extensions of conjunctions. Artificial cognitive agents that carry out the grounding are assumed to observe their external worlds and store results of observations in the so-called *base profiles*. Two levels of fuzzification are introduced to an original model for grounding: the first one deals with fuzzification of atom observations in particular base profiles and the second deals with fuzzification of grounding sets by introducing a fuzzy membership of base profiles to grounding sets. Both levels of fuzzification are applied to make the original mechanism of grounding more context sensitive.

1 Introduction

The language grounding problem belongs to basic research issues considered by Artificial Intelligence and Cognitive Sciences [2, 10, 14]. This problem is studied in relation to the concept of communicative cognitive agent due to the fact that all spoken sentences are always uttered by some intelligent agents. These agents are responsible for choosing sentences that suit best to the outside worlds and use them as external language representations of their opinions. The relation between states of an actual world and their partial or complete descriptions is called the *grounding*. This class of relations is always computed by an agent mediating between external environment and the symbols of a language of communication. The agent, the symbol and the state of world constitute the so-called semiotic triangle [1,11,12].

Models for grounding have been proposed to define the grounding relation for various classes of agents and various interpreted communication languages. The size and complexity of these models depend strongly on the internal organization of agents and syntactic and semantic properties of languages to be grounded. Many cases of grounding have already been partly implemented in practical applications [10]. In this paper an original theory of grounding is cited and further extended. This theory has been formulated in [9] for a language of three classes of modal formulas: simple modalities [4], modal conjunctions [5-7] and modal alternatives [9]. Below the case of modal conjunctions is used as an example.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 341–350, 2008.

The artificial cognitive agent is assumed to be located in a world built from objects $\Omega=\{o_1,o_2,...o_N\}$. At each time point $t\in T=\{t_1, t_2, ...\}$ each object of the world can exhibit properties from a set $\Im=\{P_1,P_2,...P_K\}$. The agent observes its outside world and stores results of these observations in the so-called base profiles $BP(t)$, where t is a time point in which a particular observation was completed. The collection $KS(k)$ of all base profiles defines the overall knowledge state developed by the artificial cognitive agent up to the time point k.

In order to communicate its private opinions on current states of the world the artificial cognitive agent is equipped with its possibility to utter three types of modal formulas [9]. The class of formulas considered in this paper consists of non-modal and modal conjunctions defined as follows [5-7]:

$$\phi, Pos(\phi), Bel(\phi), Know(\phi) \tag{1}$$

where:

$$\phi\in \{p(o)\wedge q(o), p(o)\wedge\neg q(o), \neg p(o)\wedge q(o), \neg p(o)\wedge\neg q(o)\} \tag{2}$$

and symbols p and q refer to some properties P, $Q\in\Im=\{P_1,P_2,...P_K\}$, $P\neq Q$, and are interpreted as "Object o exhibits property P" and "Object o exhibits property Q", respectively. The usual commonsense interpretation is assigned to logic symbols used in points (1) and (2). Namely, the modal operators *Pos*, *Bel* and *Know* are interpreted as '*It is possible that*…', '*I believe that*…' and '*I know that*…', the connectives $\neg$ and $\wedge$ are the natural language negation and conjunction, respectively.

The original theory of grounding, built for formulas given in (1), has been already described in papers [5-7] and monograph [9], where the states of cognitive agent's knowledge are defined in which the modal conjunctions are treated as properly connected to (grounded in) the outside world. The main target of this work is to extend the original definition for grounding modal conjunctions by introducing additional fuzzy mechanisms that make the grounding of modal conjunctions more context sensitive.

2 The Non-fuzzy Theory of Grounding Modal Conjunctions

2.1 The Outline of Original Theory

The following general assumptions are made in the original theory of grounding modal conjunctions [5-7]:

At first, each modal conjunction given in point (1) is assigned with its intuitive meaning. For instance, the modal conjunction $Bel(p(o)\wedge q(o))$ is assumed to represent the sentence "I believe that object o exhibits property P and object o exhibits property Q".

At second, in each cognitive agent four dedicated knowledge structures called *mental models* are required to relate signs of modal conjunctions to their meaning. For instance, the sign of modal conjunction $Bel(p(o)\wedge q(o))$ captures the meaning of the sentence "I believe that object o exhibits property P and object o exhibits property Q". However, in the theory of grounding it is assumed that this meaning is accessible for the agent if and only if the so-called mental model for $p(o)\wedge q(o)$ has already been extracted from the agent's knowledge base [3]. This mental model is induced by the so-called grounding set that consists of all stored observations in which both properties P and Q were observed in the object o. This mental model is denoted by m^{C}_{1} and the related grounding set by C^1. Similarly, pairs of concepts (m^{C}_{2}, C^2), (m^{C}_{3}, C^3) and (m^{C}_{4}, C^4) are defined for conjunctions $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$, $\neg p(o)\wedge\neg q(o)$, respectively (see Tab. 1).

Table 1. Conjunctions and related knowledge structures

Conjunction	Intuitive interpretation	Mental model	Grounding set
$p(o)\wedge q(o)$	Object o exhibits property P and exhibits property Q.	m^{C}_{1}	C^1
$p(o)\wedge\neg q(o)$	Object o exhibits property P and does not exhibit property Q.	m^{C}_{2}	C^2
$\neg p(o)\wedge q(o)$	Object o does not exhibit property P and exhibits property Q.	m^{C}_{3}	C^3
$\neg p(o)\wedge\neg q(o)$	Object o does not exhibit property P and does not exhibit property Q.	m^{C}_{4}	C^4

At third, the original theory of grounding assumes that each mental model m^{C}_{i}, i=1,2,3,4 should always be analyzed with strict relation to its complementary counterparts m^{C}_{j}, $j\neq i$, j=1,2,3,4. For instance, m^{C}_{1} assigned to $p(o)\wedge q(o)$ is always strongly related to m^{C}_{2}, m^{C}_{3}, and m^{C}_{4} assigned to $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$ and $\neg p(o)\wedge\neg q(o)$, respectively. This property of cognitive processes is treated as an inherent one for this class of agents.

At fourth, it is assumed that mental models are assigned with their intensity. This intensity is usually different for each mental model due to the fact that cardinalities of related grounding sets are usually not the same. For instance, the richer the collection of observations stored in C^1, the higher the relative intensity assigned to m^{C}_{1} and the lower the intensity assigned to at least some of m^{C}_{2}, m^{C}_{3} or m^{C}_{4}.

At fifth, the theory of grounding assumes that relative intensity of conjunctions $p(o)\wedge q(o)$, $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$, $\neg p(o)\wedge\neg q(o)$, given by relative strength of related mental models, determines these modal operators which should be used to extend an non-extended in order to connect it properly to the outside world. A dedicated system consisting of the so called modality thresholds is defined for each cognitive agent to capture this additional cognitive concept.

Detailed definitions for the above introduced concepts as well as for their fuzzified versions are given below.

2.2 A Definition for Non-fuzzy Grounding of Modal Conjunction

As it has already been partly mentioned in [3], in order to model the phenomenon of non-fuzzy grounding of modal conjunctions the following concepts have to be introduced: the state of agent's (empirical) knowledge, the grounding sets, the relative strength of grounding sets, the system of modality thresholds, the distribution of (empirical) knowledge and the grounding relation.

The state of agent's (empirical) knowledge. The state of agent's empirical knowledge $KS(t)$ is defined by overall empirical material that has ever been collected by an artificial cognitive agent in its internal databases. At each time point $t \in T$ this state is given as a set of all base profiles stored up to the time point t [9]:

$$KS(t) = \{BP(l): l \in T \text{ and } l \leq^{TM} t\}. \quad (3)$$

where:

1. $BP(l)$ denotes this particular base profile that is related to time points l,
2. $l \leq^{TM} t$ denotes temporal precedence**,**
3. the structure of base profile is given as:

$$BP(l) = <\Omega, P^+_1(l), P^-_1(l), P^+_2(l), P^-_2(l), \ldots P^+_K(l), P^-_K(l) > \quad (4)$$

with the following constraints and interpretations:

a) l denotes a time point to which the observations captured by l-related base profile is related,
b) for i=1,2,…, K the conditions $P^+_i(l) \subseteq \Omega$ and $P^-_i(t) \subseteq \Omega$ hold,
c) for i=1,2,…, K and for $o \in \Omega$ condition $o \in P^+_i(l)$ holds if and only if the agent observed o as exhibiting property P_i at the time point l,
d) for i=1,2,…, K and for $o \in \Omega$ condition $o \in P^-_i(t)$ holds if and only if the agent observed o as non-exhibiting property P_i at the time point l.

The grounding sets. The grounding sets for modal conjunctions are these subsets of $KS(t)$ which contribute to the subjectively induced interpretations of conjunctions used by a particular artificial agent. The non-fuzzy grounding sets for possible binary conjunctions $p(o) \wedge q(o)$, $p(o) \wedge \neg q(o)$, $\neg p(o) \wedge q(o)$, $\neg p(o) \wedge \neg q(o)$ are as follows:

$$C^1(t) = \{BP(l): l \leq^{TM} t \text{ and } BP(l) \in KS(t) \text{ and } o \in P^+(l) \text{ and } o \in Q^+(l)\}, \quad (5)$$

$$C^2(t) = \{BP(l): l \leq^{TM} t \text{ and } BP(l) \in KS(t) \text{ and } o \in P^+(l) \text{ and } o \in Q^-(l)\}, \quad (6)$$

$$C^3(t) = \{BP(l): l \leq^{TM} t \text{ and } BP(l) \in KS(t) \text{ and } o \in P^-(l) \text{ and } o \in Q^+(l)\}, \quad (7)$$

$$C^4(t) = \{BP(l): l \leq^{TM} t \text{ and } BP(l) \in KS(t) \text{ and } o \in P^-(l) \text{ and } o \in Q^-(l)\}, \quad (8)$$

where $P, Q \in \Im$, $P \neq Q$.

Sets $C^i(t)$ are always related to a time point t. It means that their content evolves over the time depending on changes in the world observed by the agent. Below abbreviated symbols C^i, i=1,2,3,4 will be used instead of $C^i(t)$ i=1,2,3,4. As it has already been mentioned for each i=1,2,3,4 the set C^i is this collection of empirical experiences from which the related mental model $m^C{}_i$ is extracted to represent the grounded meaning of conjunction [3,9].

The relative strength of grounding sets. The original theory of grounding [3,5-7,9] assumes that the strength with which particular modal extensions of conjunctions are connected to external worlds is strongly shaped by differences in the content of their grounding sets. For instance the richer the set C^1 comparing to the remaining sets C^2,

C^3 and C^4, the stronger the intention of artificial cognitive agent to use $Bel(p(o)\wedge q(o))$ instead of $Pos(p(o)\wedge q(o))$. This concept is supported by a measure called the relative strength of grounding. Let for each i=1,2,3,4 the following symbols be introduced:

$$G^{C}{}_{i} \stackrel{def}{=} card(C^i). \tag{9}$$

For each C^i, i=1,2,3,4 their relative strength of grounding is given as follows:

$$\lambda(t, p(o)\wedge q(o)) = \frac{G_1^C}{G_1^C + G_2^C + G_3^C + G_4^C}\ , \tag{10}$$

$$\lambda(t, , p(o)\wedge\neg q(o)) = \frac{G_2^C}{G_1^C + G_2^C + G_3^C + G_4^C}\ . \tag{11}$$

$$\lambda(t, \neg p(o)\wedge q(o)) = \frac{G_3^C}{G_1^C + G_2^C + G_3^C + G_4^C}\ . \tag{12}$$

$$\lambda(t, \neg p(o)\wedge\neg q(o)) = \frac{G_4^C}{G_1^C + G_2^C + G_3^C + G_4^C}\ . \tag{13}$$

The system of modality thresholds. This concept is introduced into the theory to model additional and necessary constraints applied to grounding of knowledge, belief and possibility extensions of $p(o)\wedge q(o)$, $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$ and $\neg p(o)\wedge\neg q(o)$ in cognitive agents. Let such system be given as tuple $(\lambda^{K}{}_{minPos}, \lambda^{K}{}_{maxPos}, \lambda^{K}{}_{minBel}, \lambda^{K}{}_{maxBel})$ with an additional general requirement:

$$0 < \lambda^{K}{}_{minPos} < \lambda^{K}{}_{maxPos} \leq \lambda^{K}{}_{minBel} < \lambda^{K}{}_{maxBel} \leq 1. \tag{14}$$

The role of modality thresholds is fundamental to the theory of grounding modal conjunctions. For instance, if the condition of $\lambda^{K}{}_{minPos} < \lambda(t, \neg p(o)\wedge q(o)) < \lambda^{K}{}_{maxPos}$ is not fulfilled, then the cognitive agent is not allowed to extend non-modal conjunction $\neg p(o)\wedge q(o)$ with modal operator *Pos*. Similar conditions are defined for knowledge and belief operators.

In [4-7,9] *an original list of additional basic requirements was introduced and studied to constraint the system of modality thresholds and make the resulting implementations of grounding consistent with related natural language behavior.* Multiple theorems were formulated and proved to capture basic properties of these requirements.

The distribution of empirical knowledge. This concept is another important element of the theory. It is used to reflect the natural way the stored empirical material contributes to grounding. This concept relates directly to the division of agent's memory into short and long term memory or equivalently into conscious and non-conscious levels of knowledge processing. The general assumption is that at each time point t the state of knowledge processing is defined by both the overall content of $KS(t)$ and current distribution of this content in short and long term memory [3,9]. The following commonsense interpretation of long and short term cognitive subspaces is assumed in the original theory. The working memory $CM(t)$ consists of these pieces of knowledge which represent the currently processed meaning and refer the agent to all empirical material related to this meaning. The remaining part of this material is always located in the long term memory denoted by $NM(t)$. Such time point - related distribution should be always taken into account when grounding sets C^i, i=1,2,3,4 are used. Let the following symbols be introduced:

$$DC(t)=\{RC^1(t), TC^1(t), RC^2(t), TC^2(t), RC^3(t), TC^3(t), RC^4(t), TC^4(t)\}, \tag{15}$$

where for each i=1,2,3,4 the following requirements are fulfilled:

$$RC^i(t) = CM(t) \cap C^i(t), \tag{16}$$

$$TC^i(t) = NM(t) \cap C^i(t), \tag{17}$$

$$RC^i(t) \cap TC^i(t) = \emptyset, \tag{18}$$

$$RC^i(t) \cup TC^i(t) = C^i(t). \tag{19}$$

This distribution is used in detailed definition of conditions that make particular modal conjunctions properly grounded. This definition is given below.

The relation of grounding. Let a particular time point $t\in T$, a state of knowledge described by distribution $KS(t)=DC(t)=\{RC^1(t), TC^1(t), RC^2(t), TC^2(t), RC^3(t), TC^3(t), RC^4(t), TC^4(t)\}$ and a system of modality thresholds $0 < \lambda^K{}_{minPos} < \lambda^K{}_{maxPos} < \lambda^K{}_{minBel} < \lambda^K{}_{maxBel} <1$ be given. For each P, $Q\in\{P_1,...,P_K\}$ such that $P\neq Q$, and for each object o the following relations of grounding is defined [5-7,9]:

- $KS(t)|=_G Pos(p(o)\wedge q(o))$ holds if and only if conditions (20) hold:

$$o\in O\backslash(P^+(t)\cup P^-(t)) \wedge o\in O\backslash(Q^+(t)\cup Q^-(t)) \wedge RC^1(t)\neq\emptyset \wedge \lambda_{minPos}\leq\lambda(t,p(o)\wedge q(o))\leq\lambda_{maxPos}. \tag{20}$$

- $KS(t)|=_G Bel(p(o)\wedge q(o))$ holds if and only if conditions (21) hold:

$$o\in O\backslash(P^+(t)\cup P^-(t)) \wedge o\in O\backslash(Q^+(t)\cup Q^-(t)) \wedge RC^1(t)\neq\emptyset \wedge \lambda_{minBel}\leq\lambda(t,p(o)\wedge q(o))\leq\lambda_{maxBel} \tag{21}$$

- $KS(t)|=_G Know(p(o)\wedge q(o))$ and $KS(t)|=_G p(o)\wedge q(o)$ hold if and only if conditions (22) or (23) hold:

$$o\in P^+(t) \wedge o\in Q^+(t) \tag{22}$$

$$o\in O\backslash(P^+(t)\cup P^-(t)) \wedge o\in O\backslash(Q^+(t)\cup Q^-(t)) \wedge RC^1(t)\neq\emptyset \wedge \lambda(t,p(o)\wedge q(o))=1 \tag{23}$$

The epistemic satisfaction relation for the remaining three conjunctions is defined in the similar way, provided that appropriate grounding sets are used.

This definition for grounding can be used to design a particular agent-encapsulated cognitive mechanism for relating modal conjunctions to the world. The major property of this mechanism is that it can be used to produce language behaviour consistent with the way all natural language interpretations of modal conjunctions are processed. This issue was discussed in [5-7,9]: where it is presented as a detailed system of multiple theorems. However, this model for grounding should be called context-independent due *o* the fact that such grounding is supported by all stored experiences without considering any influence the current situation can have on the role of past experiences. In [8] four general strategies were proposed to introduce such additional context oriented dimension to the original mechanism of grounding. Below one of these strategies is further developed and captured as a fuzzified model of grounding.

3 Context Sensitive Fuzzification of Grounding

3.1 Decomposition of Base Profiles

It is assumed in this approach that the fuzzification of grounding sets can result from two facts.

At first, each base profile stored in the agent's knowledge base can be naturally decomposed into a related set of atom observations of the type "$o \in P^{+}$" and "$o \in P^{-}$", $P \in \Im$.

At second, for each individual atom observation "$o \in P^{+}$" (or "$o \in P^{-}$") located in a particular base profile $BP(t)$, its contribution to the overall content of $BP(t)$ depends on all occurrences of "$o \in P^{+}$" ("$o \in P^{-}$") in remaining base profiles from the knowledge base. Let the following symbols be introduced for sets of all possible atom observations:

The set $E^{+}(t)$ consists of all *positive atomic observations* that are captured by a particular base profile $BP(t)$:

$$E^{+}(t) =^{def} \{\text{"}o \in P^{+}\text{"}: o \in O, P \in \Im \text{ and } o \in P^{+}(t)\} \tag{24}$$

The set $E^{-}(t)$ consists of all *negative atomic observations* that are captured by a particular base profile $BP(t)$:

$$E^{-}(t) =^{def} \{\text{"}o \in P^{-}\text{"}: o \in O, P \in \Im \text{ and } o \in P^{-}(t)\} \tag{25}$$

Obviously, for each base profile $BP(t)$ there are some virtual atom observations that are not included to $E^{+}(t) \cup E^{-}(t)$. The mechanism of base profiles' decomposition makes it possible to treat each base profile as a set consisting of related atom observation. Therefore each base profile can be treated as fuzzy set, provided that related atom observations are treated as its elements. An example of membership function of atom observations in particular base profiles is given below.

3.2 The Subjective Importance of Atom Observations in Stored Base Profiles

Let us start with the case of positive atom observations. Let a certain distribution of positive atom observations "$o \in P^{+}$" in grounding sets C^{i}, i=1,2,3,4 be given, for each i=1,2,3,4 let $N_{i} =^{def} card(C^{i})$ and let n_i denote the number of all base profiles with the atom observation "$o \in P^{+}$" belonging to C^{i}. Let $w(\text{"}o \in P^{+}\text{"} | C^{i})$ denote the fuzzy membership value assigned to each atom observation "$o \in P^{+}$" stored in at least one base profile in all C^{i}, i=1,2,3,4. The following requirements are assumed to constrain the way the actual contribution of each atom observation to a particular base profile from C^{i}, i=1,2,3,4 should be computed:

$$\text{If } \frac{n_i}{card(C^i)} \text{ increases then } w(\text{“}o\in P^{+}\text{“} \mid C^i) \text{ increases.} \tag{26}$$

$$\text{If } \frac{n_i}{\sum\limits_{\substack{k\neq i \\ k\in\{1,2,3,4\}}} n_k} \text{ increases then } w(\text{“}o\in P^{+}\text{“} \mid C^i) \text{ increases.} \tag{27}$$

$$\text{If } n_{i=}0 \text{ then } w(\text{“}o\in P^{+}\text{“} \mid C^i) = 0. \tag{28}$$

$$\text{If } \frac{n_i}{card(C^i)} = \frac{n_k}{card(C^k)} \text{ then } w(\text{“}o\in P^{+}\text{“} \mid C^i) = w(\text{“}o\in P^{+}\text{“} \mid C^k). \tag{29}$$

It can be easily proved that the following measure fulfills all requirements (26-29):

$$w(\text{“}o\in P^{+}\text{“} \mid C^i) = \frac{n_i^2}{N_i \cdot \sum\limits_{k=1}^{4} n_k}. \tag{30}$$

Similar definitions can be formulated for the case of negative atom observations.

In consequence for a particular base profile $BP(t)\in C^i$ its decomposition into a fuzzy set *Fuzzy-BP*(t) can be defined as follows:

$$Fuzzy\text{-}BP(t)=\{(e,w(e \mid C^i)): e\in E^{+}(t)\cup E^{-}(t)\}. \tag{31}$$

3.2 The Subjective Similarity of Stored Perceptions to Current Context

The above fuzzification of base profiles resulted from internal distribution of particular atom observations over the family of sets C^i, i=1,2,3,4.

The second level of fuzzification is however introduced to capture additional influence the observed context could have on the actual role of particular grounding sets. This step of fuzzification is based on the idea of similarity measure defined to capture differences between the latest (current) observation $PB(t_{now})$ and other base profiles stored in grounding sets to be fuzzified. The way this measure can be computed is based on an idea of atom similarities *asim* defined for triples consisting of the latest observation (current base profile) $BP(t_{now})$, a previous base profile $PB(t)$ and a particular atom observation $e\in E^{+}(t_{now})\cup E^{-}(t_{now})$.

Let us consider a particular set C^i, i=1,2,3,4 and an atom observation “$o\in P^{+}$”$\in E^{+}(t_{now})$. The following symbols are introduced:

$n^{+}_{i=}$ the number of base profiles in C^i in which the atom observation “$o\in P^{+}$” is captured,

$n^{-}_{i=}$ the number of base profiles in C^i in which the atom observation “$o\in P^{-}$” is captured.

Values of atom similarities can be computed on the base of the following formula:

$$asim(BP(t_{\text{now}}), BP(t), \text{o}) = \tag{32}$$

$$= \begin{cases} w("o \in P^{+}" | C^{i}) & iff \quad o \in P^{+}(t_{now}) \, and \, o \in P^{+}(t) \\ \dfrac{n^{+}_{i} \cdot w("o \in P^{+}" | C^{i})}{n^{+}_{i} + n^{-}_{i}} & iff \quad o \in P^{+}(t_{now}) \, and \, o \in P^{\pm}(t) \\ w("o \in P^{-}" | C^{i}) & iff \quad o \in P^{-}(t_{now}) \, and \, o \in P^{-}(t) \\ \dfrac{n^{-}_{i} \cdot w("o \in P^{-}" | C^{i})}{n^{+}_{i} + n^{-}_{i}} & iff \quad o \in P^{-}(t_{now}) \, and \, o \in P^{\pm}(t) \\ 0 & otherwise \end{cases}$$

The commonsense premises underlying the above definition of atom similarities are rather obvious. The result is that a dedicated algorithm based on the following definitions (33-34) can be developed to yield the related fuzzy versions of each grounding set C^{i}, i=1,2,3,4:

$$sim(BP(t_{\text{now}}), BP(t)) = \sum_{o \in \Omega} asim(BP(t_{\text{now}}), BP(t), \text{o}) \tag{33}$$

$$Fuzz\text{-}C^{i} = \{ \, (BP(t), sim(BP(t_{\text{now}}), BP(t))) : BP(t) \in C^{i} \, \} \tag{34}$$

3.3 The Strength of Fuzzified Grounding Sets and Fuzzy Grounding

The above fuzzification of grounding sets C^{i}, i=1,2,3,4 results in an additional requirement to apply an appropriate measure for capturing the cardinality of particular grounding material. Multiple approaches are available to realize this target [13]. Let the most commonly used scalar cardinality of fuzzy sets be applied to define the cardinality of $Fuzz\text{-}C^{i}$, i=1,2,3,4:

$$G^{C}{}_{i} \stackrel{\text{def}}{=} \sum_{BP(t) \in C^{i}} sim(BP(t_{\text{now}}), BP(t)) \, . \tag{35}$$

This measure is simple and it is not constrained with any additional requirements. However, in real practical applications it is possible to define additional properties of fuzzy grounding sets which can influence the choice of fuzzy cardinality measure.

The definition of relative grounding strength does not need to be changed and is still given by (10)-(13).

4 Final Remarks

In this paper some ideas generally described in [8] have been further developed. In particular an original fuzzy approach to a contextualization of grounding modal conjunctions has been proposed. The basic assumption underlying this approach was captured within a system of definitions, that the strength of grounding material should be always modified according to a current observable context. In particular, it has been assumed that all stored base profiles (i.e. past observations of the world) which are more similar to the current context should be of higher importance when relevant empirical material is retrieved by grounding processes. This model was defined for the case of modal conjunctions. However, it could also be applied for realizing fuzzification of grounding simple modalities [3,9] and modal alternatives [9], provided that required modifications are introduced.

References

1. Eco, U.: La struttura assente, Tascabili Bompiani, Milano (in Italian) (1991)
2. Harnad, S.: The symbol grounding problem. Physica D 42, 335–346 (1990)
3. Katarzyniak, R.: The language grounding problem and its relation to the internal structure of cognitive agents. Journal of Universal Computer Science 11(2), 357–374 (2005)
4. Katarzyniak, R.: On some properties of grounding simple modalities. Systems Science 31(3), 59–86 (2005)
5. Katarzyniak, R.: On some properties of grounding non-uniform sets of modal conjunctions. Int. Journal of Applied Mathematics and Computer Science 16(3), 399–412 (2006)
6. Katarzyniak, R.: On some properties of grounding uniform sets of modal conjunctions. Journal of Intelligent & Fuzzy Systems 17(3), 209–218 (2006)
7. Katarzyniak, R.: Some notes on grounding singletons of modal conjunctions. Systems Science 32(1), 45–55 (2005)
8. Katarzyniak, R., Nguyen, T.N., Jain, L.C.: Soft Computing Approach to Contextual Determination of Grounding Sets for Simple Modalities. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 230–237. Springer, Heidelberg (2007)
9. Katarzyniak, R.: Grounding modal languages of communication in artificial agents, EXIT Warsaw 2007 (in Polish) (2007)
10. Roy, D., Reiter, E.: Connecting language to the world. Artificial Intelligence 167(1-2), 1–12 (2005)
11. Vogt, P.: Anchoring of semiotic symbols. Robotics and Autonomous Systems 43, 109–120 (2003)
12. Vogt, P.: The Physical Grounding Symbol Problem. Cognitive Systems Research 3, 429–457 (2002)
13. Wygralak, M.: An axiomatic approach to scalar cardinalities of fuzzy sets. Fuzzy Sets and Systems 110, 175–179 (2000)
14. Ziemke, T.: Rethinking Grounding. W. In: Riegler, A., Peschl, M., von Stein, A. (eds.) Understanding Representation in the Cognitive Sciences, pp. 177–190. Plenum Press, New York (1999)

An Integrated Argumentation Environment for Arguing Agents

Taro Fukumoto[1], Syuusuke Kuribara[1], and Hajime Sawamura[2]

[1] Graduate School of Science and Technology, Niigata University
8050, 2-cho, Ikarashi, Nishiku, Niigata, 950-2181 Japan
+81-25-264-7490
{fukumoto, kuribara}@cs.ie.niigata-u.ac.jp
[2] Institute of Natural Science and Technology Academic Assembly, Niigata University
8050, 2-cho, Ikarashi, Nishiku, Niigata, 950-2181 Japan
sawamura@ie.niigata-u.ac.jp

Abstract. This paper provides an Integrated Argumentation Environment (IAE) for arguing agents, the system can inclusively support argumentation like an integrated environment for the software development. IAE yields three types of knowledge base designer, knowledge base editor, argument template and AMLLMA transform module and three types of argumentation visualization tools for analyzing, live argumentation viewer, argumentation and argument tree viewer. Agents can use argument which is written in Argument Markup Language (AML) as own knowledge base via transform module. Additionally IAE can launch argumentation and judge which argument is dominance in an argumentation. After the argumentation, user can also upload the argumentation via transform module to argument DB. IAE will build a bridge between research domain of natural argumentation like a computer supported argumentation visualization and formal argumentation.

1 Introduction

Arguing is a powerful tool we use individually and socially in the daily life. Recently it has been recognized by computer scientists, and in particular AI researchers that argumentation has clear correlation with the interests of the newly emerging field of social computation such as multi-agent systems research. Such arguing agents can apply to various complex problems, hence the application domain of the argument-based agent system is very wide. The implementation and experimental uses over the Internet suggest that argument-based agent system could have a great potential for computer-supported argumentation in pedagogy, future e-government, public policy making, as well as a function for the Web technology such as Web search engines and the Semantic Web [10].

However, unfortunately the way of designing agent's knowledge and the way of analyzing argumentation have been often neglected in the domain of formal argumentation. The design of knowledge base has remained as a big problem yet. On the other hand, in the domain of Computer Supported Argumentation

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 351–360, 2008.

Visualization, there are much researches and products such as Argument Mapping Systems for designing and analyzing argument (e.g., gIBIS[3] Belvedere[6] Araucaria[8] Reason!able[13]). We can design and analyze argument described by natural language through these systems. Although it is natural for us to want to know final conclusion of argument when there are a lot of opinions, these systems do not intend to judge which argument is dominance.

If we can integrate these two systems, argument-based agent system and argument mapping system, we can get great system which can inclusively support argumentation like an integrated environment for the software development. The system not only helps us to design agent's knowledge base, decides which argument is prevail and help to analyze argumentation, but also may help to generate various argument-based agent systems.

For the argument-based agent system, we have already developed a computational argumentation framework that basically consists of EALP and LMA [11]. EALP (Extended Annotated Logic Programming) is an expressive logic programming language we formalized for argumentation, as they are most important and primary actions what we actually do in argumentation. The basic language constituents are literals associated with annotations as truth-values or epistemic states of agents. LMA is a logic of Multiple-valued Argumentation constructed on top of EALP. It has three notions of negation to yield a momentum or driving force for argumentation that allows us to specify various types of truth values depending on application domains, and to deal with uncertain arguments. Such a feature bring us the extensive applicability of LMA that is considered the most advantageous point in comparison to other approaches to argumentation [1] [7].

In this paper, we describe an integrated argumentation environment (IAE) for arguing agents, which supports both humans and agents in participating in argumentation in the Internet environment, designing truth-values for uncertainty designing and editing knowledge bases under the specified truth-values, making and analyzing arguments, lively visualizing arguments and argument flows and so on. It should be noted that our aim at developing an integrated environment proper to argumentation is similar to that of the integrated development environment (IDE) for productive and reliable software development (e. g., eclipse, xcode, etc.). However, there seems to be not yet such an attempt to develop an integrated environment for argumentation framework to allow us to deal with various uncertain arguments under users-specified truth-values. of course, we could see some specific argumentation systems developed so far, each of whose underlying argumentation framework has its own truth-values for uncertainty [7].

This paper is organized as follows. In section 2, we outline our EALP and LMA as a preliminary for the succeeding section. In section 3, we describe the main features of an integrated argumentation environment (IAE) for LMA along the line of our argumentation scenario. In section 4, we discuss about application domain of IAE. The final section summarizes contributions of the paper and discuss some intriguing future plans.

2 Overview of EALP and LMA [11]

Here we introduce some notions of EALP and LMA to make the paper self-contained. Following KB_A and KB_B show an example of EALP with a complete lattice of truth value $\mathcal{T} = \mathcal{R}[0.0,\ 1.0]$, which represents agent's confidence.

Example 1. $KB_A = \{\ agree(GMO):1.0 \leftarrow resoluve(GMO, food\_problem):1.0,$
$resolve(GMO, food\_problem):1.0 \leftarrow productive(GMO):0.8,$
$productive(GMO):0.8 \leftarrow modified(anti\_agrichemical):0.6,$
$\sim contaminate(GMO, land):0.7 \leftarrow restrict(law, GMO):1.0,$
$modified(anti\_agrichemical):1.0 \leftarrow, restrict(law, GMO):1.0 \leftarrow\ \}$
$KB_B = \{\ \sim agree(GMO):1.0 \leftarrow contaminate(GMO, land):0.8,$
$contaminate(GMO, land):0.9 \leftarrow use(agrichemical):0.7,$
$use(agrichemical):0.8 \leftarrow modified(anti\_agrichemical):0.8,$
$\sim agree(GMO):0.7 \leftarrow \mathbf{not}\, safety(GMO, humans):0.9,$
$modified(anti\_agrichemical):0.9 \leftarrow\ \}$

For instance in KB_A,

$$\sim agree(GMO):0.7 \leftarrow \mathbf{not}\, safety(GMO, humans):0.9\ \&\ output\_poison(GMO, pest):0.9$$

is rule of EALP. Each real number on the right side of colon are annotations, and $output\_poison(GMO, pest):0.9$ and all that, are called annotated atoms. $output\_poison(GMO, pest):0.9$ means "GMO outputs poison against pest". The symbol **not** stands for the default negation. Then, annotated atoms with default negation are called annotated default atoms. $\mathbf{not}\, safety(GMO, human):0.9$ means "There is no evidence that GMO is safe for human". The symbol $\sim$ stands for the ontological explicit negation. $\sim agree(GMO):0.7$ means "I carry some hesitation toward agreeing GMO" The rule says "If GMO outputs a poison against the pest and there is no evidence that poison dose not effect to human, then I carry some hesitation toward agreeing GMO".

Definition 1. (**Argument**[11]). *Let P be an EALP. An argument in P is a finite sequence $Arg = [r_1, \ldots, r_n]$ of rules in P such that for every i $(1 \leq i \leq n)$, i) r_i is either a rule in P, ii) For every annotated atom $A:\mu$ in the body of r_i, there exists a r_k $(n \geq k > i)$ such that $A:\rho$ $(\rho \geq \mu)$ is head of r_k. iii) For every ontological explicit negation $\sim A:\mu$ in the body of r_i, there exists a r_k $(n \geq k > i)$ such that $\sim A:\rho$ $(\rho \leq \mu)$ is head of r_k. iv)There exists no proper subsequence of $[r_1, \ldots, r_n]$ which meets from the first to the third condition, and includes r_1.*

In KB_A, we can make an argument $Arg_1 = [\ agree(GMO):1.0 \leftarrow resoluve\ (GMO, food\_problem):1.0,\ resolve(GMO, food\_problem):1.0 \leftarrow productive\ (GMO):0.8,\ productive(GMO):0.8 \leftarrow modified(anti\_agrichemical):0.6,\ modified(anti\_agrichemical):1.0 \leftarrow\ ]$ Arg_1 says, "GMO has a quite highly productive capacity since it is modified anti-agrichemicals, and it can resolve the food problem, so I agree with GMO very confidently". The conclusions of rules in Arg are called conclusions of Arg, and the assumptions of rules in Arg are called assumptions of Arg. We write $concl(Arg)$ for the set of conclusions and $assm(Arg)$ for the set of assumptions of Arg. An attack relation plays a role of a driven force of argumentation or dialogues in general. It is a conflict relation among arguments, and their "*rebut*" and "*undercut*" are typical one so far. In case of multiple-valued argumentation based on EALP, much complication is to be involved into the rebuttal relation under the different concepts of negation [11]. now we present notion of attacks.

Definition 2. (**Rebut**[**11**]). *Arg_α rebuts Arg_β $\Leftrightarrow$ there exists $A\!:\!\mu_1 \in concl(Arg_\alpha)$ and $\sim A : \mu_2 \in concl(Arg_\beta)$ such that $\mu_1 \geq \mu_2$, or exists $\sim A : \mu_1 \in concl(Arg_\alpha)$ and $A\!:\!\mu_2 \in concl(Arg_\beta)$ such that $\mu_1 \leq \mu_2$.*

Definition 3. (**Undercut**[**11**]). *Arg_α undercuts Arg_β $\Leftrightarrow$ there exists $A\!:\!\mu_1 \in concl(Arg_\alpha)$ and* **not** *$A : \mu_2 \in assm(Arg_\beta)$ such that $\mu_1 \geq \mu_2$, or exists $\sim A : \mu_1 \in concl(Arg_\alpha)$ and* **not** *$\sim A\!:\!\mu_2 \in assm(Arg_\beta)$ such that $\mu_1 \leq \mu_2$.*

Definition 4. (**Defeat**[**11**]). *Arg_α defeats Arg_β $\Leftrightarrow$ Arg_1 undercuts Arg_2, or Arg_1 rebuts Arg_2 and Arg_2 does not undercut Arg_1.*

Definition 5. (**Argumentation**[**11**]). *An argumentation is a finite nonempty sequence of moves $move_i = (Player_i, Arg_i)$, $(i \geq 1)$ such that, i) $Player_i = P$ (Proponent) $\Leftrightarrow$ i is odd; and $Player_i = O$ (Opponent) $\Leftrightarrow$ i is even. ii) If $Player_i = Player_j = P$ $(i \neq j)$ then $Arg_i \neq Arg_j$. iii) $(Arg_i, Arg_{i-1}) \in defeat$.*

Justified arguments can be dialectically determined from a set of arguments by the dialectical proof theory.

Definition 6. (**Argumentation tree**[**11**]). *An argumentation tree is a tree of moves such that every branch is an argumentation, and for all moves $move_i = (P, Arg_i)$, the children of $move_i$ are all those moves (O, Arg_{i+1}) such that $(Arg_{i+1}, Arg_i) \in defeat$.*

Definition 7. (**Justified**][**11**]). *An argumentation A is a winning argumentation $\Leftrightarrow$ the termination of A is a move of proponent. An argumentation tree T is a winning argumentation tree $\Leftrightarrow$ every branch of T is a winning argumentation. An argument Arg is a justified argument $\Leftrightarrow$ there exists a winning argumentation tree with Arg as its root.*

For instance, suppose agent society $\mathcal{MAS} = \{KB_A,\ KB_B\}$ as shown in Example 1. Fig. 1 shows all possible arguments in $\mathcal{MAS}$ and attack relations among them. In this example, dotted frame shows argument and solid frame shows justified argument.

3 Features of an Integrated Argumentation Environment

Principal components of integrated argumentation environment are following three points: i)*An editor for knowledge base and truth value* helps user to design

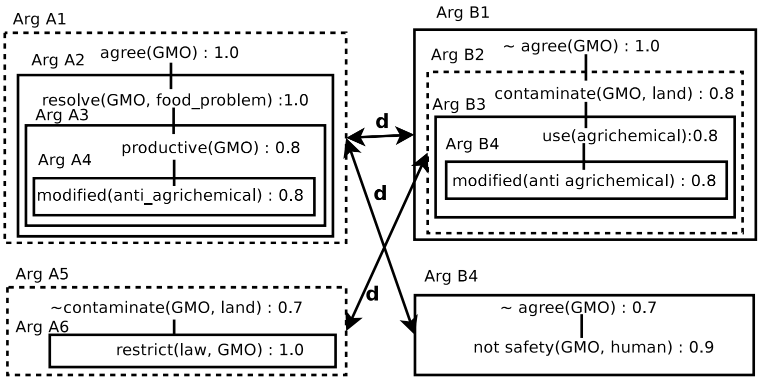


Fig. 1. All argument and attack relation among them

agent's knowledge base. The editor also supports argument template which supports the input by using the structure of the argument like toulmin model and expert opinion scheme. ii)*Live Argumentation Viewer* shows dynamic argumentation. iii)*Argumentation Tree Viewer* generates diagrammatized argument and argumentation automatically and also can convert it to marked up argumentation. Our argumentation scenario consists of the following phases: 1) Registering agents with the argument server so that they can commit to argumentation. 2) Preparing a lattice of truth values for dealing with uncertainty depending on application domains. 3) Designing knowledge bases under the specified truth values in terms of EALP. 4) Starting argumentation on submitted issues/claims in LMA. 5) Analyzing argumentation. 6) Storing arguments and their results in the argument repository for the future reuse. In this section, we briefly describe our system features and how they are related to the argumentation scenario.

3.1 Registration of Agent

The argument server plays a role of a coordinator in argumentation. It manages argument flows, keeping track of dialectical trees and judges a final status of arguments. First of all, user should generate an agent as own avatar.In Fig. 2, right upper small frame shows user agent. User can set agent name and agent image via this frame. Next, user resists own agent to take part in argumentation. In Fig. 2 the Earth labeled button is for connecting the server.

3.2 Preparing a Lattice of Truth Value

We prepared an editor for specifying truth values as a complete lattice. Actually this is a normal editor with which a complete lattice of truth values are stipulated in terms of Prolog. To economy of time and effort to make up and verify the truth value, IAE ships with three types of truth value which are commonly used in multiple-valued argumentation domain, that is, $\mathcal{FOUR}$, real number interval and

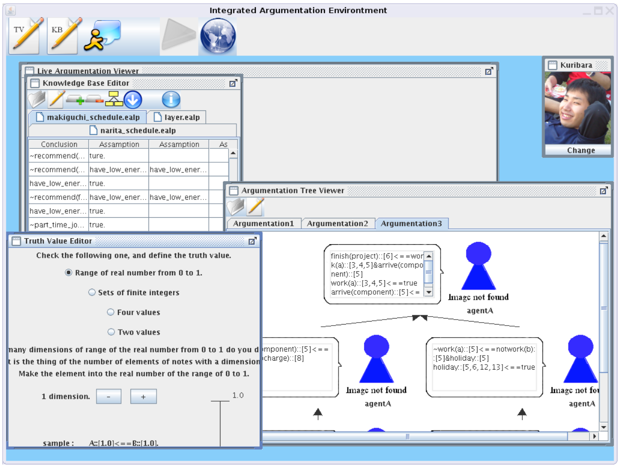


Fig. 2. Screen Shot of IAE

natural number interval, and IAE also supports classical two value. $\mathcal{FOUR} = (\{\mathbf{t}, \mathbf{f}, \top, \bot\}, \leq)$, $\forall x, y \in \{\mathbf{t}, \mathbf{f}, \top, \bot\} x \leq y \Leftrightarrow x = y \vee x = \bot \vee y = \top$ [11]. $\top$ means inconsistency and $\bot$ means undefined respectively. This is convenient for representing inconsistent knowledge, like $agree(nuclear\_power) : \top$ means "both agree and disagree with nuclear power generations". Real number interval is more expressive than $\mathcal{FOUR}$. It can represent a degree of confidence or probability, and that, $R[0,\ 1]^2$ can represent the degree of truth as well as inconsistency and undefined like a fuzzy. $agree(death) : (0.2,\ 0.4)$ means "I disagree with kind of death penalty". Natural number is used to represent time information, such that $\sim open(cafe) : \{4, 11, 18\}$, which means the cafe dose not open on 4th, 11th and 18th in this month. In Fig. 2, left lower frame is Truth value editor. What user have to do is only choice of truth value. Of course users can define original truth values in our system.

3.3 Designing Agent's Knowledge Base

Designing knowledge is one of important part of argumentation. Actually we prepare own knowledge very carefully before the argumentation to persuade friends or beat down enemies. In Fig. 2, left upper frame and Fig. 3 shows a knowledge base editor with which agents can prepare their own knowledge bases to be used for argumentation. To reduce effort and archive high error-resistivity, we provide a light-weight structure editor that helps agents specify rules with annotations, separating the right and left parts of rules. Additionally the knowledge base editor supports argument template, user can also designs agent's knowledge base via template. As another way to help agents prepare knowledge bases for argumentation, we take into account a way to transform natural arguments analyzed in Araucaria [9] to formal arguments in LMA [12].

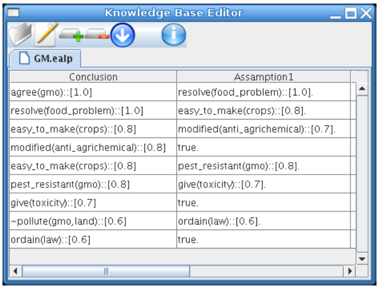


Fig. 3. Screen shot of Knowledge Base Editor

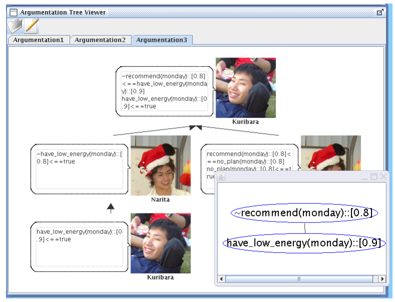


Fig. 4. Screen shot of Argumentation Tree Viewer and Argument Tree Viewer

3.4 Launching Argumentation

In Fig. 2 the triangle labeled button is for launching an argument on some issue or claim which is submitted by an arbitrary agent, from the small box in the Issue input window. In the argument attack pane, agents can set an appropriate rebuttal relation for each of proponents and opponents by choosing from among various rebuttal relations LMA provides [11]. Then, according to the dialectical proof theory of LMA, arguments proceed to produce an argument status: justified, overruled, or defensible. On its way, we can watch the course of an argument on the Live Argumentation Viewer, and we can also watch the appearance for which the dynamic argumentation tree is constructed automatically on the Argumentation Tree Viewer. Our Live Argumentation Viewer shows agent's name, image, locution and icon which presents simple emotion or locution type. During an argumentation, every locution appears in agent own speech balloon.

3.5 Analyzing Argumentation

The argumentation tree viewer frame (Fig. 2, right lower frame) displays the live argumentation process based on the dialectical proof theory, with arguments and counter-arguments put forward during argumentation. Generally, we often discover new knowledges through arguing or analyzing the argumentation note at a later date. For the analyzing, Diagramming is effective. Such a diagramming is the most important part of the integrated argumentation environment since we are concerned with not only argument results but also the overall structure and flow of an argument now developing. Fig. 4 shows an Argumentation Tree Viewer. Each node has three components: agent name, agent image and speech balloon, and root node has argument status label additionally. Each arrow represents attack relation. It also has a label presents type of attack relation. We further can see the structures of arguments themselves diagrammatically by clicking the nodes in the dialectical tree (Fig. 4, right lower pane). All argumentation trees can be saved and loaded, and they can be exported as XML files or JPEG images.

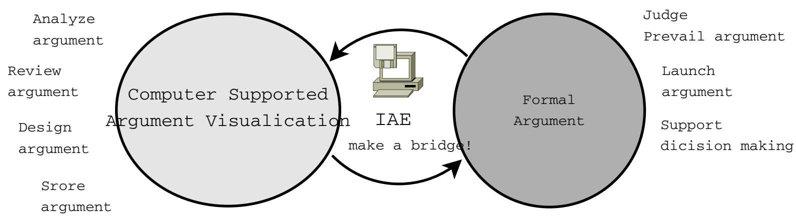


Fig. 5. Overall story of Integrated Argumentation Environment

3.6 Repository of Arguments

Arguments should be reused for the future argumentation as a component of other arguments. Current version of IAE saves argument and argumentation by original form locally. Now we have a plan to introduce argument DB and a common format for arguments like Argument Markup Language (AML) [9] and Argument Interchange Format (AIF) [2]. Arguments of this form are to be distributed on the computer network and reused in the sites on the network where arguing agents reside.

4 Application Domains

So far, two field of researches, Computer Supported Argumentation Visualization (CSAV) and Formal Argument, have not intersected. However there is a great potential when these domains are integrated. The foremost application is computer-supported argumentation in pedagogy. There is huge amount of arguments on the Web. If all of arguments are represented by a common format like a XML, and argument-based agent system can deal with the argument as own knowledge base, agents able to decide prevail argument based on formal argument. It may be effective not only enhancing student's critical thinking skill but also improving argumentative skill. Recently, the research on a common format of argument and argumentation attract attentions, such as Argument Markup Language and Argument Interchange Format[2]. Using these techniques, IAE ensures that argumentation animation and diagrammatized argumentation from stored natural arguments via translation component [12]. Such argumentation visualization may help to understand the structure and the flow of the argumentation. Furthermore, using argument DB, IAE solves problem how to design agent's knowledge. Fig. 5 shows overall story of our ultimate goal. IAE supplements what both CSAV and formal argument lack mutually and tightly integrats function of these two systems into one.

5 Concluding Remark and Future Work

In this paper, we provide Integrated Argumentation Environment (IAE). The system has been designed so that it reflects such an argument metaphor or scenario from issues to agreements that agents first get together at the roundtable-like argument field (place), lively argue with each other on a specific issue, share the visualization of the argument process, and finally reach to an agreement (if any).

The assessment of its effectiveness should be done only through various empirical uses. The overall sentiment expressed by the users of our laboratory reflected the need for facilitation the use of significance of metaphor in the computer world be obvious as can been seen in many windows-based desktop OSs.

IAE is a first step towards support tool for building argument-based agent systems where argumentative reasoning is central to agent-oriented problem solving. Although IAE is generic, it constitutes just an argument engine that facilitates knowledge construction and argument construction with the proper graphical user interface. In particular, plug-in based extensible architecture that allows for incorporating a number of other tools for argumentation and agent systems are most desirable.

In relation to argumentation, there is still a lot of room for the improvement and further development, Here we briefly discuss three intriguing future plans which can make our system more powerful and versatile.

1. *Hybridizing symbolic argumentation with neural net argumentation:* Symbolic argumentation frameworks usually rule out notorious arguments such as self-defeating or more pathological arguments [7] from the public argumentation. We devised a more comprehensive argumentation framework with the help of neural network [5], which tolerates those notorious arguments. The future IAE will incorporate its idea in its underlying argumentation framework LMA [11].

2. *Argument mining from argument DB and argument markup language:* We argue all the time in our daily life, scientific communities, parliaments, courts, online discussion boards and so on. Humans knowledge an wisdom produced there usually have the form of arguments that are built up from more primitive knowledge of the form of facts and rules. Those repositories or treasuries of knowledge are now about to be organized to argument date bases or corpora that and argument discovery technology are paricularlu needed to allow us to create new ideas, opinions and thoughts by finding out meaningful arguments and reusing them from such large-scale argument repositories. Then, a common format for argument representation, like Argument Markup Language (AML) [9], will be required and plays an important role in argument mining and argument discovery technology.

3. *Combining IAE with argument mapping systems:* Argument mapping systems (e.g., [3] [4] [6] [8] [13]) produce diagrams of reasoning and argumentation for especially complex arguments and debates. It improves our ability to articulate, comprehend and communicate reasoning, and hence help us promoting critical thinking. If those diagrammatized arguments have been exported to formal knowledge and augmentation that formal argumentation systems can deal with, they could benefit a huge amount of knowledge for argumentation, and reciprocally argument mapping systems could gain a formal argumentation systems. We just started a small step towards transforming natural arguments in

Araucaria to formal arguments in LMA [12]. We will promote this attempt further to a fully developed IAE.

This is the first step toward realizing impressive and attractive agent-based argumentation or dialogue system.

References

1. Chesñevar, C.I., Maguitman, G., Loui, R.P.: Logical models of argument. ACM Computing Surveys 32, 337–383 (2000)
2. Chesñevar, C.I., McGinnis, J., Modgil, S., Rahwan, I., Reed, C., Simari, G., South, M., Vreeswijk, G., Willmott, S.: Towards an argument interchange format. The Knowledge Engineering Review 21, 293–316 (2006)
3. Conklin, J., Bageman, M.L.: gibis: A hypertext tool for exploratory policy discussion. In: Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW 1988), pp. 140–152 (1988)
4. Kirschner, P.A., Shum, S.J.B., Carr, C.S. (eds.): Visualizig Argumentation, Software Tools for Collaborative and Educational Sense-Making. Springer, Heidelberg (2003)
5. Makiguchi, W., Sawamura, H.: A hybrid argumentation of symbolic and neural net argumentation. In: Proceedings of the Fourth International Workshop on Argumentation in Multi-Agent Systems (2007)
6. Paolucci, M., Suthers, D., Weiner, A.: Belvedere: Stimulating students' critical discussion. In: CHI 1995 Conference Companion, pp. 123–124 (1995)
7. Prakken, H., Vreeswijk, G.: Logical systems for defeasible argumentation. In: Gabbay, D., Guenther, F. (eds.) Handbook of Philosophical Logic, pp. 219–318. Kluwer, Dordrecht (2002)
8. Reed, C., Norman, T.J. (eds.): Argumentation Machines. Kluwer Academic Publishers, Dordrecht (2004)
9. Reed, C., Rowe, G.: Araucaria: Software for argument analysis, diagramming and representation. International Journal of AI Tools 14(3–4), 961–980 (2004)
10. Sawamura, H., Wakaki, T., Nitta, K.: The logic of multiple-valued argumentation and its applications to web technology. In: Dunne, P.E., Bench-Capon, T.J.M. (eds.) Computational Models of Argument, pp. 291–296. IOS Press, Amsterdam (2006)
11. Takahashi, T., Sawamura, H.: A logic of multiple-valued argumentation. In: Proceedings of the third international joint conference on Autonomous Agents and Multi Agent Systems (AAMAS 2004), pp. 800–807. ACM, New York (2004)
12. Takahashi, Y., Sawamura, H., Zhang, J.: Transforming natural arguments in araucaria to formal arguments in lma. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, Hong Kong, China, December 18-22, 2006, pp. 668–678. IEEE Computer Society, Los Alamitos (2006)
13. van Gelder, T.: Enhancing deliberation through computer-supported argument visualization. In: Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, pp. 97–115. Springer, Heidelberg (2002)

Agent Collaboration for Multiple Trading Strategy Integration*

Longbing Cao[1], Dan Luo[2], Yanshan Xiao[3], and Zhigang Zheng[4]

[1,2] Faculty of Information Technology, University of Technology, Sydney, Australia
[3] Department of Information Technology, Guangzhou Asian Games Organizing Committee, China
[4] Department of Information Technology, PICC, Property And Casualty Ltd, China
lbcao@it.uts.edu.au

Abstract. The collaboration of agents can undertake complicated tasks that cannot be handled well by a single agent. This is even true for excecuting multiple goals at the same time. In this paper, we demonstrate the use of trading agent collaboration in integrating multiple trading strategies. Trading agents are used for developing quality trading strategies to support smart actions in the market. *Evolutionary* trading agents are armed with evolutionary computing capability to optimize strategy parameters. To develop even smarter trading strategies (we call *golden strategies*), multiple *Evolutionary* and *Collaborative* trading agents negotiate with each other for m loops to search multiple local strategies with best parameter combinations. They also integrate multiple classes of strategies for trading agents to achieve the best global objectives acceptable for trader needs. Tests of five classes of trading strategies in ten years of five markets of data have shown that agent collaboration for strategy integration can achieve much better performance of trading compared with that of either individually optimized or randomly chosen strategies.

1 Introduction

Collaboration, coordination, cooperation and negotiation are some of key organizational activities in a multi-agent system. Agent collaboration can undertake complicated tasks that usually cannot be handled well by a single agent. This is even true for executing multiple goals at the same time. Another key concept used in this paper is trading agent. Trading agent [5,6,13,14] is a concept developed to design and simulate market mechanisms, auction strategies, and supply chain management etc. The collaboration of trading agents is interesting because in this way it may develop trading support information that cannot be achieved by single agents.

This paper demonstrates the use of trading agent collaboration for integrating multiple classes of trading strategies to support smart trading. We develop

* This work is sponsored by Australian Research Council Discovery Grant (DP0773412, LP0775041, DP0667060), and UTS internal grants.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 361–370, 2008.

Evolutionary trading agents through arming agents with evolutionary computing capabilities. Evolutionary trading agents are therefore able to identify most interesting strategies for smart trading. Further, *Collaborative* agents negotiate with *Representative* trading agents for m loops if necessary to search for the locally optimal trading strategies and then aggregate some of these locally optimal strategies to generate a global optimization objective.

In developing the integrative trading agents, our concern is *to what extent the trading agents powered by the existing approaches can support both technical significance and business decision making [11,8] in real-life marketplaces*. To this end, we customize trading agents with the tasks of satisfying both *trader preference* and taking trading positions in favor of trader's expectation. In general, there are three types of positions: +1, −1, 0 for a trading agent to take in the market. Position +1 indicates a *buy* or *holding buy* action in the market. Position −1 reflects either a *sell* or *holding sell* action. Position 0 indicates none of actions.

Any positions undertaken by a trading agent are associated with certain benefit and cost/risk. Trading agents need to carefully select those trading strategies that can guide them to take positions in the market with higher benefits while controlling costs. To balance benefit and cost, the aim of trading agents is to maintain the highest unit of benefit per cost $\gamma_{\alpha\beta}$.

Definition 1 (Benefit-Cost Ratio). Benefit-cost ratio $\gamma_{\alpha\beta,s}$ *measures the the unit of benefit per cost of a trading agent in undertaking position sequences* $\{b_i\}$ *determined by a trading strategy s.*

$$\gamma_{\alpha\beta,s} = \alpha_s / \beta_s \tag{1}$$

We use stock trading agents to illustrate the development of actionable trading strategies and multi-strategy integration. Evolutionary trading agents for optimizing parameters and trading agent collaboration for strategy integration are proposed to develop smart trading agents for actionable strategies. Seven categories including 36 types of trading strategies have been developed for stock trading agents. Ten years of historical data from five markets are used to evaluate the actionability of the stock trading agents. Massive experiments have shown that our trading agents present excellent performance that can not only beat naive strategies, but also financial market benchmarks.

2 Designing Smart Trading Agents

The idea of *designing smart trading agents* is to endow trading agents with capabilities of searching strategies in constrained market environment to satisfy trader preference. In this section, we introduce two approaches to designing smart trading agents. One is to design *evolutionary trading agents*, which are equipped with evolutionary computing capabilities, and can search strategies from a large candidate strategy space targeting higher *benefit-cost ratio*. The other is to integrate [7] optimal instances from multiple classes of trading strategies into one combined powerful strategy through *collaborative trading agents*.

2.1 Evolutionary Trading Agents for Parameter Optimization

Evolutionary trading agents have capabilities of evolutionary search computing. They can search trading strategies based on given optimization fitness and specified optimization objectives. Their roles consist of *optimization requests* (including base strategies and arguments), *creating strategy candidates* (namely chromosomes), *evaluating strategy candidates*, *crossing over candidate strategies*, *mutating candidate strategies*, *re-evaluating candiate strategies*, and *filtering optimal strategies*, etc.

The strategy optimization mechanism of evolutionary trading agents is as follows. A *User Agent* receives optimization requests from user-agent interaction interfaces. It forwards the request to *Coordinator Agents*, *Coordinator Agents* check the availability and validity of optimized *Strategy Agent* class with *strategyClassID*. If a *Strategy Agent* class is available and optimizable, *Coordinator Agents* call the *Evolutionary Agents* to perform corresponding roles, for instance, *createStrategyCandidates*, *evaluateStrategyCandidates*, *crossoverCandidateStrategies*, *mutateCandidateStrategies*, *re-evaluateCandidateStrategies*, or *returnOptimalStrategies* to optimize the strategy. After the optimization process, *Evolutionary Agents* return *Coordinator Agents* the searched optimal *Strategy Agent* with *strategyID* and corresponding parameter values. *Coordinator Agents* further call the *User Agents* to present the results to traders by invoking *Presentation Agents*. Fig. 1 illustrates the workflow of the above evolutionary trading agents and their relevant collaboration process in searching actionable trading strategies.

The following descriptive notations further illustrate one of the above roles: *mutateCandidateStrategies*.

Role *R_mutateCandidateStrategies*
Statement *Mutation is a process that parts of a chromosome are to be changed. This role determines to what extent the parts of a chromosome in a trading agent are to be mutated. The extent is the mutation rate.*
Agent *A_EvolutionaryAgent*
Agent *A_UserAgent*
Agent *A_StrategyAgent*
Agent *A_CoordinatorAgent*
Attribute *aea:A_EvolutionaryAgent*
Attribute *constant mutrate:MutationRate*
Attribute *paraid[]:A_InParameters*
Attribute *aua:A_UserAgent*
Attribute *asa:A_StrategyAgent*
Attribute *constant strid:asa*
Attribute *aca:A_CoordinatorAgent*
Protocol *receiveStrategyMutationRequest*
Protocol *checkStrategyAgentValidity*
Protocol *openMutateSettingInterface*
Protocol *submitStrategyMutationRequest*

Protocol *returnStrategyMutationResponse*
Responsibilities
Liveness
$\forall$ *strid:aca.checkStrategyAgentValidity()* $\rightarrow$
aua.openMutateSettingInterface(aea, asa.paraid[])
$\rightarrow$ *aea.receiveStrategyMutationRequest(aua)*
$\rightarrow$ *aca.submitStrategyMutationRequest(aua)*
$\rightarrow$ *aea.executeStrategyMutation(aua, mutrate, aca)*
$\rightarrow \Diamond_{\leqslant t}$ *aea.returnStrategyMutationResponse(aua, aca)*
Safety (Invariant) $0 < mutrate < 1.0$

2.2 Trading Agent Collaboration for Strategy Integration

In real-life trading, trading strategies can be categorized into many classes. To financial experts, different classes of trading strategies indicate varying fundamental principles of market models and mechanisms. A trading agent often takes series of positions generated by a specific trading strategy, which instantiates a trading strategy class. Trading agents can collaborate to take concurrent positions created by multiple trading strategies to take advantage of varying strategies.

The idea of *trading agent collaboration for strategy integration [7]* is as follows. There are a few *Representive Trading Agents* in the market. Each *Representive Agent* invokes an *Evolutionary Agent* to search for optimal *Strategy Agent* from a strategy class. *Coordinator Agents* then negotiate with these *Representive Agents* and *Evolutionary Agents* to integrate the identified optimal strategies of *Strategy Agents*. An appropriate integration method is negotiated and chosen by *Coordinator Agents*, *Representive Agents* and *Evolutionary Agents* based on globally optimal output.

For instance, the following notations describe one of the goals of *Coordinator Agents*. The goal is to achieve the globally maximal *benefit-cost ratio* through negotiating with all *Representive Agents*.

Goal *integrateStrategy*
Statement Coordinator *agents discuss with* Representative *trading agents to get maximally* global benefit-cost ratio*.* Representative *trading agents invoke* n Evolutionary *trading agents to execute* n *classes of* Strategy *agents for maximally* local benefit-cost ratio*, respectively. The following describes the objective of agents fulfilling such a task.*
Role *R_StrategyOptimizer*
Agent *A_StrategyAgent*
Agent *A_UserAgent*
Agent *A_RepresentativeAgent*
Agent *A_EvolutionaryAgent*
Agent *A_CoordinatorAgent*
Attribute $aea : A\_EvolutionaryAgent_i$
Attribute $aua : A\_UserAgent$

Attribute asa : $A\_StrategyAgent_i$
Attribute ara : $A\_RepresentativeAgent$
Attribute aca : $A\_CoordinatorAgent$
Attribute $constantstrid$: asa
Attribute $constantstrid$: asa
Attribute an : $AlgoName$
Attribute ac : $AlgoCode$
Attribute $ain[]$: $AlgoParameters$
Attribute $aout[]$: $AlgoOutputs$
Creation condition $\neg Existed(ac)$
Invariant condition $ac.actor = ActorID$
Fulfillment condition
$\forall$ *ac:AlgorithmComponent (*
ac.algo = *algo* $\rightarrow$
$\Diamond_{\leqslant t_1} \exists$ *cpi:CallPluginInterfaces (cpi.actor* = *actor* $\wedge$ *Fulfilled(cpi))*
$\wedge$ $\Diamond_{\leqslant t_2}$($\exists$ *faro:FillinAlgoRegisterOntologies*
(faro.depender = *actor* $\wedge$ *Fulfilled(faro))*
$\wedge$ $\exists$ *uac:UploadAlgoComponent*
(uac.depender = *actor* $\wedge$ *Fulfilled(uac)* $\wedge$ *ac.uploaded)*
)
)

Fig. 1 further describes the process of trading agent negotiation for strategy integration. As shown in the diagram, there are two steps of optimzation. First, locally optimal strategies are searched through *Evolutionary* agents on request of *Representative Agents* if the strategy achieves the highest *benefit-cost ratio* σ. *StrategyManager Agent* stores the golden strategies. Second, *Coordinator* agents call *StrategyIntegrator Agents* to work out the requested global goal. *Coordinator* agents check *StrategyManager Agent* and invoke *Evolutionary* agents if necessary to recalculate the golden strategies based on negotiation model. *StrategyIntegrator Agents* select the best golden strategies for each loop and accumulate all promising strategies for m loops to achieve the requested globally optimal goal by following the agreed negotiation model.

3 Real-World Experiments

Since 2002, we have been working on developing trading agents and strategies with industrial partners' support, say CMCRC, SIRCA and SMARTS [1,3,2]. Massive experiments have been conducted on many years of multi-markets of data. An agent service-based platform F-Trade [4,9] has been built to support this effort. Some of our results have been delivered to partners. In this section, we illustrate the process and results in optimizing strategies through *Evolutionary Trading Agents* and integrating strategies via *Collaborative Trading Agents*.

Given a trading strategy s, a trading strategy class S_i (i=*1, 2, ...*), $s \in S_i$, α_s and β_s are the *benefit* and *cost* of a trading agent in executing the

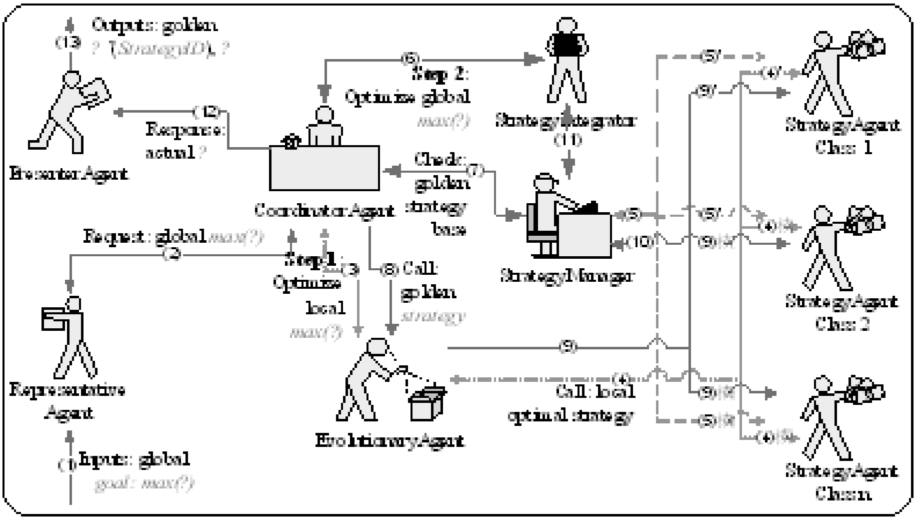


Fig. 1. Workflow of trading agent collaboration for strategy integration

strategy s. The development process of integrating strategies through trading agent collaboration is as follows.

Part A. *Data Manager Agent* prepares data:
0). *UserAgent* receives trader's input requests;
1). *DataManager* agent splits two years of data for training;
2). *RepresentativeAgent* invokes *EvolutionaryAgents* to identify locally golden trading strategies with highest $\gamma_{\alpha\beta,s}$ as discussed in part B;
3). *DataManager* agent splits another three years of data following the training windows for testing;
4). *RepresentativeAgent* invokes *EvolutionaryAgents* to test the identified golden strategies as discussed in part B;
5). *DataManager* agent slides the 2-year training and the 3-year deploying data windows one year forward to extract data sets as in A:1) and A:3);
6). *RepresentativeAgent* invokes *EvolutionaryAgents* to repeat the operations of searching golden strategies;
Part B. *Evolutionary Trading Agents* searching golden strategies:
1). *EvolutionaryAgent* calls a *StrategyAgent* s in class S_i and searches strategy instance s' with $max(\alpha_{s'})$ for s' positions;
2). *EvolutionaryAgent* calls a *StrategyAgent* s and searches strategy s'' with $max(\gamma_{\alpha\beta,s})$ when s'' positions are executed;
3). *RepresentativeAgent* invokes *EvolutionaryAgents* to search all strategies s''_i $((i = 1, 2, \ldots))$ in all strategy classes satisfying conditions in step B:2) respectively;
Part C. *Collaborative Trading Agents* aggregate golden strategies:
1). *PositionAgents* extract all positions from *EvolutionaryAgents* with all

strategies identified in step B:3) for *RepresentativeAgent*;
2). *EvaluationAgents* check the *benefits*, *costs* and *benefit-cost ratio* of each *RepresentativeAgent* executing the above positions;
3). *DecisionAgents* filter out strategies with low $\gamma_{\alpha\beta,s}$ for each strategy class i;
4). *CoordinatorAgents* call all *RepresentativeAgents* to execute the above filtered strategies concurrently to generate the final outcomes.

Experiments of trading agent collaboration for multi-strategy integration in stock market data have been conducted as follows:

- Five classes of trading strategies are developed: MA, FR, CB, SR, and OBV;
- Five stock markets: ASX, Hongkong, London, New York, and Japan;
- Interday data from 1/11/1998 to 31/10/2007: date, price, volume as shown in Table 1;
- Training data: 2-year sliding window, say 1/11/1998-31/10/1999;
- Testing data: 1-year sliding window, say 1/11/1999-31/10/2000.

Table 1. Data sample

Date	Price	Volume
2006-12-14	16.39	239943
2006-12-15	16.74	183908
2006-12-18	17.25	203883
2006-12-19	16.97	178483

Fig. 2 illustrates some results of evolutionary trading agent for optimizating the *Filter Rule Base Strategy* $FR(x)$. $FR(x)$ indicates a generic class of correlated trading strategies, by which you go long at the time that the price rises by x% and hold until the price falls x%, at which time you close out and go short, where $x \in [0, 1]$ is the percentage price movement of highest high and lowest low.

Even though there is only one parameter d in this rule, it is hard to find the most appropriated 'x' in real-life market. Evolutionary trading agent is helpful for searching the golden 'x'. As shown in Fig. 2, the cumulative payoff with $x = 0.04$ always outperform other ds from 14 July 2003 in trading the listed security CBA (Australian Commonwealth Bank) in Australian Stock Exchange (ASX) in 2003-2004.

Table 2 shows the signals, positions, benefits and costs of trading agents following *MA-BMN Strategy*, which is an identified golden strategy by evolutionary trading agent in 2004 Hongkong Exchange data.

Table 3 shows the positions recommended by each golden strategy identified by collaborative trading agents in 2006 Hongkong Exchange data.

Fig. 3 shows the cumulative benefits *MA_Ben, FR_Ben, CB_Ben, SR_Ben, OBV_Ben* of trading agents taking positions recommended by golden trading strategies *MA, FR, CB, SR, OBV*, as well as that (Int_Ben) of the *Collaborative*

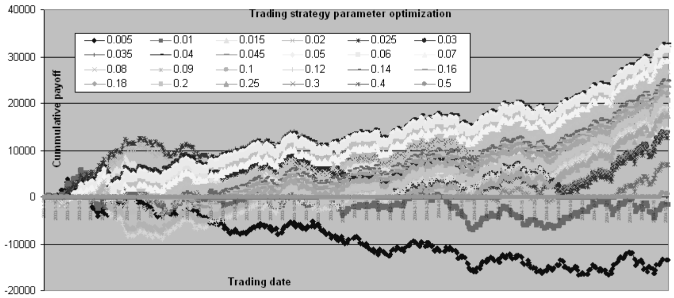


Fig. 2. Some results of evolutionary trading agent for strategy optimization

Table 2. Output excerpt of a trading strategy

Date	Price	Sell	Buy	Position	Benefit($)	Cost($)
$2004-8-16$	3466	-1	0	-1	9200	103
$2004-8-17$	3480	-1	0	-1	8850	106.5
$2004-8-18$	3472	-1	0	-1	9150	108.5
$2004-8-19$	3481	-1	0	-1	8825	110.75
$2004-8-20$	3494	0	0	-1	8500	114

Table 3. Trading agent positions recommended by five trading strategy classes (excerpt)

Date	MA Pos	FR Pos	CB Pos	SR Pos	OBV Pos
$2006-11-16$	1	1	0	1	1
$2006-11-17$	1	1	0	1	1
$2006-11-20$	1	1	0	1	1
$2006-11-21$	-1	-1	0	1	1
$2006-11-22$	-1	-1	0	1	1

Trading Agent executing all golden positions by integrating individual golden strategies concurrently in 2003-2006 Hongkong United Exchange data.

A large amount of tests in stock data of five markets have shown trading agents following all golden trading strategies can obtain higher benefits and benefit/cost ratios (except FR in the first few days). In particular, collaborative trading agents concurrently executing positions from all individual golden strategies can greatly increase benefits while control very low costs compared with those taking positions recommended by either an individually golden strategy or randomly chosen strategies only (see Table 4, *lift* [10] measures how much good a trading strategy is in all split data sets).

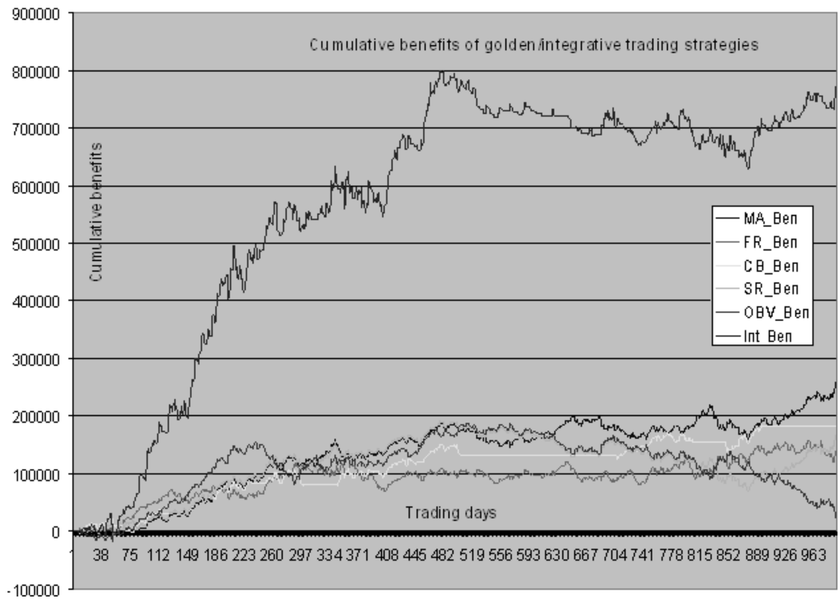


Fig. 3. Cumulative benefits of trading agents following golden trading strategies

Table 4. Lift comparison between random chosen strategies and golden strategies

Lift	MA-CMN	FR-XY	OBV-B	CB-XNC	SR-NC
Random	10%	0	20%	10%	10%
Optimized	70%	80%	80%	90%	100%

4 Conclusion

Agent collaboration and negotiation is very helpful for solving complicated tasks and achieve multiple goals at the same time. Trading agent has demonstrated its potential in simulating market mechanism design and strategy development. This paper has demonstrated the use of agent collaboration for optimizing and aggregating multiple classes of trading strategies. Trading agent can contribute to traders with trading strategies that can support their action-taken in the market. First of all, trading agents are armed with evolutionary computing capability. The evolutionary capability enables trading agents to search for parameter combinations with the most appropriate performance. Further, evolutionary and collaborative trading agents collaborate with each other to generate locally optimal trading strategies, and then produce integrative and globally optimal strategies. The integrative trading strategies enable trading agents with trading positions that can lead to higher profit but lower costs.

References

1. CMCRC: www.cmcrc.com
2. SIRCA: www.sirca.org.au
3. SMARTS: www.smarts.com.au
4. F-Trade: www.f-trade.info
5. Trading Agent Competition: http://www.sics.se/tac/
6. Trading Agent Design and Analysis: http://www.sics.se/tada05/
7. Cao, L.B.: Multi-strategy integration for actionable trading agents. In: Proceedings of the International Conference on Intelligent Agent Technology, IEEE Computer Society Press, Los Alamitos (2007)
8. Cao, L.B., Zhang, C.Q.: Two-Way Significance of Knowledge Actionability. Int. J. of Business Intelligence and Data Mining (2007)
9. Cao, L.B., Zhang, C.Q.: F-Trade: An Agent-Mining Symbiont for Financial Services. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, ACM Press, New York (2007)
10. Cao, L.B., Luo, C., Zhang, C.Q.: Developing Actionable Trading Strategies for Trading Agents. In: Proceedings of the International Conference on Intelligent Agent Technology (2007)
11. Cao, L.B., Luo, D., Zhang, C.Q.: Fuzzy Genetic Algorithms for Pairs Mining. In: Yang, Q., Webb, G. (eds.) PRICAI 2006. LNCS (LNAI), vol. 4099, pp. 711–720. Springer, Heidelberg (2006)
12. Chan, T.: Artificial Markets and Intelligent Agents, PhD thesis, Massachusetts Institute of Technology (2001)
13. Wan, H.A., Hunter, A.: On Artificial Adaptive Agents Models of Stock Markets. Simulation 68, 279–289 (1997)
14. Wellman, M.P., et al.: Designing the Market Game for a Trading Agent Competition. IEEE Internet Computing 5(2), 43–51 (2001)

Strategy Acquisition on Multi-issue Negotiation without Estimating Opponent's Preference

Shohei Yoshikawa, Yoshiaki Yasumura, and Kuniaki Uehara

Dept.of Computer Science and Systems Engineering Graduate School of Engineering
Kobe University, Japan

Abstract. In multi-issue negotiation, an opponent's preference is rarely open. Under this environment, it is difficult to acquire a negotiation result that realizes win-win negotiation. In this paper, we present a novel method for realizing win-win negotiation although an opponent's preference is not open. In this method, an agent learns how to make a concession to an opponent. To learn the concession strategy, we adopt reinforcement learning. In reinforcement learning, the agent recognizes a negotiation state to each issue in negotiation. According to the state, the agent makes a proposal to increase own profit. A reward of the learning is a profit of an agreement and punishment of negotiation breakdown. Experimental results showed that agents could acquire a negotiation strategy that avoids negotiation breakdown and increases profits of both sides. Finally, the agents can acquire the action policy that strikes a balance between cooperation and competition.

1 Introduction

As the Internet spreads throughout the world, the needs of e-commerce is growing. In e-commerce, we need transaction systems using the network and technologies for effectively supporting users. One of transaction systems is negotiation agents that support negotiation automatically instead of users. For supporting the user, negotiation agents should have good strategies. Negotiation have been researched in the field of the game theory such as the divorce game[11]. In game theory, both players know complete information of the opponent players. However, in actual negotiation, both players do not know those information completely.

So far, two types of negotiation have been researched. The first type is single-issue negotiation [1,2] which deals with only a price. In actual negotiation, negotiators generally deal with not only a price. The second type is multi-issue negotiation that considers multiple issues in addition to a price [3,4,5,6]. The pleading game[10] deals with multiple issues. This theory is a formal normative model of argumentation. The goal of the game is identifying issues, rather than deciding them.

In multi-issue negotiation, as the number of issues increases, the negotiation becomes more complex. Besides, in multi-issue negotiation, it is possible to reach

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 371–380, 2008.

an agreement that is mutually beneficial for both sides. It is called win-win agreement. In order to lead a win-win agreement, negotiation strategies for agents are required in this type of negotiation.

Yoshikawa et.al proposed a method for acquiring win-win strategy when an opponent's preference is open[9]. However, in an actual negotiation, an opponent's preference is rarely open. One of the methods for acquiring strategies without knowing opponent's preference is to estimate the preference from opponent's proposals[4]. By using this estimation, an agent can make a proposal that meets an opponent's request. However, it is difficult to precisely estimate an opponent's preference by learning because an opponent changes his/her negotiation strategy according to negotiation state. Moreover, it is difficult for an agent to change its strategy after estimating preference.

In this paper, we present a novel method for learning negotiation strategies without estimating opponent's preference. In this method, an agent learns concession rate to each negotiation issue from an opponent's concession rate. For learning a concession rate, we adopt reinforcement learning to increase own profit. In order to increase own profit, the agent should acquire the strategy that increases opponent's profit because the opponent do not agree to a selfish proposal. That is, the learning agent should acquire the strategy that increases both sides' profit. As a result, the agent can acquire a win-win strategy by learning.

2 Negotiation Model

In this section, we define a basic model of negotiation. First, we present the negotiation setting. Second, we describe the negotiation agent for negotiating with an opponent. Under this environment we propose a novel method for getting an optimal strategy.

2.1 Negotiation Setting

Here we define a setting of multi-issue negotiation. In this negotiation, two agents make proposals alternately. In multi-issue negotiation, products targeted for negotiation is represented as n values $(x_1, ..., x_n)$. Thus, products has n issues. To evaluate a value of an issue, each agent has an evaluation function $v_i(x_i)$. In addition, each agent has preference as a weight vector $W = \{w_1, w_2, ..., w_n\}$ over the issues, where w_j is the relative importance that an agent attaches to issue j. These weights are normalized as $\sum_{i=1}^{n} w_i = 1$. Thus, a utility U of an agent is defined as follows.

$$U = \sum w_i v_i(x_i) \tag{1}$$

As weight vectors of agents are various, each agent can acquire a different utility for the same proposal. This difference enables agents to realize win-win negotiation.

2.2 Negotiation Agent

Here, we describe a negotiation agent. A negotiation agent is a software that negotiates with an opponent instead of a user. The goal of a negotiation agent is to find win-win agreement for both sides. The actions that an agent can use are a proposal, agreement, and breakdown. In this research, agents' preference is not open in negotiation. The only information that an agent can be available is opponent's proposals. There are many types of agents that have various methods for making a proposal, such as a self-interest agent and a cooperative agent.

We describe a condition of an agreement. The tth proposal denotes the opponent's last proposal and the $(t-1)$th proposal denotes the last own proposal. An agent makes an agreement when the own utility at $(t-1)$th proposal is more than that at tth proposal. Using the utility function U, the condition of an agreement is expressed as $U_t^{own} - U_{t-1}^{own} > 0$. U_t^{own} represent the own benefit on tth proposal.

We define two kinds of breakdown : breakdown by deadline and breakdown by rejection. In breakdown by deadline, negotiation is finished when the number of proposals is more than the deadline. With tth proposal, the condition of breakdown by deadline is expressed as $t > T_{deadline}$. $T_{deadline}$ represents the limited number of proposals that an agent can make in a negotiation.

In breakdown by rejection, negotiation is finished when an agent decides to break down the negotiation. The decision is based on the prediction value at the deadline. This prediction value is calculated by a linear regression. An agent has an reservation utility. An reservation utility represents the minimum value that is acceptable for an agent. The negotiation is broken down before deadline when the prediction value is less than the reservation utility. If the utility at the agreement is lower than the reservation utility, an agent suffers a loss in this negotiation. With the utility function U, the condition of breakdown by rejection is denoted as $U_{deadline}^{own} < RU^{own}$. $U_{deadline}^{own}$ represents a predicted own utility at the deadline. RU^{own} expresses an own reservation utility.

3 Acquisition of Strategy

In this section, we propose a novel method for acquiring strategies to deal with opponents that have various preferences.

3.1 Acquisition of Strategy

We assume that an opponent's preference is not open in this research because it is rarely open in an actual negotiation. Under this environment, it is difficult to acquire a negotiation result that reflect each agent's preference. For this reason, win-win negotiation is not easy in this environment. Although an opponent's preference is not open, it is required for an agent to acquire strategies to realize win-win negotiation.

One of the methods for acquiring strategies is to estimate an opponent's preference from opponent's proposals. By using this estimation, an agent can make a proposal that meet an opponent's request. However, it is difficult to estimate an opponent's preference by learning precisely because an opponent changes his/her negotiation strategy according to negotiation state.

In this research, we present a novel method that an agent learns strategies without estimating an opponent's preference. In this method, an agent learns concession rate to each negotiation issue from an opponent's concession rate. An opponent's concession rate reflects an opponent's preference. By learning concession rate to each negotiation issue, it is possible to make the proposal that take into account of opponent's preference. For acquiring a negotiation strategy, we apply the reinforcement learning. As an agent acquires the better strategy with reinforcement learning, it learns the action policy that increases its profit.

You might guess that an agent learns the action that increase only its profit. However, an agent is expected to learn differently when we apply breakdown in negotiation. An agent need to learn the action policy that makes concession and increases an opponent's profit in order to avoid breakdown. At the same time, an agent need to learn the action policy that increases its profit. As a result, it can acquire a strategy that increases both sides' profits. Finally, the agent acquire the win-win strategy.

3.2 Reinforcement Learning

Reinforcement learning refers to a class of problems in machine learning. Reinforcement learning algorithms attempt to find a policy for maximizing cumulative reward for the agent over the course of the problem. Q-learning is famous reinforcement learning method. In Q-learning, we use a tuple (s, a, s', r), which is a condition s, action a, next condition s', and reward r. An agent learns by updating the evaluation of action $Q(s, a)$ denoted as equation (5).

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha[r + \gamma Q(s', a')] \tag{2}$$

where $\alpha \in (0\ 1]$ is a learning rate parameter. $\gamma \in (0\ 1]$ is the discount factor. $Q(s, a)$ is denoted as the expected value of the reward given by taking action a in a state s. An agent chooses an action according to $Q(s, a)$. Updating $Q(s, a)$, the agent can find an optimal action policy.

3.3 Learning Concession Strategy

A negotiation agent learns an action based on a reward of an agreement proposal and negotiation breakdown. For choosing the next actions, we adopt ϵ-greedy strategy in learning. The goal of this learning is to acquire an optimal concession strategy for leading a good agreement. For this learning, states, action, and reward should be defined appropriately.

$\langle$state$\rangle$ A current state in a negotiation is represented as $ConcessionBalance$, and $NegotiationStage$. The ConcessionBalance is expressed as the difference of own total concession and opponent's total concession on each issue. By observing the ConcessionBalance, the agent can recognize which issue to be conceded. With the total concession α_j, the ConcessionBalance for each issue is

$$ConcessionBalance = \sum_{i=1}^{t-1} \alpha_{j,i}^{own} - \sum_{i=1}^{t-1} \alpha_{j,i}^{opp}. \tag{3}$$

where $\sum_{i=1}^{t-1} \alpha_{j,i}^{own}$ is the own total concession on j issue. $\sum_{i=1}^{t-1} \alpha_{j,i}^{opp}$ is the opponent's total concession on j issue.

The NegotiationStage means how close both of proposals are. This is expressed as the difference between own utility in an own proposal and that in a opponent's proposal. With an utility function U, difference of the utilities is

$$NegotiationStage = U_{t-1}^{own} - U_t^{own}. \tag{4}$$

where U_{t-1}^{own} is denoted as the utility in the last own proposal. U_t^{own} is denoted as the utility in an opponent's proposal. With the three-dimensional state space, an agent can recognize a negotiation state.

⟨action⟩ An action of an agent is to decide a concession rate for each issue : how much an agent makes a concession to an opponent in the next proposal. This rate is expressed as eight discrete values. An agent decides the value for each issue.

⟨reward⟩ An agent makes a proposal by deciding a concession rate in action space. In this process, when the agents reach an agreement, the agent obtains a reward that is defined as the equation (8).

$$Reward = U_t^{own} - RU^{own} \tag{5}$$

where U_t^{own} is the own utility on the agreement proposal. When the own utility is higher than the reservation utility RU^{own}, the agent gets a positive reward. On the other hand, when an own utility is lower than the reservation utility RU^{own}, an agent gets a negative reward. A negotiation breakdown gives an agent a negative reward.

4 Experiment

In this section, we provide an experimental analysis to evaluate the performance of our method. We conduct some experiments.

4.1 Experimental Setting

The basic setting of these experiments is denoted as follow.

- Bilateral negotiation is conducted.
- The opponent's preference is not open.
- Proposals are submitted each other.
- The number of issues is three.

- The first proposal of an own agent is {8, 8, 8}.
- The first proposal of an opponent is {2, 2, 2 }.
- The reservation utility of both agents is 6.0.
- The learning rate α is 0.8, and the discounting rate γ is 0.5.
- An agent and an opponent have random preferences in every negotiation.
- A learning agent makes a next proposal based on the concession rate acquired by learning.

The strategies of opponents in experiments are three types of concession strategies ,High-Low strategy, Linear strategy, Low-High strategy.

First, we describe a High-Low strategy. The feature of this strategy is that the agent makes a large concession in the early stage of the negotiation, and gradually decreases concession rate. The agent decides the concession rate α by equation $\alpha = \beta(U_{MAX} - U_t^{opp})$,where U_{MAX} is the maximum of utilities. U_t^{opp} is opponent's utility on the own proposals. That is, U_t^{opp} indicates the total concession that an agent makes to an opponent. β is denoted as the speed of negotiation.

Next, we describe a linear strategy. This is the simple strategy that makes a constant concession to an opponent. In this strategy, the amount of concession α is decided as this equation $\alpha = c$,where c is denoted as a constant.

Finally, We describe a Low-High strategy that is contrary of the High-Low strategy. The agent decides a concession rate α using this equation $\alpha = \beta U_t^{opp}$.

4.2 Experiment 1

We focus on how an agent learns behavior without breakdown. The opponent agent has linear strategy that chooses a random concession at the proposal. An agent learns 50000 times.

Figure 1 shows the result of this experiment. The vertical axis indicates each agent's utility at the agreement. Each agent's utility is averaged in every one hundred learning. The horizontal axis indicates the number of learning. The straight line denotes the learning agent's utility at the agreement. The dashed line denotes the opponent agent's utility at the agreement.

In Figure 1, at the early stage of learning the agent makes a loss. However, as the number of learning increases, the agent gradually acquired the action that increases the own utility. Finally, the agent learned the action that makes the own utility much higher than that of the opponent. As a result, though the agent can learn to increase the own utility, the opponent makes a loss largely at the agreement. This means that the agent simply reinforced the action that makes little concession to increase the own utility.

4.3 Experiment 2

We performed the experiments for demonstrating the effect of breakdown. In the experiment 2, agents use breakdown by deadline and breakdown by rejection. In the experiment 2-1, we use a linear strategy agent that makes random concessions

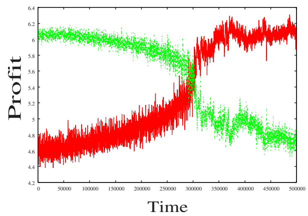


Fig. 1. Experiment 1

as an opponent. In the experiment 2-2, we use the opponent agent that has three strategies, linear strategy, High-Low strategy, and Low-High strategy. The opponent chooses one of three strategies randomly.

Experiment 2-1

Here we show the result in the case of applying breakdown by deadline and breakdown by rejection. Figure 2 shows the result of this experiment. Figure a) shows the change of both agents' utilities. The vertical axis indicates each agent's utility at the agreement. Each agent's utility is averaged in every one hundred learning. The horizontal axis indicates the number of learning. The straight line denotes the learning agent's utility at the agreement. The dashed line denotes the opponent agent's utility at the agreement. Figure b) shows the change of the sum of the both agents' utilities at the agreement. Figure c) shows the change of the total number of breakdown in every one hundred learning.

In Figure a) and Figure c), due to small number of breakdown at the early stage of learning, the agent reinforces the action making a little concession to increase the own utility. The number of breakdown increases from the early stage to the middle stage of learning. This means that little concession causes breakdown. From the middle stage to the late stage of learning, the number of breakdown decreases. This means that the negative rewards at breakdown revises the actions of the learning agent. By receiving the negative reward, the agent reinforces the action making a large concessions to avoid breakdown. At the same time, the agent reinforces the action that increases the own utility. Finally, the agent finds the condition that maximizes the own utility and avoids breakdown. As a result, the agent obtains the more profit and the opponent agent also obtains the more profit. In Figure b), as the learning progresses, the sum of both agents' utilities increases. By reinforcing the action, the agent can increase the sum of utilities. As a result, the agent acquires the strategy that makes a win-win agreement. We find that the agent can deal with the opponent that uses a linear strategy although the opponent's preference is not open.

Experiment 2-2

Here we perform an experiment in the setting that an opponent has various strategies. An opponent randomly chooses one of three strategies in each negotiation such as linear strategy, High-Low strategy, Low-High strategy.

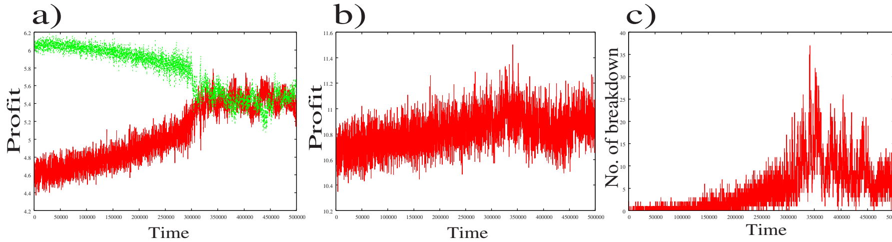


Fig. 2. Experiment 2-1

Figure 2 shows the result of this experiment. Figure a) shows the change of the utilities of both agents. The vertical axis indicates each agent's utility at the agreement. The horizontal axis indicates the number of learning. The straight line denote the learning agent's utility at the agreement. The dashed line denote the opponent agent's utility at the agreement. Figure b) shows the change of the sum of the both agents' utilities at the agreement. Figure c) shows the change of the total number of breakdown in every one hundred learning.

In Figure a), as well as experiment 2-1, the agent reinforces the action to increases the own utility until both agents' utilities are almost equal. In Figure b), by controlling the both agents' utilities, the agent can increase the sum of utilities. However, in Figure c), the number of breakdown does not decrease after learning. One of the cause of this problem is as follows. The opponent agent that uses High-Low strategy tends to make a breakdown by rejection because of the little concession in the early stage of negotiation. The opponent agent that uses Low-High strategy also tends to make a breakdown by deadline because of the little concession in the late stage of negotiation. Moreover, since the agent also reinforces the action that makes a little concession as the learning progresses, the agent tends to make a breakdown by rejection. As a whole, the number of breakdown increases as the learning progresses. Although the agent acquires the strategy that increases the own utility, it can't avoid to make a breakdown. As a result, the agent acquires the strategy that makes a win-win agreement. However, the number of breakdown does not decrease. This is the problem to be solved in the future work.

4.4 Experiment 3

In this experiment, we show how an agent learns actions against a learning agent. Figure 4 show the results of this experiment. Figure a) shows the change of both agents' utilities. In Figure a), the vertical axis indicates the utility at the agreement. The horizontal axis indicates the number of learning. The straight line denote the learning agent's utility at the agreement. The dashed line denote the opponent agent's utility at the agreement. Figure b) shows the change of the sum of the both agents' utilities at the agreement. Figure c) shows the change of the total number of breakdown in every one hundred learning.

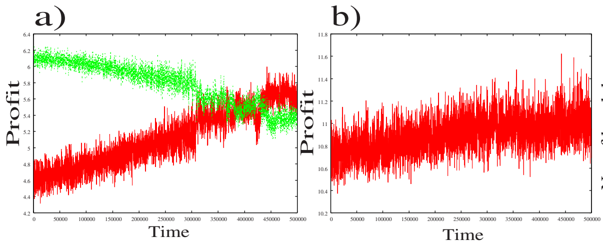


Fig. 3. Experiment 2-2

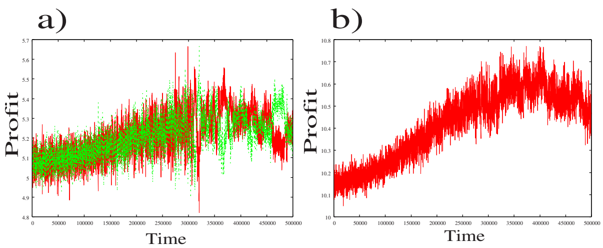


Fig. 4. Experiment 3

In the early stage of learning, both of agents learn to make a little concession because of small number of breakdown. In Figure a), as the learning progresses, the sum of both agents' utilities increases. This means that each agent learns the action to increase the own utility. In Figure b), from the early stage to the middle stage of learning, the number of breakdown increases. This means that each agent makes breakdown by rejection to the other agent that makes a little concession. Receiving the negative rewards by breakdown, each agent revises the action. From the middle stage to late stage of learning, the number of breakdown decreases. Finally, the agents learns action policies that increase the both agents' utility and avoid to breakdown. As a result, both agents can acquire the strategy to realize win-win negotiation. Thus, using our method, we can make negotiations more effective.

5 Conclusion

In this paper, we presented a learning method of strategies in multi-issue negotiation under the situation where an opponent's preference is not open. In this method, we use reinforcement learning to each issue for acquiring a concession strategy. Our experiments showed that the agent could acquire the strategies that both agents can obtain more profit. Under a constraint of breakdown by deadline and breakdown by rejection, the agent could revise a competitive action and acquire a cooperative action. Finally, we found that the learning agent realizes the win-win agreement in multi-issue negotiation under the situation where an opponent's preference is not open.

References

1. Tsang, E., Gosling, T.: Simple Constrained Bargaining Game. In: Proc. of First International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS-2002 (2002)
2. Huang, P., Sycara, K.: A Computational Model For Online Agent Negotiation. In: Proc. of 35th Hawaii International Conference on System Sciences, HICSS-2002 (2002)
3. Fatima, S.S., Wooldridge, M.J., Jennings, N.R.: Multi-issue negotiation under time constraints. In: Proc. of First International Joint Conference on Autonomous Agents and Multi-agent systems, AAMAS-2002, pp. 143–150 (2002)
4. Coehoorn, R.M., Jennings, N.R.: Learning an opponent's preferences to make effective multi-issue negotiation tradeoffs. In: Proc. of 6th International Conference on E-Commerce, pp. 59–68 (2004)
5. Jonker, C.M., Robu, V.: Automated Multi-Attribute Negotiation with Efficient Use of Incomplete Preference Information. In: Proc. of Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS-2004, pp. 1056–1063 (2004)
6. Bosse, T., Jonker, C.M., Treur, J.: Experiments in Human Multi-Issue Negotiation: Analysis and Support. In: Proc. of Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS-2004, pp. 671–678 (2004)
7. Lau, R.Y.K., Valdal, E.: Discovering Negotiation Knowledge for a Probabilistic Negotiation Web Service in e-Business. In: Proc. of Web Intelligence 2004, pp. 264–270 (2004)
8. Brzostowski, J., Kowalczyk, R.: Predicting partner's behaviour in agent negotiation. In: Proc.of the fifth international joint conference on Autonomous agents and multiagent systems, pp. 355–361 (2006)
9. Yoshikawa, S., Kamiryo, T., Yasumura, Y., Uehara, K.: Strategy Acquisition of Agents in Multi-Issue Negotiation. In: Proc. of International Conference on Web Intelligence 2006, pp. 933–939 (2006)
10. Gordon, T.F.: The pleadings game: formalizing procedural justice. In: Proc. of the 4th international conference on Artificial intelligence and law ICAIL, pp. 10–19 (1993)
11. Lundberg, Pollak, S., Robert, A.: Noncooperative Bargaining Models of Marriage. The American Economic Review, 132–137 (1994)

Predictor Agent for Online Auctions

Deborah Lim, Patricia Anthony, and Chong Mun Ho

School of Engineering and Information Technology,Universiti Malaysia Sabah
Locked Bag 2073, 88999, Kota Kinabalu, Sabah, Malaysia
deborahlim.05@gmail.com, panthony@ums.edu.my, cmho@ums.edu.my

Abstract. In the last few years online auctions have become a popular method in purchasing and selling goods over the Internet. Bidding in an online auction is a challenging task since we do not know the outcome of our bid until the auction is closed. It is difficult to predict the winning bid of any particular auction. Hence, many investors have been trying to find a better way to predict auction closing price accurately. Knowing the closing price of a given auction would be an advantage since this information will be useful and can be used to ensure a win in a given auction. This information is beneficial to bidders since the outcome of the auction is dependent on several factors such as the number of auctions selling the same item, the number of bidders participating in that auction as well as the behaviour of every individual bidder. If the closing price of an auction is known, then bidder could decide which auction to participate and at what price. This paper reports on the development of a predictor agent that attempts to predict the online auction closing price. The performance of this predictor agent is compared with two well known techniques which are the Simple Exponential Function and the Time Series in a simulated auction environment.

Keywords: Agent, Grey Theory, Time Series, Simple Exponential Function.

1 Introduction

The design and conduct of auctioning institutions have occupied the attention of many people over thousands of years. Auction is a market institution with an explicit set of rules in determining resource allocation and the price on the basis of bid from the market participants [14]. Auction mechanisms have become very popular within electronic commerce and it has been implemented in many domains with assorted environments. Unlike traditional auction houses, online auction websites offer a better place for people to purchase and publicize their merchandise through a bidding process. The demands for online auction are increasing. Some examples of popular online auction houses include eBay, Amazon, Yahoo!Auction, Priceline, UBid, and FirstAuction. Online auction has also given consumers a "virtual" flea market with all the new and used merchandises from around the world. The auctioneers also have the ability to market their valuable items globally.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 381–390, 2008.

There are four main types of single-sided auctions that are commonly used in traditional auctions [14] which are ascending-bid auction (also called the open, oral, or English auction), descending-bid auction (also called Dutch auction), first-price sealed bid auction and second-price sealed bid auction (also called Vickrey auction). The English auction begins with the lowest price and bidders are free to raise their bid successively until there are no more offers to raise the bid and the bidder with the highest bid will be the winner. The Dutch auction is the opposite of an English auction, where the auctioneer will start with an initial high price and is progressively lowered until there is an offer from a bidder to claim the item. In the first price sealed bid, each bidder submits their offer for an item privately. The highest bidder gets the item and pays for the item based on his bid value. The Vickrey auction is similar to the first-price sealed bid auction, as the item goes to the highest bidder but he only pays a price equal to the second highest bid. Online auctions are similar to the traditional auctions but most auctions are constrained by time. Online auctions usually last for days and week depending on the seller's requirement.

Due to proliferation of these online auctions, consumers faced problems on monitoring multiple auction houses, picking which auction to participate in, and making the right bid to ensure that they get the item under conditions that are consistent with their preferences [1]. Hence, having a good prediction model to predict the closing price of a given auction would definitely be an advantage as the bidder can decide the amount and when to bid. Unfortunately, predicting a closing price for an auction is not easy since it is dependent on many factors such as the behaviour and the number of the bidders who are participating in that auction.

As a result, many investors have been trying to find a better way to predict auction closing price accurately. Neural Network, Fuzzy Logic, Evolutionary Computation, Probability Function and Genetic Algorithm, has been integrated to become a more commendable practical model for prediction purposes. A large body of analysis techniques has been developed, particularly from methods in statistics and signal processing such as the Time Series methods and the exponential function. The Grey Theory is a new method which applies to the study of unascertained problems with few or poor incoming information [11]. It has been successfully applied to economical,management, social systems, industrial systems, ecological systems, education, traffic, environmental sciences, and geography [12]. Some of the work on Grey Theory includes forecasting the Li-ion battery charge [10], predicting the gas-in-oil concentrations in oil-filled transformer [16] and analyzing dynamic and future customer requirements based upon historical, present and predicted data sets [7].

In this paper, we describe a predictor agent that utilizes Grey Theory to predict the closing price of online auction. We will also investigate and compare the agent's effectiveness against Time Series and Simple Exponential Function in predicting the closing price of online auction. In Section 2, the design of Grey Theory is explained. Section 3 discusses the Grey Theory's prediction algorithm. Section 4 elaborates the design of the Simple Exponential Function Design. The

Time Series Prediction Model and ARIMA prediction algorithm are described in Section 5. In Section 6, the electronic simulated marketplace that we used in this work is described in detail. The experimental results are elaborated in Section 7 and finally the conclusion and future work is discussed in Section 8.

2 Grey Theory Design

The Grey System Theory [4] works on unascertained systems with partially known and partially unknown information by drawing out valuable information and also by generating and developing the partially known information where it helps in describing correctly and monitor effectively on the systemic operational behaviour [11]. Basically, the Grey System Theory was chosen based on colour [12]. For instance, "black" is used to represent unknown information while "white" is used to represent for complete information. Those partially known and partially unknown information is called the "Grey System Theory".

The Grey System Theory has been successfully applied to various fields and had made a success in analyzing uncertain systems that have multi-data inputs, discrete data, and insufficient data. Traditional prediction methods, such as time series, usually require a large amount of historical data and process a known statistical distribution in order to make an accurate assessment and prediction of the required parameters [7]. In contrast to the traditional prediction method, the main attributes of the grey theory, which is the core of the grey forecasting theory, are it does not need to make strict assumptions about the data set and is used successfully to analyse uncertain systems that have multi-data inputs, discrete data, and insufficient data. These simplify data collection and allow for timely predictions to be made. Grey Systems Theory explores the law of subject's motivation using functions of sequence operators according to information coverage. It is different from fuzzy logic since it emphasizes on objects with definite external extensions and vague internal meanings. Table 1 shows the Grey Prediction Model compared to other traditional forecasting models [2]. It can be seen that this model only requires short-term, current and limited data in order to predict a given value.

Grey prediction is a quantitative prediction based on Grey generating function, $GM(1,l)$ model, which uses the variation within the system to find the relations between sequential data and then establish the prediction model. The

Table 1. Attributes of Traditional Forecasting Methods

Mathematical model	Minimum observation	Type of sample	Sample interval	Mathematical requirements
Simple exponential function	5-10	Interval	Short	Basic
Regression analysis	10-20	Trend	Short	Middle
Casual regression	10	Any type	Long	Advanced
Box-Jenkins	50	Interval	Long	Advanced
Neural network	Large number	Interval or not	Short	Advanced
Grey prediction model	4	Interval	Long	Basic

Grey Forecasting Model is derived from the Grey System, in which one examines changes within a system to discover a relation between sequence and data. After that, a valid prediction is made to the system. Grey Prediction Model has the following advantages: (a) It can be used in situations with relatively limited data down to as little as four observations, as stated in Table 1. (b) A few discrete data are sufficient to characterize an unknown system. (c) It is suitable for forecasting in competitive environments where decision-makers have only accessed to limited historical data [3].

3 Grey Theory Prediction Algorithm

In this section, we describe our predictor agent algorithm which focuses on the Grey generating function, GM used in grey prediction [12]. The algorithm of $GM(1,1)$ can be summarized as follows.

Step 1. Establish the initial sequence from observation data. In this case, the data used is the previous values of the online auction closing price observed over time.

$$f^0 = \{f_1^0, f_2^0, f_3^0, ..., f_n^0\},\ where\ n \geq 2 \tag{1}$$

Step 2. Generate the first-order accumulated generating operation (AGO) sequence

$$f^1 = \{f_1^1, f_2^1, f_3^1, ..., f_n^1\},\ where\ f_t^1 = \sum_{k=1}^{t} f_k^0\ \ and\ \ t = 1, 2, ..., n. \tag{2}$$

Step 3. The grey model $GM(1,1)$

$$f_{t+1}^0 = a[-\frac{1}{2}(f_{t+1}^1 + f_t^1)] + b \quad \forall t \geq 1 \tag{3}$$

Step 4. Rewrite into matrix form

$$\begin{bmatrix} f_2^0 \\ f_3^0 \\ \vdots \\ f_n^0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(f_2^1 + f_1^1) & 1 \\ -\frac{1}{2}(f_3^1 + f_2^1) & 1 \\ \vdots & \vdots \\ -\frac{1}{2}(f_n^1 + f_{n-1}^1) & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \tag{4}$$

Step 5. Solve the parameter a and b

$$\begin{bmatrix} a \\ b \end{bmatrix} = (B^T B)^{-1} B^T F^0,\ where\ F^0 = \begin{bmatrix} f_2^0 \\ f_3^0 \\ \vdots \\ f_n^0 \end{bmatrix}\ and\ B = \begin{bmatrix} -\frac{1}{2}(f_2^1 + f_1^1) & 1 \\ -\frac{1}{2}(f_3^1 + f_2^1) & 1 \\ \vdots & \vdots \\ -\frac{1}{2}(f_n^1 + f_{n-1}^1) & 1 \end{bmatrix} \tag{5}$$

Step 6. Estimate AGO value

$$\hat{f}_{t+1}^0 = \left[f_1^0 - \left(\frac{b}{a}\right)\right] e^{-at} + \left(\frac{b}{a}\right),\ \forall t \geq 1 \tag{6}$$

Step 7. Get the estimate $IAGO$ value or the estimated closing price for a given auction.

$$\hat{f}_t^0 = \hat{f}_t^1 - \hat{f}_{t-1}^1, \ \forall t \geq 2 \tag{7}$$

Step 8. We use the average residual error for each set of data to calculate the accuracy of the predicted data. The formula for the average residual error (ARE) is given as

$$\left(\frac{1}{n} \sum_{t=1}^{n} \right) \frac{|f_t^0 - \hat{f}_t^0|}{f_t^0} \times 100\% \tag{8}$$

where f_t^0 =real value of exchange rate at time t
$\hat{f}_t^0$ = estimated value of exchange rate at time t
n = total observation

4 Simple Exponential Function

The exponential function is one of the most important functions in mathematics. The application of this function to a value x is written as $exp(x)$. Equivalently, this can be written in the form e^x. The exponential function is nearly flat (climbing slowly) for negative values of x, climbs quickly for positive values of x, and equals 1 when x is equal to 0. Its y value always equals the slope at that point. As a function of the real variable x, the graph of $y = e^x$ is always positive (above the x axis) and increasing (viewed left-to-right). It never touches the x axis, although it gets arbitrarily close to it (thus, the x axis is a horizontal asymptote to the graph). Sometimes, especially in the sciences, the term exponential function is more generally used for functions of the form ke^x. In general, the variable x can be any real or complex number. The best thing about exponential functions is that they are so useful in real world situations. Exponential functions are used to model populations, carbon date artifacts, help coroners determine time of death, compute investments, as well as many other applications.

5 Time Series Design

A Time Series design is a sequence of data collected from some systems by sampling a system property, usually at regular time intervals [8]. The analysis of time series has a long and rich history with recorded examples which had gone back more than a millennium [13]. They appear in such diverse fields as astronomy, meteorology, seismology, oceanography, signal processing, plant operations and economics among others [8]. Time series models may for instance be used for forecasting, option pricing and risk management [9].

Models for Time Series data can have many forms. Three broad classes of practical importance are the autoregressive (AR) models, the integrated (I) models, and the moving average (MA) models. These three classes depend linearly on previous data points and are treated in more detail in the articles autoregressive moving average models (ARMA) and autoregressive integrated moving average

(ARIMA). In the past, conventional statistical techniques such as ARIMA models have been extensively used for forecasting [5]. The application of the ARIMA methodology used for the study of time series analysis is due to Box and Jenkins [6]. The model is generally referred to as an ARIMA (p, d, q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively [15]. ARIMA model formula is given as follow:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \tag{9}$$

where y_t = Value forecasting at time t
ϕ_p = Parameter for autoregressive model at order p
δ_q = Parameter for autoregressive model at order p
ε_t = Residual at time t

6 The Electronic Simulated Marketplace

In order to test the effectiveness of our predictor agent, we compare its performance against the performance of an agent that utilizes a Simple Exponential Function and another agent that uses ARIMA Time Series. An electronic simulated marketplace was set up to simulate the real online auctions environment. This simulated electronic marketplace consists of a number of auctions that run concurrently. There are three types of auctions running in the environment: English, Dutch and Vickrey. The English and Vickrey auctions have a finite start time and duration generated randomly from a standard probability distribution, the Dutch auction has a start time but no pre-determined end time. At the start of each auction (irrespective of the type), a group of random bidders are generated to simulate other auction participants. These participants operate in a single auction and have the intention of buying the target item and possessing certain behaviours. They maintain the information about the item they wish to purchase, their private valuation of the item (reservation price), the starting bid value and their bid increment. These values are generated randomly from a standard probability distribution. Their bidding behaviour is determined based on the type of auction that they are participating in. The auction starts with a predefined starting value; a small value for an English auction and a high value for a Dutch auction. There is obviously no start value for a Vickrey auction. The marketplace is flexible and can be configured to take up any number of auctions and any value of discrete time. All the auctions are assumed to be auctioning the item that the consumers are interested in and all auctions are selling the same item.

7 Experimental Evaluation

We use three different agents that utilize three different techniques to predict the closing price of the online auctions in the simulated marketplace. The first one is the Grey Theory Agent that uses Grey Theory to predict the closing price. The second one is the Time Series agent that forecasts the closing price using the ARIMA (1, 0, 1). The third agent is the Simple Exponential Agent that uses the Simple Exponential Function technique. Here, we would like to compare the performance of our predictor agent (Grey Theory Agent) against agent (Simple Exponential Agent) that uses a minimum number (4 - 10 observations) of data and an agent that uses a large number (10 - 100 observations) of data (Time Series Agent). Using the simulated marketplace, we ran the auction from t = 1 until t = 150. We have also set most of auctions to close after t = 30. In one particular run, the closing price history for all auctions running in a marketplace are shown from t = 41 until t = 150. From t = 41 onward, consistency of obtaining the closing price at each t is preserved. The accuracy of predicted values by the three agents is measured using the ARE.

Table 2. Performance of the Three Agents

Time (t)	Original Data	(GT) (6 Historical Data)	(SEF) (5 Historical Data)	ARIMA (1, 0, 1) (40 Historical Data)
101	79	79.52	79.59	80.12
102	80	79.55	79.10	80.06
103	77	78.10	78.61	80.01
104	78	78.17	78.13	79.98
105	76	76.76	77.65	79.95
ARE (%)		**0.77**	**1.26**	**2.63**

The result of the agents' performance is shown in Table 2 in which the best ARE recorded for each agent is compared. In this particular experiment, it was observed that the Grey Theory Agent performs well using 6 historical data, the Simple Exponential Agent uses 5 historical data whilst the Time Series Agent performs best with 40 historical data. Based on these results, we can conclude that the Grey Theory Agent, produced better result (ARE = 0.77%) than Simple Exponential Agent (ARE = 1.26%) and Time Series Agent (ARE = 2.63%). Even with 6 historical data, the Grey Theory Agent is able to predict a more accurate auction's closing price in our simulated auction environment.

Figure 1 shows how far off the three predictions against the actual closing price. The predicted values by the Grey Theory Agent follow the trend of original data very closely whereas the predicted values for both the Time Series Agent and the Simple Exponential Agent are almost constant. In this particular run, the Simple Exponential Agent gives more accurate prediction than the Time Series Agent when only 5 historical data were used. This means that the Simple Exponentiation Agent is able to predict data using a minimal number of historical. However, the Grey Theory Agent was able to predict a more accurate closing price.

As mentioned earlier, at every time steps, there are auctions that will be closing. This simply means that, if one wants to predict the future data, one has to take into account the auctions that will be closing between the current time

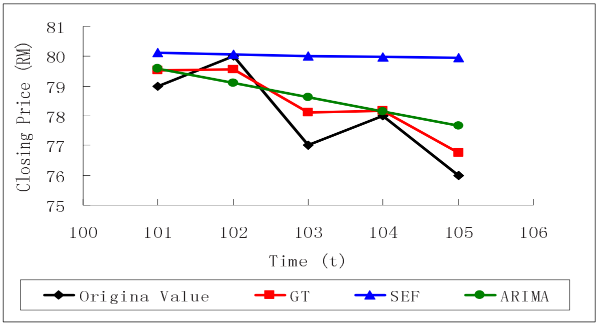


Fig. 1. Results Obtained by the Three Agents Over Time

and the next time steps. Hence, we performed the following experiments in which we compared the result based on moving historical data. Table 3 shows the result obtained by using 6 moving historical data for the Grey Theory Agent, 5 moving historical data for the Simple Exponential Agent and 40 moving historical data for the Time Series Agent compared with the original data generated by the simulated auction. These values are used because of their performance in the previous experiment. The value on the right hand side box is the ARE ranking of each individual agent. By using ARE ranking value, we can clearly see the sequence of each predicted value for different agents.

Table 3. Result Obtained by the Three Agents for Moving Historical Data

Time (t)	Original Data	(GT) Forecast 6 Moving Data (ARE %)	ARE Ranking	(SEF) Forecast 5 Moving Data (ARE %)	ARE Ranking	(ARIMA) Forecast 40 Moving Data (ARE %)	ARE Ranking
102	80	79.55 (0.56)	2	78.23 (2.21)	3	79.85 (0.19)	1
103	77	78.12 (1.45)	1	79.32 (3.01)	2	79.87 (3.73)	3
104	78	78.60 (0.77)	1	76.93 (1.37)	2	79.37 (1.76)	3
105	76	77.41 (1.86)	1	77.50 (1.97)	2	79.24 (4.26)	3
106	79	78.45 (0.70)	2	75.42 (4.53)	3	78.75 (0.32)	1
107	80	80.02 (0.03)	1	77.29 (3.39)	3	79.09 (1.14)	2
108	82	82.24 (0.29)	1	79.95 (2.50)	2	79.50 (3.05)	3

Figure 2 shows the predicted values over time for the three agents using fixed historical data and moving data. It can be observed that the values predicted by the Grey Theory Agent is still able to follow the real value very closely followed by the Simple Exponential Agent. Based on these experiments it can be concluded that the Grey Theory Agent is able to predict the closing price of an auction more accurately when compared to the other two agents even with a minimal number of historical data. It was also shown that the Grey Theory Agent is able to predict a value that is very close to the real value over time by making use of moving historical data. It has consistently outperformed the other two agents in terms of ARE when predicting the values either using fixed historical data or moving historical data. In the context of online auction, user

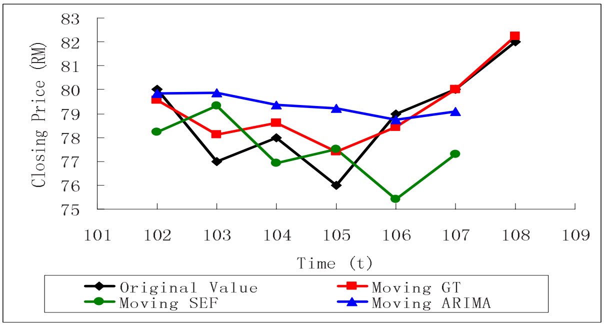


Fig. 2. Predicted Values for the Three Agents using Fixed Historical Data and Moving Data

may only be able to access or view several values from past history, so in this case the Grey Theory Agent can be used to make a quick prediction on the closing price of a given auction.

8 Conclusion and Future Work

This paper elaborated on the predictor agents that make use of the Grey Theory Agent, the Simple Exponential Agent and the Time Series Agent to predict the closing price of an online auction. It has been shown that using all three methods, the accuracy rate always exceeds 90%. The Grey Theory Agent gives better result when less input data are available while the Time Series Agent can be used with the availability of a lot of information. The experimental results also showed that using moving historical data produces higher accuracy rate than using fixed historical data for all three agents. This is important since, bidders in an online auction need to take into accounts all the auctions that are going to close within the prediction period. This closing price knowledge can then be used by the bidder to decide which auction to participate, when and how much to bid. This information will also allow the bidder to maximize his chances of winning in an online auction.

For future work, we will investigate how we can combine the Grey Theory with Time Series in our predictor agent by taking into account the amount of information that is available at any given time. Apart from that, we would also look into assessing and comparing the performance of Grey Theory with other AI techniques such as Artificial Neural Network. Finally, we would also like to apply our prediction method to predict on the real auction data in eBay and other online auctions.

Acknowledgments. The author wishes to acknowledge Ministry of Science and Technology Malaysia (MOSTI) for funding this project.

References

1. Anthony, P., Jennings, N.R.: Agent for Participating in Multiple Online Auctions. ACM Transaction on Internet Technology 3(3), 1–32 (2003)
2. Chiang, J.S., Wu, P.L., Chiang, S.D., Chang, T.I., Chang, S.T., Wen, K.L.: Introduction to Grey System Theory. Gao-Li Publication, Taiwan (1998)
3. Chiou, H.K., Tzeeng, G.H., Cheng, C.K., Liu, G.S.: Grey Prediction Model for Forecasting the Planning Material of Equipment Spare Parts in Navy of Taiwan. In: Proceedings of the World Automation Congress, pp. 315–320. IEEE XPlore, New York (2003)
4. Deng, J.: Control Problem of Grey System. System Control Letter 1(1), 288–294 (1982)
5. Delurgio, S.: Forecasting: Principles and Applications. McGraw-Hill, New York (1998)
6. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis Forecasting and Control, 3rd edn. Prentice Hall, New Jersey (1994)
7. Wu, H.H., Liao, A.Y.H., Wang, P.C.: Using Grey Theory in Quality Function Deployment to Analyse Dynamic Customer Requirements. The International Journal of Advanced Manufacturing Technology 25(11) (2005)
8. Nunn, I., White, T.: The application of antigenic search techniques to time series forecasting. In: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, pp. 353–360. ACM, New York (2005)
9. Aas, K., Dimakos, X.K.: Statistical modelling of financial time series: An introduction. Norwegian Computing Center, Norway (2004)
10. Chen, L.R., Lin, C.H., Hsu, R.C., Ku, B.G., Liu, C.S.: A study of Li-ion battery charge forecasting using Grey theory. In: The 25th International Telecommunications Energy Conference, pp. 744–749. IEEE XPlore, New York (2003)
11. Lin, Y., Liu, S.: A historical introduction to grey systems theory. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, pp. 2403–2408. IEEE XPlore, New York (2004)
12. Liu, S., Lin, Y.: Grey Information: Theory and Practical Applications. Springer, London (2005)
13. Kendall, M., Ord, J.: Time Series. Griffin, London, and Oxford University Press, New York (1993)
14. McAfee, R.P., McMillan, J.: Auctions and Bidding. Journal of Economic Literature 25(2), 699–738 (1987)
15. Pindyck, R.S., Rubinfeld, D.L.: Econometric Models and Economic Forecasts. McGraw-Hill/Irwin, New York (1997)
16. Yuan, B., Guo, J., Tang, G., Wang, L.: Using grey theory to predict the gas-in-oil concentrations in oil-filled transformer. In: Proceedings of the 6th International Conference on Properties and Applications of Dielectric Materials, pp. 217–219. IEEE XPlore, New York (2000)

Discovering Behavior Patterns in Muti-agent Teams

Fernando Ramos and Huberto Ayanegui

ITESM Campus Cuernavaca
Reforma 182-A, Col Lomas De Cuernavaca,
62589, Temixco Morelos, Mexico
{fernando.ramos, huberto.ayanegui}@itesm.mx

Abstract. Three of the most important open research areas in multi-agent cooperative systems are the construction of models related with the communication between agents, the multiple interactions and the behaviors adopted by agents during a task [5]. This work deals with the discovery of behaviors in multi-agent teams, more precisely in soccer teams. It is an extension of a previous work presented in [1]. The extension is focused on the discovery of tactical plays adopted by soccer-agents during a match within the context of formations. Due to the nature of team work in soccer-agent domains, the discovery of tactical behaviors should be done within the context of team formations. Nevertheless, the dynamic nature and the multiple interactions between players at each instant of the game difficult the tracking of formations, which at the same time difficult the discovery of tactical plays. In this work is proposed an original and efficient way of discovering tactical plays supported by a robust tracking of formations, even though they are submitted to dynamic changes of the world, based on the construction of topological structures. Successful results, derived from the analysis of an important number of matches of the RoboCup Simulation league matches, valid the efficiency of the model presented in this work.

Keywords: Soccer-agents, behavior pattern recognition, tactics.

1 Introduction

Soccer-agent domains are characterized by dynamic environments with multiple interactions between agents. These conditions difficult strongly the construction of models capable of discovering behaviors of teams playing soccer matches [11,12]. It can be distinguished several levels of behaviors of players during a match: individual behaviors, tactical behaviors, where two or more players are involved and formations which are related with the whole team.

On the one hand, a team that presumes to play by following certain strategies should play under the context of formations to assure order, discipline and organization during a match [10]. On the other hand, tactical plays occur, most of the time, under the context of formations. A formation is a structure based on the positions and a set of relations established between the soccer-agents. The

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 391–400, 2008.

formation represents the relation among defenses, midfielders and forwards of a team, for instance, 5:3:2 (five defenders, three mid-fielders, two forwards).

A current interest of research in soccer-agent domains is the discovery of formations and tactical plays of a team under observation. This information can be very useful by the opponent team to reduce the performance of the observed team.

Nevertheless, due to the dynamic nature of a match and the multiple interactions between players the discovery of formations, and consequently of tactical plays, is a very difficult task. In a previous work [1], we proposed a model based on the construction of topological structures able to track correctly the formations and recognize real changes even though they are submitted to the dynamic conditions and multiple interactions such as those played in soccer-agent matches. The main contribution of this work is the discovery of tactical behavior patterns supported by the topological structure model. The topological structure is represented by a planar graph which in turn represents relevant relations such as: the relation between agents that belong to the same zone; the relation between agents that belong to neighbor zones. The planar property of the graph is monitored during the following simulation cycles. The intersection of two edges of the graph breaks the planar property of the graph [2] indicating a possible change of formation.

The tactical plays to be discovered are mainly characterized by the path of the ball, the position of players involved in such plays and the zones associated with the plays. The topological structures of formations in this case serve to track the players involved in a tactical play at each instant of the match. The test domain for this research is the simulated robotic soccer, specifically, the Soccer Server System [9], used in the Robot World Cup Initiative [8]. Tactical plays within offensive formations have been discovered. Successful results, supported by the experimental work, valid the efficiency of the model.

2 Related Works

Bezek and Bratko [3] present a method to discover pass patterns incorporating domain knowledge and providing a graphic representation for detected strategies. This work is focused on the increase of human comprehension. Although their approach obtains tactical behavior patterns, they only consider the players involved in the passes without taking into account the notion of team behaviors related with formations.

The work from Kaminka et al. [7] addresses the problem of modeling the behavior of a team. The main idea is to take time-series of continuous observations and then parse and transform them into a single-variable time-series. The authors use a set of behavior recognizers that focus only on recognizing simple and basic behaviors of the agents (e.g. pass or dribble). The data are then represented in a trie (a tree-like data structure) to support two statistical methods: (a) frequency counting and (b) statistical dependency detection. Experiments showed that the latter method is more suitable to discover sequential behavior.

The Kaminka's method did neither consider the variants of behavior patterns nor the formation behavior of a team. Raines and colleagues [13] presented a system called ISAAC which analyzes a game in an off-line mode using a decision tree method to generate rules about the success of players in situation of shooting to goal. It tries to discover patterns in each level based on 'key events'. Key events are events that affect potentially the result of the game. Some differences between ISAAC and our work are: we build a model of a team based on behavior patterns, independently of success or failures events; the global behavior discovered by ISAAC is not based on the formation of a team, like ours, but in a general performance of the team. Thus, ISAAC is unable to discover the tactical behavior of a team and ISAAC does not integrate formation models.

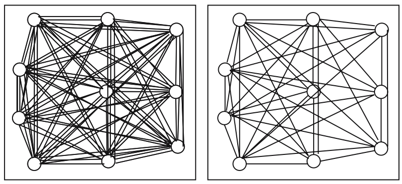
Devaney and Ram in [4] are focused on the problem of identifying particular repeated patterns of movement rather than trying to select a particular model which captures the movements of all of the agents. Such approach did not provide enough information to infer structures of formations. Visser and colleagues [14] recognized the formation of the opponent team using a neural networks model. In the Visser's work observed player positions were the input of a neural network. The output was a predefined set of formations. If a classification can be done, the appropriate counter-formation is looked up and communicated to the players. The main difference with our approach is that Visser and colleagues did not represent relations between players. As Visser mentioned in his work [14], his approach is unable of tracking the changes of formations. This is because the lack of structures due to the absence of relations between players. Kuhlmann, Stone and Lallinger [10] learned team formations using home positions. As they mentioned, they modeled the formation as a home position (x,y) and introduce a ball attraction vector (Bx,By) for each player. The X and Y values are calculated as the average x and y coordinates of the observed player during the course of the game. Values for BX and BY were handpicked for each position and were found through brief experimentation. A weakness of Kuhlmann's work is that there were cases where the ball attraction caused the forwards to play too far to the opponent goal; then, they adjusted the home positions manually.

3 Discovering Behavior Patterns

This section presents an efficient model able to discover tactical behaviour patterns under the context of formations. Firstly, an overview of the topological structure model oriented to track team formations is presented. The topological structures of formations in this case serve to track the players involved in a tactical play at each instant of the match. Secondly, the model of discovering tactical behaviour patterns is presented.

3.1 Topological Structure Model

The relations between soccer-agents occur through the multi-interactions taking place between them at each instant of the game. For instance, an agent A can be related with the agent B because they interact through the exchange of passes to accomplish tactical plays. Another example of interactions occurs when several agents perform together a tactical play. In this situation, some of the relevant relations are determined by the structure that represents the distribution of soccer-agents accompanying the play. Thus, the relations construct a structure that can be represented by a graph, where nodes represent the position of soccer agents and arcs the relations between them.



(a) (b)

Fig. 1. All posible relations between players of a soccer team. (*a*) 90 relations and (*b*) 45 relations based on a specific formation.

Figure 1(a) shows the relations of soccer-agents for a 4:3:3 formation (four defenders, three mid-fielders, and three attackers). A total of 90 relations are obtained given by the formula: $n(n-1)$, where n represents the number of players. This formula considers two relations by each pair of players. For instance, player A is related with player B and player B is related with player A. Then, the same relation is repeated two times. Thus, one of these relations should be eliminated, in such a way that the 90 relations are reduced to 45. Figure 1(b) illustrates these 45 relations. Even that the number of relations has been reduced it is still a complicated model unable to track formations and tactical plays successfully, it would be desirable to dispose of a graph model capable of tracking correctly the formation even though the dynamic conditions and the multiple interactions. Otherwise, if the graph model is very sensible to the changing conditions mentioned before, apparent changes could be interpreted as real changes. This work uses a topological structure model that represents relevant relations. A relevant relation is one that soccer-agent uses to exchange passes and positions in a strategic way. Thus, an agent will be related with his closer neighbor belonging to his zone (defensive, medium and attack) and his closer neighbor belonging to the neighbor zone as illustrated in Figure 2(a) and Figure 2(b). Figure 2(c) shows the integration of both kinds of relations.

As it is shown, the graph representing the formation is composed of triangular sub-graphs, each node represents an agent and arcs represent relations between agents that belong to the same zone and relations between agents belonging to neighbor zones. Due to the topological nature of these graphs, they are enough robust to support the changing conditions avoiding the apparent changes.

Tactical plays can be represented by a sequence of graphs where agents are represented by nodes and arcs by the relations between them. These graphs are

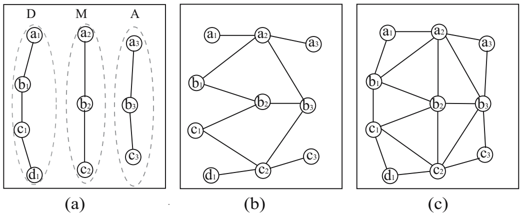


Fig. 2. (a) Neighbor nodes of the same zone are linked. (b) Neighbor nodes fo neighbor zones are linked. (c) Topological graph including both kind of links.

necessary sub-graphs of the graph representing a formation under which the tactical play occurs.

3.2 Tactical Behavior Patterns

The model to discover tactical behavior patterns proposed in this work is presented in Figure 3, which is composed of: The extraction of a set of similar paths from real matches; the set of paths are coded to be represented by a sequence of numbers using the code of Freeman; the coded sequences using the code of Freeman are then recoded to obtain a more abstract code; a method based on a generalization of a tree is applied to discover the general behavior patterns representing the paths of tactical offensive plays; finally the players and zones are associated with paths by using the topological structures of formations.

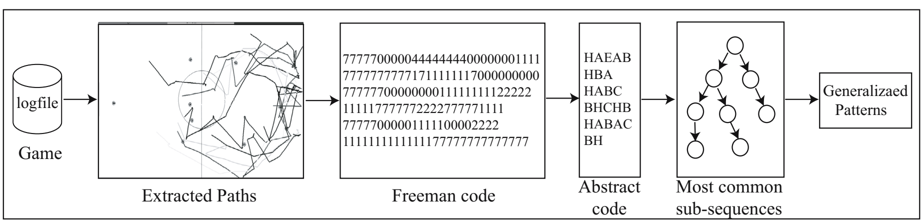


Fig. 3. The steps to discover the tactical behavior patterns

Extracting similar paths. The analysis of similar paths of the ball can lead us to discover promising tactical behavior patterns. A set of paths are similar if they occur under similar contexts. Although this work is focused on the analysis of attack situations, this model could be extended to defensive ones. In particular, paths arriving close to the opposite goal reveal tactical plays of attacking. Usually paths do not start at the same field zone but the reached zone is more relevant that the initial one.

Freeman codification. The set of extracted paths are coded to be represented by a sequence of orientations using a Freeman codification [6] which is composed of

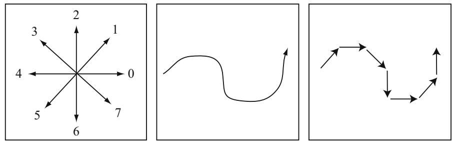


Fig. 4. Freeman code

eight orientations shown by Figure 4(a). Such representation enables an efficient management of the paths reducing the whole complex space into eight possible orientations. For instance, Fig. 4(c) shows the Freeman code representation which corresponds to the path shown in Fig. 4(b). In this case, the Freeman code is 1-0-7-6-0-1-2.

Given that a path is characterized by a sequence of segments, similar shapes of the paths tend to be different using the Freeman codification. That is, two paths coded as 11110000000000002222 and 1111111100002222222222 have the same sequence of segments: first a sequence of 1's, second a sequence of 0's and finally, a sequence of 2's. However, representations in Freeman code seem to be different in despite of holding the same shape. To cope with this problem, Freeman code paths are recoded into more abstract codes representing the shapes of paths. Let $\{A,B,...,H\}$ be a set of abstract segments where each element of this set represents a sequence of codes, based on the Freeman codification, with the same orientation, such that, A represents a sequence of 0's, B represents a sequence of 1's, and so on. Thus, a path coded as 1111111100002222222222 can be represented by the code BAC. Therefore, this new codification represents the shape of paths without taking into account the length of them, facilitating the matching of different potential tactical plays.

*Identifying most common sub-sequences.*In pursuit of discovering generalized patterns, an augmented suffix trie [7,9] is used. *A* trie is a tree-like data structure which efficiently stores sequences such that duplicated sub-sequences are stored only once, but in a way that allows keeping a count of how many times they had appeared. In the trie, every node represents a trajectory in the abstract code discussed before and the node's children represent trajectories that have appeared following this trajectory. Thus, a path from the root to a node represents a sub-sequence that has appeared in at least one, but possibly more than one, path of the ball. The next example has been taken from [7]: Let's take two paths: *ABCA* and *BCAD*. Let's suppose that the trie is empty. It would first insert *ABCA* into it, resulting in a trie in Figure 5(a). It would then insert the three remaining suffixes of *ABCA*: $\{BCA, CA, A\}$, resulting in the trie in Figure 5(b). Next, it would then insert the next path and its suffixes: $\{BCAD, CAD, AD, D\}$, into the trie, resulting in the trie in Figure 5(c). Note that in this final trie, the most common single sub-sequence is *A*, the most common two sub-sequences is *CA* and the most common three sub-sequences is *BCA*.

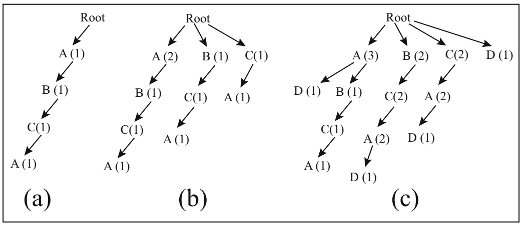


Fig. 5. A trie under construction. Numbers in parenthesis are the nodes counts.

Applying the steps mentioned before, the tactical behavior patterns have been obtained.

4 The Analysis of Results

In this section, important experimental results are analyzed. They are derived from two teams: The TsinghuAeolus soccer team, who won the Simulation RoboCup Championship in 2002. It is presented an analysis of the match between TsinguAeolus vs. Everest; And the WrightEagle team, who won the second place in the same competition that held in 2007. The model has been proven in nine matches, but for the relevance of the teams, we present the analysis of results of two matches, one for the TsinguAeolus and one for the WrightEagle.

Figure 6 shows a sequence of sub-graphs that represents the structure involving the soccer-agents in a path of a tactical play. Because of the lack of space it is shown some of the sub-graphs that compose the total sequence of sub-graphs representing the path (in fact, there are approximately 50 sub-graphs for this tactical play). As can be seen, the shadowed sub-graphs contain the soccer agents involved in the tactical plays. They are in this case: the middle center, the right middle, the right forward, the center forward and the left forward.

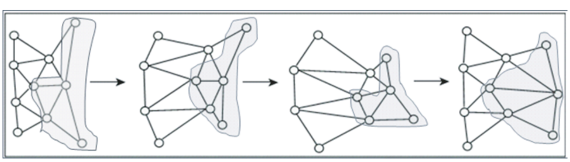
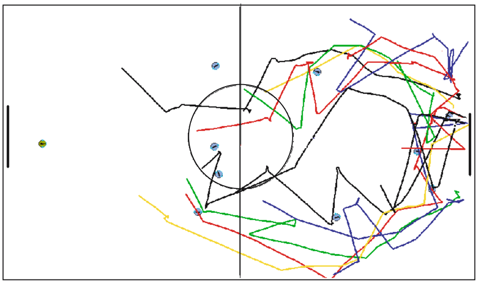


Fig. 6. Sequence of sub-graphs representing a tactical play incorporating agents and field zones

As first step, the paths of the ball were extracted to be analyzed and coded by the code of Freeman. Figure 7 shows the extracted paths and their corresponding Freeman code for TsinghuAeolus team by considering that players are approaching to the opposite goal. In this way the set of paths can be compared numerically by measuring the similarity between them. Another advantage of this codification is that we can have an idea about how long the paths are. However, what is interesting in this analysis is not exactly how long a path is, but, from the point of view of behavior, the form adopted by the path and obviously the properties associated with the intention or purpose of it, in this case to get close to a position of shooting to the goal.



(a)

7777770000000002444440000000000000001111111111 0011111
777770007777777777777777111111111111111111110000
77777700000000000000111111222222222222200000
111112677777777722222216777777111111111111107
777777777777777777777777000000000111111000000222222221
111111111111111111111111117777777777777777777777777
7777777777777777777777222221110000077777777777777
00000000000111111100066666666611111111000000000777777777
77777777774444444444111111111777777777777755555557777 7
7777775555555511111111111116666

(b)

Fig. 7. (a) Paths indicating approaches of the left team to the opposite goal. (b) The corresponding Freeman code of the paths shown in (a).

Due to these reasons, it is proposed in this work a more abstract representation. Then the paths coded by the code of Freeman have been recoded to obtain a more abstract code, as shown in Section 3.2. The paths represented by abstract codes have facilitated the application of the model to discover behavior patterns related with tactical plays. It is important to point out that similar paths are not necessary those to end in a goal, but those that assume a similar behavior from the start of the path to the final objective. Figure 8 illustrates two shapes of generalized paths of tactical behaviors played through the right and left side of the terrain. These generalized paths correspond to the TsinghuAeolus team.

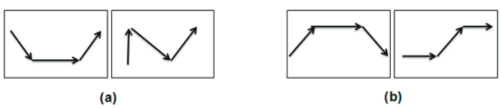


Fig. 8. Generalized paths of tactical behaviors: a) Attacks by right side and b) Attacks by left side

For the case of the WrightEagle team, they played in the right side, Figure 9 shows the extracted paths that get close to the opposite goal and Figure 10 shows two shapes of discovered generalized paths.

Based on the results obtained, it is observed that the model to obtain the paths representing the tactical plays do not depend on the analyzed team. The topological structures used to track formations have been a very good support to determine the players participating in tactical plays, as well as the zones through which the plays have taken place.

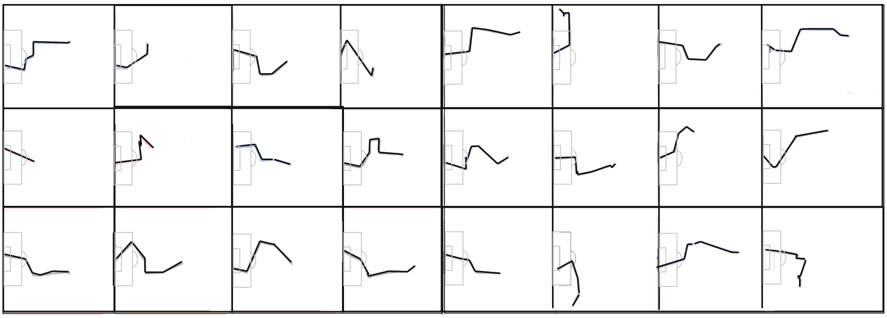


Fig. 9. Extracted paths that get close to the opposite goal. The team is attacking from left to right side.

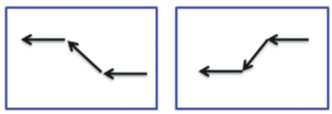


Fig. 10. Two shapes of generalized paths of tactical behaviors

5 Conclusion

This work has pointed out the importance of discovering tactical plays within the context of formations. Tactical plays were characterized not only by a sequence of the ball from an initial to a final position. It is remarked in this work that soccer-agents and the zones through which the paths pass should be included. Therefore, a model able to track correctly the formations was, even though the changing environment nature of soccer-agent domains, quite desirable. The model of topological structures that supported the discovery of tactical plays has been determinant to extract the sequence of sub-graphs that contain the soccer-agent involved in the tactical plays.

The behavior of the paths representing tactical plays should be more related with the form of the paths than with the length of them. To cope with this problem, the original paths have been coded two times, using the code of Freeman, where the codification of a second path is more abstract than the first one. The difference between them is that the first one contains the length of the path and the second one the form of it. However, it is difficult to manipulate if the purpose is to discover generalized patterns. The second path is shorter and it represents only the form of the path. But it is easier to manipulate to discover the most common sequences or generalized pattern behaviors.

Offensive tactical plays were discovered in this work, a future work concerns the discovery of defensive tactical plays. In addition, the discovery of variants seems to be important if counter strategies are considered.

References

1. Ayanegui, H., Ramos, F.: Recognizing patterns of dynamic behaviors based on multiple relations in soccer robotics domain. In: Ghosh, A., De Rajat, K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 33–40. Springer, Heidelberg (2007)
2. Berge, C.: Graphes. Guathier-Villars (1983)
3. Bezek, A., Gams, M., Bratko, I.: Multi-agent strategic modeling in a robotic soccer domain. In: Nakashima, H., Wellman, M.P., Weiss, G., Stone, P. (eds.) AAMAS, pp. 457–464. ACM Press, New York (2006)
4. Devaney, M., Ram, A.: Needles in a haystack: Plan recognition in large spatial domains involving multiple agents. In: AAAI/IAAI, pp. 942–947 (1998)
5. Ferber, J.: Multi-Agent System: An Introduction to Distributed Artificial Intelligence. Addison-Wesley Longman, Amsterdam (1999)
6. Freeman, H.: On the encoding of arbitrary geometric con?gurations. In: IRE Transactions, editor (1973)
7. Kaminka, G.A., Fidanboylu, M., Chang, A., Veloso, M.M.: Learning the sequential coordinated behavior of teams from observations. In: Kaminka, G.A., Lima, P.U., Rojas, R. (eds.) RoboCup 2002. LNCS (LNAI), vol. 2752, pp. 111–125. Springer, Heidelberg (2003)
8. Kitano, H., Tambe, M., Stone, P., Veloso, M.M., Coradeschi, S., Osawa, E., Matsubara, H., Noda, I., Asada, M.: The robocup synthetic agent challenge 97. In: Kitano, H. (ed.) RoboCup 1997. LNCS, vol. 1395, pp. 62–73. Springer, Heidelberg (1998)
9. Knuth, D.E.: The Art of Computer Programming. Addison-Wesley, Reading (1973)
10. Kuhlmann, G., Stone, P., Lallinger, J.: The ut austin villa 2003 champion simulator coach: A machine learning approach. In: Nardi, D., Riedmiller, M., Sammut, C., Santos-Victor, J. (eds.) RoboCup 2004. LNCS (LNAI), vol. 3276, pp. 636–644. Springer, Heidelberg (2005)
11. Lattner, A.D., Miene, A., Visser, U., Herzog, O.: Sequential pattern mining for situation and behavior prediction in simulated robotic soccer. In: Bredenfeld, A., Jacoff, A., Noda, I., Takahashi, Y. (eds.) RoboCup 2005. LNCS (LNAI), vol. 4020, pp. 118–129. Springer, Heidelberg (2006)
12. Nair, R., Tambe, M., Marsella, S., Raines, T.: Automated assistants for analyzing team behaviors. Autonomous Agents and Multi-Agent Systems 8(1), 69–111 (2004)
13. Raines, T., Tambe, M., Marsella, S.: Automated assistants to aid humans in understanding team behaviors. In: Veloso, M.M., Pagello, E., Kitano, H. (eds.) RoboCup 1999. LNCS (LNAI), vol. 1856, pp. 85–102. Springer, Heidelberg (2000)
14. Visser, U., Drcker, C., Hbner, S., Schmidt, E., Weland, H.-G.: Recognizing formations in opponent teams. In: Stone, P., Balch, T.R., Kraetzschmar, G.K. (eds.) RoboCup 2000. LNCS (LNAI), vol. 2019, pp. 391–396. Springer, Heidelberg (2001)

Fuzzy Logic for Cooperative Robot Communication

Dingyun Zhu and Tom Gedeon

Department of Computer Science, The Australian National University, Acton, Canberra, ACT 0200, Australia
DingyunZhu@gmail.com, Tom.Gedeon@anu.edu.au

Abstract. This paper proposes a new approach which applies a recently developed fuzzy technique: Fuzzy Signature to model the communication between cooperative intelligent robots. Fuzzy signature is not only regarded as one of the key solutions to solve the rule explosion in traditional fuzzy inference systems, but also an effective approach for modeling complex problems or systems with a hierarchical structure. Apart from the application of fuzzy signatures, another modeling structure of pattern-matching with possibility calculation is designed for the further intentional inference of cooperative robot communication. By the combination of these two theoretical issues, a codebook for intelligent robot decision making has been developed, as well as its implementation - a Cooperative Robot Communication Simulator.

Keywords: Fuzzy Logic; Cooperative Robots; Codebook; Fuzzy Signature; Possibility Calculation.

1 Introduction

Scenario of Co-operating Intelligent Robots [8]: There is a set of identical oblong shaped tables in a room. Various configurations can be built from them, such as a large U shape, a large T shape, a very large oblong, rows of tables, etc. A group of autonomous intelligent robots is supposed to build the actual configuration according to the exact instructions given to the "Robot Foreman" (R_0). The other robots have no direct communication links with R_0, but they are able to observe the behavior of R_0 and all others, and they all posses the same codebook containing all possible table configurations. The individual tables can be shifted or rotated, but two robots are always needed to actually move a table, as they are heavy. If two robots are pushing the table in parallel, the table will be shifted according to the joint forces of the robots. If the two robots are pushing in the opposite directions positioned at the diagonally opposite ends, the table will turn around the center of gravity. If two robots are pushing in parallel, and one is pushing in the opposite direction, the table will not move. Under these conditions the task can be solved, if all robots are provided by suitable algorithms that enable "intention guessing" from the actual movements and positions, even though they might not be unambiguous.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 401–410, 2008.

2 Fuzzy Signature

Fuzzy signature has been regarded as an effective approach to solve problem of rule explosion in traditional fuzzy inference systems: constructing characteristic fuzzy structures, modeling the complex structure of the data points (bottom up) in a hierarchical manner [6], [2], [9]. Fuzzy signatures result in a much reduced order of complexity, at the cost of slightly more complex aggregation techniques.

The original definition of fuzzy sets was $A : X \to [0,1]$, and was soon extended to *L-fuzzy sets* [3]:

$$A_S :\to [a_i]_{i=1}^{k}, a_i = \begin{cases} [0,1] \\ [a_{ij}]_{j=1}^{k_i} \end{cases}, a_{ij} = \begin{cases} [0,1] \\ [a_{ijl}]_{l=1}^{k_{ij}} \end{cases} \tag{1}$$

$A_L : X \to L, L$ being an arbitrary algebraic lattice. A practical special case, *Vector Valued Fuzzy Sets* was introduced by [5], where $A_{V,K} : X \to [0,1]^k$, and the range of membership values was the lattice of k-dimensional vectors with components in the unit interval. The general concept of fuzzy signature is a nested vector, where each vector component can be another nested vector structure. So it can be described as a generalized vectorial fuzzy set with possible recursive vectorial components, consequently, it is a generalization of valued fuzzy sets and denoted by [6]:

$$A : \ X \to S^{(n)}, \tag{2}$$

where $n \geq 1$ and

$$S^{(n)} = \prod_{i=1}^{n} S_i, \tag{3}$$

$$S = \begin{cases} [0,1] \\ S^{(m)} \end{cases} \tag{4}$$

and $\prod$ describes Cartesian product.

In fact, we can consider fuzzy signature as a special kind of multi-dimensional fuzzy data. Some of the dimensions are formed as a sub-group of variables, which jointly determine some feature on a higher level. Figure 1 illustrates an example of fuzzy signature structure.

3 Fuzzy Signatures Construction for Cooperative Robot Action Inference

The process of constructing fuzzy signature has also been discussed in [11]:

Let S_{S_0} denote the set of all fuzzy signatures whose structure graphs are sub-trees of the structural ("stretching") tree of a given signature S_0. Then the signature sets introduced on S_{S_0} are defined by:

$$A_{S_0} : X \to S_{S_0} \tag{5}$$

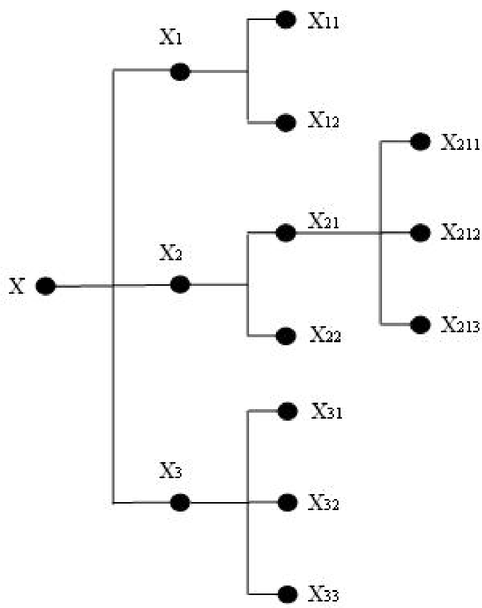


Fig. 1. Fuzzy Signature Structure

In this case, the prototype structure S_0 describes the "maximal" signature type that can be assumed by any element of X in the sense that any structural graph obtained by a set of repeated omissions of leaves from the original tree of S_0 might be the tree stretching the signature of some A_{S_0}.

In fact, there are two approaches to construct the sub-structures of the fuzzy signature, S_0 [1], [10], [11]:

1. Predetermined by a human expert in the field.
2. Determined by finding the separability from the data.

In our cooperative robots case, as we are handling complex circumstances and we actually do not have enough data, so we will only use the first approach to construct the fuzzy signatures. Based on the context of the robots scenario, we propose the use of an alternative form of fuzzy signature, which uses a better hierarchical structure where the internal nodes are simple, while the leaves are populated with small rule bases, generally of 1 variable. The effect is to retain the much reduced order of complexity, and to also substantially reduce the complexity of aggregations to simple combinations of basic fuzzy functions [7].

Before we start constructing the fuzzy signatures, we need to clarify some instructions and assumptions about the CRC framework:

1. *Instructions:*
   (a) A group of intelligent robots of size 1×1: $R_i : R_0, R_1, ..., R_n$, R_0 is the "foreman";

   (b) A set of random shape tables of size 1×2: $T_1, T_2, ..., T_n$;
   (c) A set of possible configurations made up of tables: $S_1, S_2, ..., S_n$, one of them is the final task.
2. *Assumptions:*
   (a) "Foreman" (R_0) represents a human-being (controlled by a human);
   (b) Only the "Foreman" (R_0) knows the final task;
   (c) Other robots (R_i) do not know the final task, but they know all the possible table shapes ($S_1, S_2, ..., S_n$);
   (d) Other robots (R_i) know who the foreman (R_0) is.

In order to construct the fuzzy signatures for inferring the foreman's following action, we need to figure out which "attributes" will be essentially related to foreman's intentional action based on the current situation. Since the current situation is that there are a set of tables, if the foreman is intended to do something, he should go and touch a particular table first or get closer at least. So the first "essential attribute" is the "Distance" between the foreman and each table in the environment. Figure 2 illustrates the membership function of "Distance". Actually, there exists a possible situation that can not be handled by "Distance"

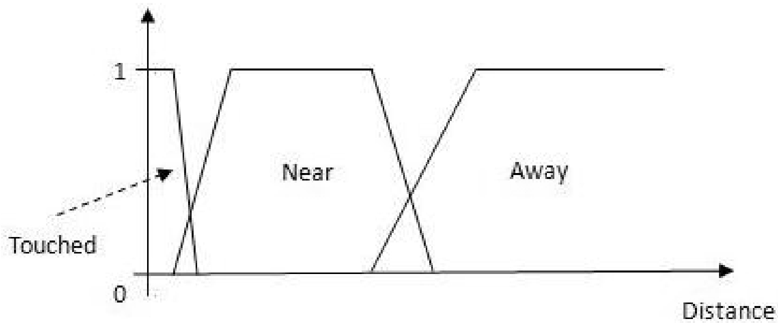
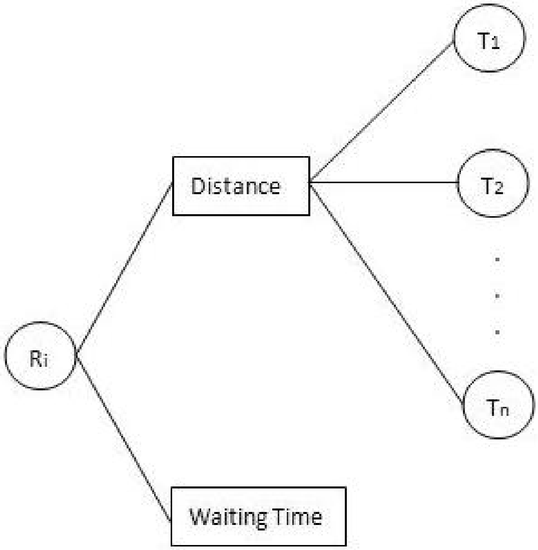


Fig. 2. Membership Function of Distance

only: if the foreman moves towards to a table then touches it, but after that he moves away or switches to another table immediately, the other robots still can not infer what the foreman is going to do. In order to solve this problem, we need to add another "essential attribute" called "Waiting Time" (the membership is similar in shape to Figure 2 and is not shown) which is used to measure how long a robot (R_i) stops at a particular spot. The reason why we need to measure the stopping time is that it is too difficult for a robot to perceive the meaning of the scene using instantaneous information (a snapshot) only [4].

By combining the "Waiting Time" with the previous item "Distance", the final fuzzy signatures for intention inference will be formed to the structure in Figure 3.

Under this circumstance, other robots will be able to infer the foreman's next action according to his current behavior. For instance, if the "Distance" between



$$R_i = \begin{bmatrix} Distance \begin{bmatrix} T_1 \\ T_2 \\ . \\ . \\ . \\ T_n \end{bmatrix} \\ WaitingTime \end{bmatrix}$$

Fig. 3. Fuzzy Signatures for CRC

the foreman (R_0) and a table (T_i) is *Touched*, meanwhile foreman's "Waiting Time" at that spot is *Long*, then it implies the foreman is "Waiting for Help" which means another robot (R_i) should go to T_i and help the foreman. Otherwise if neither of the condition is satisfied, which means other robots will not think the foreman is going to carry out any intentional action because they can not figure it out by observation of the foreman's current behavior.

4 Pattern Matching with Possibility Calculation

So far we have discussed the problem of inferring the foreman's intentional action by constructing the fuzzy signatures based on the foreman's current behavior. In some sense, it means other robots still have to count on the foreman completely and it actually does not show that these robots are intelligent enough that can help the foreman to finish the final task effectively and efficiently as well as truly reduce the cost of the communication between them.

In order to improve the modeling technique, it is important for us to consider the current situation after each movement of a table, which means other robots should be able to guess which table shape is supposed to be the most possible one according to foreman's previous actions and the current configuration of tables. The solution here is to measure how close the current table shape matches each of the possible shapes after the foreman's intentional actions. Therefore, apart from the previous fuzzy signatures, another modeling structure has been constructed for robot's further decision making (see Figure 4). The following figure shows another tree structure with all the leaves representing each possible table shape

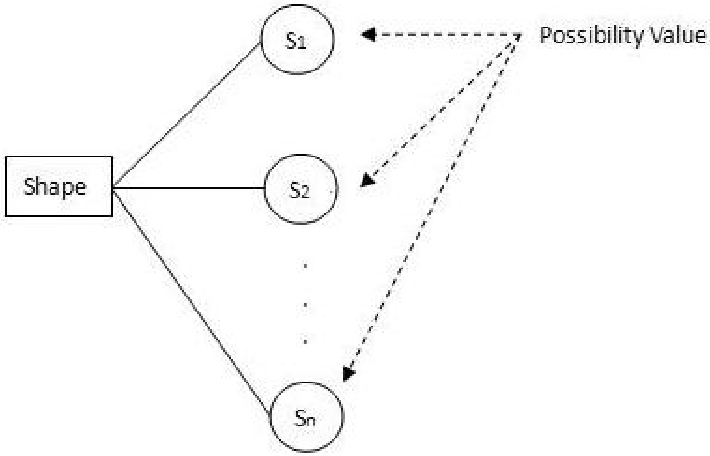


Fig. 4. Structure of Pattern Matching with Possibility Calculation

as well as its possibility value respectively. The following strategies show how this structure works:

We have a set of tables: $T_1, T_2, ..., T_n$; the total number of tables is n:

1. IF foreman and a robot push a table to a place which matches one of the possible table shapes: S_i;
   THEN increase the Possibility value of S_i: $PV_{S_i} + 1/n$;
2. IF foreman and a robot push a table to a place which does not match any of the possible table shapes;
   THEN none of the Possibility values will change;
3. IF foreman and a robot push a table which matched S_i to a place where does not match any of the possible shapes;
   THEN decrease the Possibility value of S_i: $PV_{S_i} - 1/n$;
4. IF foreman and a robot push a table which matched S_i to a place where matches another possible shape: S_j;
   THEN decrease the Possibility value of S_i: $PV_{S_i} - 1/n$;
   AND increase the Possibility value of S_j: $PV_{S_j} + 1/n$;
5. IF two robots (neither is foreman) push a table to a place where matches one of the possible table shapes: S_i;
   THEN the Possibility value of S_i, ie, PV_{S_i} will not change;

From the above strategies we can find that the possibility value of a possible shape S_i will only change when the foreman is one of the working robots who carry out the action, otherwise the possibility value will not change. The reason why we model the situation like this is due to the initial assumption mentioned that the foreman is the only robot who knows the final task so that we assume all the actions carried out by the foreman are directly related to the final task. Since other robots do not know the final task, their actions are not considered to be definitely correct and directly related to the final task so none of the possibility values will change according to these actions.

4.1 Codebook for Cooperative Robot Communication

1: IF Foreman (R_0)'s *Distance* to T_i is **NOT Touched** OR *Waiting Time* is **NOT Long** THEN Foreman (R_0) is **moving around or no action**:
- IF *Possibility Value* of S_i (PV_{S_i}) is the **highest** one THEN take S_i as the **final task**; AND calculate the distances between each table and S_i;
  - IF T_i's *Distance* to S_i is the **Shortest** THEN *Move* T_i *to* S_i;
    - IF Foreman (R_0) *Moves to* C_{ST} *(Stop)* THEN *Move back and Go to **2***;

2: IF Foreman (R_0)'s *Distance* to T_i is **Touched** AND *Waiting Time* is **Long** THEN Foreman (R_0) is **waiting for help**:
- IF R_i's *Distance* to Foreman (R_0) is the **Shortest** THEN the following action of R_i: *Move to* T_i *and Go to **3***;
  - ***3:*** Choose to *Shift* T_i (C_{SH})
    - IF R_0 *Moves away* THEN **Go back to *1***;
    - IF R_0 dose NOT carry out the action combination (*Waiting Time is Long*) THEN Choose to *Rotate* T_i (Go to ***4***);
    - IF R_0 carries out the action combination THEN **Keep Pushing**;
      - IF T_i's *stopping position* matches part of S_i THEN $PV_{S_i} + 1/n$;
      - IF T_i's *stopping position* does NOT match part of S_i THEN **Go back to *1***;
      - IF T_i's *initial position* matched part of another possible shape: S_j THEN $PV_{S_j} - 1/n$;
  - ***4:*** Choose to *Rotate* T_i (C_{CC} or C_{CW})
    - IF R_0 *Moves away* THEN **Go back to *1***;
    - IF R_0 does NOT carry out the action combination (*Waiting Time is Long*) THEN Choose to *Shift* T_i (Go to ***3***);
    - IF R_0 carries out the action combination THEN **Keep Rotating**;
      - IF T_i's *stopping position* matches part of S_i THEN $PV_{S_i} + 1/n$;
      - IF T_i's *stopping position* does NOT match part of S_i THEN **Go back to *1***;
      - IF T_i's *initial position* matched part of another possible shape: S_j THEN $PV_{S_j} - 1/n$;

5 Evaluation

The experiments we designed mainly focus on the difference between robots completely controlled by human-beings and robots working with the codebook, as well as how well these robots are able to cooperate with the foreman to finish a task in our CRC simulator. The following table is the basic setup for all the experiments.

Table 1. Basic Instructions for Experiments

Item	Description
Number of Tables	4
Test Cases ("Table Shapes")	1. Horizontal Rows (HR) 2. Vertical Rows (VR) 3. T Shape (T) 4. U Shape (U)
Test Times (Repetitions)	5
Robot's Speed	About 3 movements per second
Measurements	1. Number of robot steps 2. Number of table movements (Shifting or Rotating) 3. Time to finish a task

5.1 Experiment Description

Experiment 1: Two robots are manually controlled by two players. Players are allowed to have verbal communications.

Experiment 2: One robot with Codebook cooperates with the Foreman who is manually controlled by a player.

Experiment 3: Two robots with Codebooks cooperate with the Foreman who is manually controlled by a player. Foreman and one robot move one table to a place which fits into the final task, then the two robots finish the rest of the work.

5.2 Results of Experiments

Although we allowed players to have verbal communications in experiment 1, the human-controlled robots still took the most steps on average to finish each of the test tasks. The reason for this phenomenon is that players might have different decisions in dynamic situations. Therefore, it is possible for them to decide to move different tables at the same time rather than aiming at the same target, or placing the same table with different route plans, which will cost them extra steps to reach the common target or correct previous incorrect actions. That is, notwithstanding the explicit communication (talking) possible, it may be only

Table 2. Average Robots Steps, Table Movements and Time: 2 Humans

Experiment 1	Horizontal Rows	Vertical Rows	T Shape	U Shape
Robot A (Controlled by human)	163.0	136.8	149.2	127.4
Robot B (Controlled by human)	141.6	159.0	151.2	143.4
Total Robots Steps	**304.6**	**295.8**	**300.4**	**270.8**
Shifting Movements	40.0	43.0	42.8	38.6
Rotating Movements	7.2	6.8	7.2	5.6
Total Movements	**47.2**	**49.8**	**50.0**	**44.2**
Time (s)	74.6'	75.0'	77.6'	62.2'

Table 3. Average Robots Steps, Table Movements and Time : 1 Human + 1 Robot

Experiment 2	Horizontal Rows	Vertical Rows	T Shape	U Shape
Foreman (Controlled by human)	112.4	110.6	113.6	108.4
Robot A	153.6	141.4	156.4	143.2
Total Robots Steps	**266.0**	**252.0**	**270.0**	**251.6**
Shifting Movements	39.2	40.6	41.0	36.8
Rotating Movements	6.8	4.8	4.8	4.8
Total Movements	**46.0**	**45.4**	**45.8**	**41.6**
Time (s)	66.0’	56.0’	61.8’	55.8’

Table 4. Average Robots Steps, Table Movements and Time : 1 Human + 2 Robots

Experiment 3	Horizontal Rows	Vertical Rows	T Shape	U Shape
Foreman (Controlled by human)	28.6	26.8	29.0	24.4
Robot A	115.6	103.8	118.2	106.8
Robot B	143.4	142.8	150.0	132.0
Total Robots Steps	**287.6**	**273.4**	**297.2**	**263.2**
Shifting Movements	42.0	40.2	43.6	41.8
Rotating Movements	7.4	4.8	7.2	6.0
Total Movements	**49.4**	**45.0**	**50.8**	**47.8**
Time (s)	69.0’	65.0’	71.4’	64.0’

after incompatible moves that humans notice that they are following different plans.

The result in experiment 2 is quite good compared with the other two experiments. Since the robot with the codebook could infer the human-controlled foreman robot’s action by observation and cooperate with it, it is not necessary for the player to communicate with the other robot directly, which is different from the situation in experiment 1. So the player can make his own decision without any other disturbance, which leads to a big improvement in all the cost, including robots steps, table movements and time.

In most of the test cases, the total steps made in experiment 3 is more than experiment 2 but still better than robots totally controlled by humans. Apart from the second test case (Vertical Rows), the robots in experiment 3 made the most table movements in the rest of the test cases. The main reason here would be suboptimal strategies of route planning and obstacle avoidance.

Each player had a few minutes training time to become familiar with the keyboard controls and possible tasks before the real test in experiment 1, but the results show that they still took the longest time in most of the test cases. In experiment 2, with the cooperation of another robot, the foreman worked in an efficient way so that they took the shortest time in each case. One robot initially followed the Foreman to move one table to the place where fits the final task in the last experiment, then these robots finished moving the rest of the tables in a slightly longer time, but still shorter than experiment 1 in most cases.

6 Conclusion

The modeling approach and methodology provided in this paper for constructing the basic framework for cooperative robot communication is context dependent reconstructive communication. By the construction of fuzzy signatures and pattern matching with possibility calculation, we constructed codebook based on the cooperative robot communication scenario.

Through the implementation and evaluation of the CRC simulator, we safely arrive at the conclusion that we can successfully model the communication between the cooperative robots by our designed codebook. In addition, according to the results of the evaluation, the performance of the codebook for robot decision making has reached the effect and efficiency we expected. It has also proved that it is possible to improve the approach to be able to make the robots work more effectively and efficiently than one fully controlled by human-beings even with direct communication for completing cooperative tasks.

References

1. Chong, A., Gedeon, T.D., Koczy, L.T.: Projection Based Method for Sparse Fuzzy System Generation. In: Processings of 2nd WSEAS Int. Conf. on Scientific Computation and Soft Computing, pp. 321–325 (2002)
2. Gedeon, T.D., Koczy, L.T., Wong, K.W., Liu, P.: Effective Fuzzy Systems for Complex Structured Data. In: Proceedings of IASTED International Conference Control and Applications, pp. 184–187 (2001)
3. Goguen, J.A.: L-Fuzzy Sets. J. Math. Anal. Appl. 18, 145–174 (1967)
4. Inagaki, Y., Sugie, H., Aisu, H., Unemi, T.: A Study of A Method for Intention Inference from Human's Behavior. In: IEEE International Workshop on Robot and Human Communication, pp. 142–145 (1993)
5. Koczy, L.T.: Vectorial I-fuzzy Sets. In: Gupta, M.M., Sanchez, E. (eds.) Approximate Reasoning in Decision Analysis, pp. 151–156. North Holland, Amsterdam (1982)
6. Koczy, L.T., Vamos, T., Biro, G.: Fuzzy Signatures. In: Proceedings of EUROPUSE-SIC 1999, pp. 210–217 (1999)
7. Mendis, B.S.U., Gedeon, T.D., Botzheim, J., Koczy, L.T.: Generalised Weighted Relevance Aggregation Operators for Hierarchical Signatures. In: CIMCA 2006, Sydney (2006)
8. Terano, T.: Research Project at LIFE. Laboratory for International Fuzzy Engineering Research (Yokohama) (1993)
9. Vamos, T., Koczy, L.T., Biro, G.: Fuzzy Signatures in Data Mining. In: Proceedings of IFSA World Congress and 20th NAFIPS International Conference, pp. 2842–2846 (2001)
10. Wong, K.W., Chong, A., Gedeon, T.D., Koczy, L.T., Vamos, T.: Hierarchical Fuzzy Signature Structure for Complex Structured Data. In: International Symposium on Computational Intelligence and Intelligent Informatics (ISCIII), Nabeul, pp. 105–109 (2003)
11. Wong, K.W., Gedeon, T.D., Koczy, L.T.: Construction of Fuzzy Signature from Data: An Example of SARS Pre-clinical Diagnosis System. In: Processings of IEEE International Conference on Fuzzy Systems - FUZZ-IEEE 2004, Budapest, Hungary, pp. 1649–1654 (2004)

An Approach to Event-Driven Algorithm for Intelligent Agents in Multi-agent Systems

Anne Håkansson[1] and Ronald Hartung[2]

[1] Department of Information Science, Computer Science,
Uppsala University, Box 513, SE-751 20, Uppsala, Sweden
Anne.Hakansson@dis.uu.se
[2] Department of Computer Science,
Franklin University, 201 S. Grant Avenue, Columbus, Ohio 43215, USA
Hartung@franklin.edu

Abstract. Meta-level agents and intelligent agents in multi-agent systems can be used to search for solutions in networks and graphs where the meta-agents provide paths between nodes based on properties of the graph elements given at the time. A challenge with network problems is finding these search paths while extracting information in the network within an acceptable time bound. Moreover, this is especially difficult when information is extracted and combined from several different sources. Reducing time and making the agents work together requires a plan or an effective algorithm. In this paper we propose an approach to an event-driven algorithm that can search for information in networks using meta-agents in multi-agent systems. The meta-agents monitor the agents using event-driven communication, acting as a search method and extract the searching for information in networks.

Keywords: Algorithms, Event-driven algorithm, Multi-agent systems, Meta-agents, Intelligent agents, Networks, Graphs.

1 Introduction

Common tasks for agents in multi-agent systems in networks or graphs are searching and retrieving information, and combining information collected from adjacent elements. Finding solutions for tasks, often, implies deciding and applying a search strategy. Numerous effective graph search algorithm have been developed; some of them have been applied in agent systems, e.g., heuristics, breadth-first, depth-first search, best-first search, A*, Dijkstra's algorithm, Kruskal's algorithm [11], the nearest neighbour algorithm, and Prim's algorithm [14]. However, searching enormous networks still remains a challenge, with many of the interesting problems are NP-hard. The NP-hard problems include Hamiltonian cycle, travelling salesman (TSP), Clique, Vertex cover, Independent set, Graph partition, Edge cover, and Graph isomorphism [16].

Instead of applying an existing algorithm, it is possible to generate a strategy directly from an executing multi-agent system. In the system, the task of the agents can create the strategy and, in prolongation, manoeuvre the agents. Since the

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 411–420, 2008.

algorithm decides the actions, the agents at the object-level, do not, necessarily, have to be intelligent. Instead the agents can become intelligent agents when performing tasks and learning the environment while moving in the network following connections in the network. At the meta-level, on the other hand, the agents should possess the characteristics of meta-reasoning agents, which allows them to perform reasoning, plan actions and model individual agents and classify conflicts and also to resolve these [2]. Moreover, the meta-level agents can execute an algorithm by applying it to the ground-level agents.

In this paper, we propose an approach using an event-driven algorithm to guide the intelligent agents, while performing tasks in networks and graphs using meta-agents in multi-agent systems. The meta-agents are built on the intelligent agents, by monitoring and collecting the agents working with the same task. Depending on the task assigned to the system, the meta-agents can comprise everything from one agent to dozens or more agents, which will be used to create the event-driven algorithm. Using this algorithm, the meta-agents control behaviour of the ground-level agent moving between states in a network, as well as, inspect the agents reaching a result and perceive the reason for that result.

The benefit of using meta-agents for a multi-agent system is the ability to provide the optimal path between nodes, under given circumstances, and to handle a vast number of nodes in the graph and the network.

2 Related Work

Algorithms have been developed and used for multi-agent systems. For example, Ichida, [10] proposed an extended real-time search algorithm for autonomous agents. To adaptively control search process, the authors propose a search that allows suboptimal solutions with a number of errors and another search, which balance the trade-offs between exploration and exploitation [10]. The search is used in uncertain situations there the goal may change during the course of search. In our work, we do not allow the goal to change during the search.

The prototypical example of event driven programming, namely GUI's, is not an interesting or useful model when examining event driven programming for agents. Distributed algorithms have much more interesting applications. Dunkels *et al.* [4] describes an approach to building distributed algorithms on an event driven base. The authors view is that recasting an algorithm into event driven requires complex state machines to handle the events [4]. Their approach gives a sequential programming model. Our distributed algorithm executes in parallel, using a sequential programming model to handle sequential events.

A different approach is found in Pfeifer *et al.* [13]. The authors separate the programming task into two parts: a single threaded message receiver that handles incoming messages and the execution modules that make computations based on these messages. In a multi-agent system, the design needs to use a modularization strategy for designing smaller short execution time components. We implement the execution modules for event driven systems into our multi-agent system with the extension of multi-threading message receiver.

3 Graph Search Algorithms

Graph algorithms solve problems related to graph theory. The storage of the graph depends on the data structure and the algorithm used to manipulate the graph. Numerous effective graph search algorithms have been developed, where many of the algorithms are used for sequential implementation. There are two approaches for consideration here – a distributed search and a multiple local search that can recombine results.

The basis of most search algorithms is a backtracked search, either with depth-first or breadth-first searching [3]. These are of course uninformed, or brute force, search techniques. Many of the effective algorithms for graph search use these search techniques as a base. For example, greedy algorithms like Prim's [14] and Kruskal's [11] are very effective for simpler problems that are polynomial runtime algorithms [3]. However, there remain a lot of problems that do not yield computationally tractable solution, which are the ones of interest here.

For the more challenging problems, it is about searching to reach a solution faster. There are several extensions used to improve the performance. A variety of techniques are used, for instance heuristic search and breadth-first search [7; 16]. Heuristic searches are used to extend the basic depth-first search using "rules of thumb" [3]. The challenge is always to find an effective heuristic function where the heuristic is usually problem specific. Likewise, heuristics can have problems with search spaces that have features described as foothills, plateaus and ridges. Heuristics can be used to only follow the "best path" or can be employed to order the choices in depth-first search.

By a similar extension, breadth-first searches can be improved by using beam search [3]. In this search, the number of paths, expands at each level of the breadth-first search tree, is limited by applying a heuristic function. The breadth-first search evaluates the merit of each alternative path and then limits the list of paths to include only the best. This is typically accomplished by selecting some maximum number of paths to expand.

Other searching techniques often referred to as optimum techniques, such as branch and bound algorithms but also game algorithms like minmax and alpha-beta search [3]. Branch and bound extends the depth-first search. The extension is to always expand the lowest cost path at each step. This is like best-first without the heuristic function. In many cases, the current best path is highly likely to extend to a good solution. Branch and bound keeps all the paths, so it is guaranteed to find a solution. The sorting of the paths, to the current lowest cost, tends to make the search more efficient.

Branch and bound can be extended with a heuristic to get A* search [18]. A* combines an estimate of the cost to reach a solution from the current state with the cost encountered so far. Again a serious issue is finding the estimate function. If the estimation overestimates the cost, there is no guarantee that a solution will be found. The minmax and alpha-beta techniques are used in game searches [18]. Minmax assumes that alternating levels of the search tree should be selected to give advantage to the alternating players in a game. Thus, from the view of either one of the players, one level is select to maximize their gain and the next level is selected to minimize the other player gain. Alpha-beta uses bounding to prevent expanding all the alternatives. Hence, alpha limits every other level of the tree and beta is used to limit the others.

While using parallel approaches will not reduce these problems from the class NP, it is an interesting approach to explore. Agents provide an interesting mechanism to support distributed processing for search algorithms in graphs.

4 Event-Driven Algorithm

The event-driven algorithm is based on the notion of event-driven programming. Event-driven programming is a programming paradigm in which the flow of the program is determined by messages, from other programs, or by user actions [5]. The event-driven algorithm uses messages as events where the message passing is the communication that performs as a trigger function to process the event.

The event-driven algorithm accomplishes tasks in multi-agent systems by handling object-level agents and meta-level agents. The object-level agents are pre-programmed with, so-called, event loops that look for message to process. The meta-level agents are the dispatchers that call the event handlers. These agents comprise the ground-level agents that have successfully accomplished a task, which correspond to an event queue that holds the unprocessed events.

The object-level agents, also called ground-level agents, collect information about the environment, while moving between nodes, using the event-driven algorithm. The events are driving the action of the agents in the network. Simultaneously, meta-level agents, also called meta-agents, monitor the intelligent agents and, thereby, extract the event-driven algorithm. There will be a lot of different paths explored by performing a particular task, which will be tuned into an optimal search algorithm. The search algorithm will be stored by a meta-agent and reused for future information searching.

The event-driven algorithm can work in connected and directed graphs. For purposes of this work, we consider only connected graphs. This represents a useful set without much of a restriction. Additionally, we can consider either directed and digraphs, or undirected graphs. In some of our prior work, directed graphs were useful for road networks since the cost in each direction can be different. Since the goal of this work is to consider graph algorithms in general, directed graphs are the most useful.

Another interesting property to consider is the graph representation. For some graphs, a single representation, which can be distributed in whole to the agents, can work well. However, for complex graphs, the representation is distributed over multiple machines, which is an important facility. The graph itself might be large and needs to be spread out to conserve memory in a single processor or node. Of course, as memory costs drop, this seems to be a poor reason. A more interesting reason is that a single authority or source does not produce the graph. The graph is produced by multiple sources and, hence, the representation of the graph is more logically a distributed representation. Moreover, when the graph is a part of a real world dynamic system, the graph is a representation of a part of the world and contains data sourced from geographically distributed systems. An example of a geographic distributed system is a highway system with sensors. Such systems clearly call for a distributed agent architecture.

A best search for the event-driven algorithm is to examine the assigned task. Each task has its best path though the network and there is no single search algorithm that covers all the cases for tasks. The only way of find the path is to apply that task on the system and find solutions using a launching function on the agents. Several agents perform the same task and, in the meantime, the agents are checked for performance and the successful are kept as an event for future execution.

4.1 Multi-agent Systems

The multi-agents systems, in which our event driven algorithm is applied, contain intelligent agents at an object-level, i.e., ground level, in graphs and networks [8;9]. The graphs are usually to complex for one agent to perform a task and, thus, require several collaborating agents. In the graphs, the agents execute assignments between the nodes and, in particular, there is one agent per arc. The agents do not request any information about the assignment to perform the task. The only external influence is that the agent is triggered by an event as a message from another agent.

The object-level agents work in an uncomplicated task environment, in which the environment is fully observable, deterministic, episodic and discrete [15]. The intelligent agents have full access to the environment and do not change their internal state to keep track of the external world. Moreover, the agents do not react to unexpected events since they decide the next step, from the current, which makes the environment deterministic. Hence, in fully observable and deterministic environment, the agents do not need to handle uncertainty.

The task is divided into episodes and carried out, singly. The agents' choice of actions only depends on the episode, where each agent performs a single task according to the atomic episode. These episodes also limit the number of actions the agent can take. The results from the agents' actions are collected and assembled into a comprehensive solution. Thus, the results from carrying out episodes are compounded into a complete solution

The agents work in a semi-dynamic environment. The structure of the graph does not change, but the properties in the environment might change. These properties are static and dynamic where the static characteristics are constant constraints in the graph and the dynamic characteristics are temporary obstacles. The agents need to react according to both the constraints and the obstacles.

A semi-dynamic environment requires the agent to be autonomous and flexible [12] to respond to dynamic characteristics. The event sets the agents off but they act according to the environment without external governance or control. Moreover, the agents have to adapt to the environment and learn about the static characteristics since it affects the speed of the agents. Agents that acknowledge the static properties only have to relearn those occasionally.

The agents are intelligent in the sense that they observe the environment and collect data while performing tasks until the agents have finished their task. The collected data is sorted set of static characteristics and set of dynamic characteristics. This data is stored in the agent with the three-tuple <t, s, d> where *t* denotes the task and *s* denotes static characteristics and *d* denotes dynamic characteristics.

Each agent has a routine, which makes it possible to receive the data about the environment, but also unexpected activities. To start working, the agents need event-handlers to check for external events [5]. These event-handlers respond to an incoming event and dispatch the agent. When and where the event occurs is solely determined by the task to be accomplished. The user provides the launching event and the destination. From the launching position, an event triggers all the agents connected to that position to start computing. These agents carry on with the tasks until they reached an end-position, either a goal or a completed task.

The event-driven algorithm follows the different agents but, since they work in parallel, we need a facility that keeps track of and stores the data from the different agents. The facility we use is meta-level agents, or meta-agents. The main task for these meta-agents is to collect the agents that successfully achieved tasks but also to produce and maintain a chain of events that constitute the event-driven algorithm. The algorithm uses the stored events and guides the agents' actions to work in an optimal way by selecting a shorter and faster path.

4.2 Meta-agents

The meta-agents are intelligent agents developed from ground-level agents. These agents conform a horizontal layer on top of the ground-level agents to monitor the actions of the agents and collect the agents together with environment characteristics. The ground-level agents have a purely local view whereas meta-agents can capture a global view of the graph.

The different layers of ground-level agents and meta-level agents require some organisation of the agents, which can support linking between agents, directly or indirectly. Hierarchies of agents are useful [17], since these hierarchies include event handling and communication used for agent cooperation. The meta-agents work as a coordinating mechanism for the object-level agents, like in meta-level reasoning [2]. The meta-agents support communication to and from the ground-level agents and interact with other agents via a communication language for coordination [19; 6]. The communication occurs when an event is sent to the meta-agent.

For each successful ground-level agent reaching a position, the agent will be stored in the same meta-agent. At the position, another ground-level agent is connected and if it performs successfully, the agent is stored in the same meta-agent. Hence, meta-agents constitute a chain of agents computed for accomplishing a task. The chain becomes a stack of events that are triggered for a computation. Thus, the meta-agents are used for computation where each ground-level agent is an event. The events are triggered in the same order, as they are stored in the meta-agent.

Each meta-agent is a four-tuple <ag, spos, epos, c> where *ag* is agent identification, *spos* is the starting position and *epos* is the end position. The agent identification is used as triggers and the tuple *c* is cost for computation, used as a measurement for calculating the optimised solutions for the event-driven algorithm.

The cost for the agents is the ground-level agent's computed costs determined from the static and dynamic characteristics in the environment. These costs can be used to compare the performance of the ground-level agents. The meta-agents compare the performance and select the fastest solutions. The calculation is made in two steps: First, each meta-agent performs the calculation for the each ground-level agent by using all information contained in the ground-level agents, including all derived conclusions. The meta-agents are taking into account the static information and the dynamic information, which is translated into time costs, and simply adds the costs in a total time cost for the ground-level agent. Secondly, the meta-agent takes all ground-level agents and adds the costs for these, which will become the total cost for the meta-agent. All the meta-agents' total costs are compared to obtain the fastest.

As it is possible to search for multiple solutions in the network, the computation can be made in parallel. Several ground-agents can start simultaneously to search the network for solution. For each computation in parallel, several meta-agents are created, i.e., unless the first agent reaches the end node and is also the fastest. The parallel computation can provide for the ground-level agents and meta-agents to use less computational time to find the solution. The event driven algorithm is designed for handling event driven approaches to implement parallel computation in the network, using agents.

4.3 The Search Technique

A search technique for the event-driven algorithm has been tried in a multi-agent system solving problems in connected and directed graphs. The agents were given the task of finding an optimal way between two different positions. It was found that the search technique resembles a mix of depth-first and breadth-first. Each individual meta-agent follows path in a more depth-first strategy. However at a branch point, meta-agents are created, providing an overall breadth-first strategy.

The system starts to explore the ground-level agents associated to the start state with breadth-first, see Figure 1 to the left (Ag A and Ag B). Two meta-agents are created in where these ground-level agents become the first agents (M1 and M2) and then each of these meta-agents explore with a depth-first tactic.

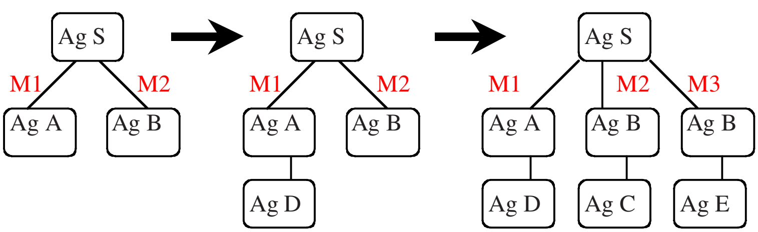


Fig 1. An illustration of the algorithm of the system

The system continues with the first explored ground-level agent (Ag A, left branch in the figure in the middle). This is also expanding the meta-agent for the current intelligent agent. Moreover, the system explores next ground-level agent (Ag B middle and right branch in the figure to the right). If there is one agent invoking several other agents, it is duplicated for each meta-agent (AgB in M2 and M3). Thus, it will be one meta-agent spawned for Ag B and Ag C (M2) and one meta-agent spawned for Ag B and Ag E (M3). The duplication of meta-agents supports building up a unique set of agents while executing the system and, thereby, being able to apply parallel computing.

The algorithm works as follows:

```
While goal not found
    Visit first neighbour
    Use an event algorithm to discover neighbours
    For all the neighbours
        Use an event algorithm to discover vertices adjacent to the neighbour
        For all vertices
            Move to adjacent vertex
            If the meta-agent does not exists
                Create one and add current visited vertex
            Else
                If one vertex adjacent is found
                    Enqueue current visited vertex
                Else
                    Duplicate meta-agent and enqueue one of the meta-agents with
                    current visited vertex
            Mark vertex as visited
        Let the event find next vertex
    Mark the neighbour as visited
    Let an event find next neighbour
```

The meta-agents are an interesting since they interact with multiple agents and they maintain more state information as the computation progresses. These meta-agents may need to migrate from processor to processor. Therefore, the meta-agents are assigned to a collection of processors, and can move from processor to process by their own choice.

5 Other Approaches for the Event Driven Algorithm

There are other interesting approaches to build multi-agent systems with event driven algorithms. The class of problems considered here, are the graph problems based on standard searching techniques. The approaches use events to communicate between the agents and meta-agents in which the meta-agents are charged with the oversight and control. By this, the meta-agents can rank the choices in a best first kind of algorithm and evaluate possible solutions.

A possible extension of the brute force breadth-first search is a best-first strategy. It is to let the meta-agents communicate to coordinate the search of the agents by applying a best-first strategy to the multiple agents. Here, the available agents can expand the best paths, up to the number of agents available. However, the communication cost quickly impose a limiting effect on the amount of parallelism that can be achieved. This is a typical problem of multiprocessing systems.

The depth-first approach is useful when the available agents are all busy. A logical approach would be to allow a group of agents to work under local coordination of single meta-agents. The meta-agent can monitor the group of agents' progress and use events to redirect their activities. This can be extended to beam or branch and bound approaches, i.e., if some of the paths are eliminated from consideration using a bound on the number of paths retained and expanded.

Another approach is to apply heuristic or estimation functions to the search. These are easy extensions, provided that the computation can be performed locally or semi-locally, for a meta-agent monitoring a small group of agents, resulting in either a heuristic or A* search strategy. The localization of the heuristic to a meta-agent may result is less than fully optimal search due to the typical problems with heretics, e.g. foothills, ridges and plateaus. For example, if the best paths pursed in some of the meta-agent groups are all very poor, it is a waste of effort. On the other hand, the parallelism can still leave the process more effective than a sequential algorithm.

We can use multiple agents to try different heuristics in parallel. One of the issues with heuristics is their problem dependence. What works well in one case may fail in another. For this possibility, we can collect a large number of possible heuristics and then start multiple searches in parallel. Each search can use a best-first or a beam search approach with a heuristic. Then, the results can be analyzed to find the best solution generated by any of the heuristics.

6 Conclusions and Further Work

In this paper, we presented an approach to an event-driven algorithm that can search for information in networks and graphs in multi-agent systems. The purpose with the event-driven algorithm is to find search paths within an acceptable time bound. Moreover, the event-driven algorithm has the benefit of driving its own way through the graph using agents as events. For this objective, we use ground-level agent and meta-agents, where the ground-level agents perform a task the meta-agents provide paths between nodes. The ground-level agent systems have an event handler, which are trigged by the meta-agents. The meta-agents monitor the agents using an event-driven acting as a search method and extract the searching for information in networks.

The pseudo code for the algorithm is presented in the paper, following the computer system built for networks. The algorithm is driven by events, searching for solutions that satisfy the goal. Next step in this research is to provide a large set of test cases and run these cases to other search algorithm. The test cases must cover more examples in graphs and network than the one we developed as a search technique in graphs.

References

1. Berman, K., Paul, J.: Algorithms: Sequential, Parallel, and Distributed. Course Technology, 1st edn. (2004), ISBN-10: 0534420575
2. Chelberg, D., Welch, L., Lakshmikumar, A., Gillen, M., Zhou, Q.: Meta-Reasoning For a Distributed Agent Architecture (2000), `http://zen.ece.ohiou.edu/~robocup/papers/HTML/SSST/SSST.html`
3. Cormen, T.H., Leiserson, C.E., Riverst, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. The MIT Press, Cambridge (2006)

4. Dunkels, A., Schmidt, O., Voigt, T., Ali, M.: Protothreads: simplifying event-driven programming of memory-constrained embedded systems. In: Proceedings of the Fourth ACM Conference on Embedded Networked Sensor Systems (SenSys 2006), Boulder, Colorado, USA (2006)
5. Ferg, S.: Event-Driven Programming: Introduction, Tutorial, History (2006-02-08), http://TutorialEventDrivenProgramming.sourceforge.net
6. Genesereth, M.R., Ketchpel, S.P.: Software Agents. Communication of the ACM 37(7) (July 1994)
7. Horowitz, E., Sahni, S., Rajasekaran, S.: Computer Algorithms C++: C++ and Pseudo code Versions, 2ndRev. edn. p. 769. W. H. Freeman, New York (1996)
8. Håkansson, A., Hartung, R.: Using Meta-Agents for Multi-Agents in Networks. In: Proceedings of ICAI 2007, The 2007 International Conference on Artificial Intelligence, 2007, WORLDCOMP 2007, June 25-28, 2007, Las Vegas, USA (2007)
9. Håkansson, A., Hartung, R.: Calculating Optimal Decision Using Meta-level Agents for Multi-Agents in Networks. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 180–188. Springer, Heidelberg (2007)
10. Ishida, T.: Real-Time Search for Autonomous Agents and Multi-Agent Systems. Journal of Autonomous Agents and Multi-Agent Systems 1(2), 139–167 (1998)
11. Kruskal, J.B.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proceedings of the American Mathematical Society 7(1), 48–50 (1956)
12. Luger, G.F.: Artificial Intelligence – Structures and strategies for Complex Problem Solving, 4th edn. Pearson Education, London (2002)
13. Pfeifer, A., Ururahy, C., Rodriguez, N., Ierusalimschy, R.: Event-Driven Programming for Distributed Multimedia Applications. In: Distributed Computing Systems Workshops Proceedings. 22nd International Conference on Distributed Computing Systems, pp. 583–584 (2002), ISBN: 0-7695-1588-6
14. Prim, R.C.: Shortest connection networks and some generalizations. Bell System Technical Journal 36, 1389–1401 (1957)
15. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall Inc, Englewood Cliffs (1995)
16. Skiena, S.: The Algorithm Design Manual, 1st edn. Springer, Heidelberg (1998)
17. Sauer, J., Appelrath, H.-J.: Scheduling the Supply Chain by Teams of Agents. In: HICSS 2003, The 36th Annual Hawaii International Conference on System Sciences, p. 81 (2003)
18. Winston, P.H.: Artificial Intelligence, 3rd edn. Addison-Wesley, Reading (1992)
19. Wooldridge, M., Jennings, N.R.: Intelligent agents: Theory and practice. Knowledge Engineering Review 10(2) (1995)

Agvs Distributed Control Subject to Imprecise Operation Times

Grzegorz Bocewicz[1], Robert Wójcik[2], and Zbigniew Banaszak[1]

[1] Koszalin University of Technology, Dept. of Computer Science and Management
ul. Śniadeckich 2, 75-354 Koszalin, Poland
zbigniew.banaszak@tu.koszalin.pl

[2] Wrocław University of Technology, Institute of Computer Engineering, Control and Robotics, 50-372 Wrocław, Poland
robert.wojcik@pwr.wroc.pl

Abstract. The subject matter of the study are the automated guided vehicle (AGV) operation synchronization mechanisms in flexible manufacturing systems, where transport processes can be modeled as a system of cyclic concurrent processes sharing common resources, i.e., some preset travelling route intervals. In this paper the problem of determination of the local rules coordinating access of the vehicles to the shared travel route intervals, ensuring the deadlock-free execution of the repetitive processes was reduced to determination of the sufficient conditions. In that context the problem considered can be seen as a problem of multi-agent coordination, and defined as the constraint satisfaction problem (CSP) subject to precise and imprecise operation times, and then solved with use of the logic programming techniques.

Keywords: knowledge engineering, constraints logic programming, deadlock avoidance, fuzzy set, scheduling.

1 Introduction

A considered class of objects covers the transport subsystems of the flexible manufacturing systems (FMS). In subsystems of that type, along given traveling routes, automated guided vehicles (AGVs) move and stop at given workstations, at strictly defined instants in order to load and/or unload appropriate product batches.

The constraints connected with the available traveling route width, the topology of traveling routes and itineraries of individual vehicles, lack of simultaneous access to the stations, etc. imply the necessity to investigate conditions leading to vehicle possible collisions and deadlocks [7]. A deadlock-avoidance considered belongs to NP-hard problems [11]. The existing approach to solving the problem base usually upon the simulation models, e.g. the Petri nets [8] or the algebraic models, e.g. upon the (max,+) algebra [10].

The problem considered in this paper reduces to determination of the rules coordinating the access of the vehicles to the shared system resources ensuring the collision-free and deadlock-free AGVs' motion. Assuming existence of local priority decision

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 421–430, 2008.

rules, the problem reduces to determination of the sufficient conditions in the form of a pair (an initial state, a set of priority rules).

Of course, since the system of autonomous AGVs can be seen as a multi-agent [12] one being aimed at the FMS workstations servicing. So, the problem considered can be treated in terms of distributed agents control. Since the quality of workstation service be treated in terms of distributed agents control. Moreover, because the due time workstation service depends on operation times the imprecise nature of transportation times have to be taken into account. In that context, the main considerations are focused on the AGVs distributed control mechanism, i.e. the conditions providing resource conflicts resolution guaranteeing the deadlock-free AGVs cooperation.

Section 2 provides basic assumptions and notations and then states the problem. Section 3 introduces to the logic-algebraic methods, and then discusses the issue of the sufficient conditions guaranteeing deadlock-free AGVs movement. Section 4 recalls concepts standing behind constraint satisfaction problem. Section 5 provides illustrative example of the approach application. The discussion of the approach extension to imprecise data based scheduling provides the Section 6. The final section summarizes this contribution and sheds light on the future direction of research.

2 Problem Formulation

The AGVs co-sharing the resources and executing repetitive tasks may be presented in a form of Cyclic Concurrent Process Systems (CCPS), wherein the cyclic processes are interconnected by use of the common resources (tracks, machine tools, etc.). A graphical representation of an exemplary CCPS is shown in Figure 1. Eight processes $P_1, P_2, \ldots, P_8$, encompass operations of relevant vehicles served by the resources R_1-R_{14}.

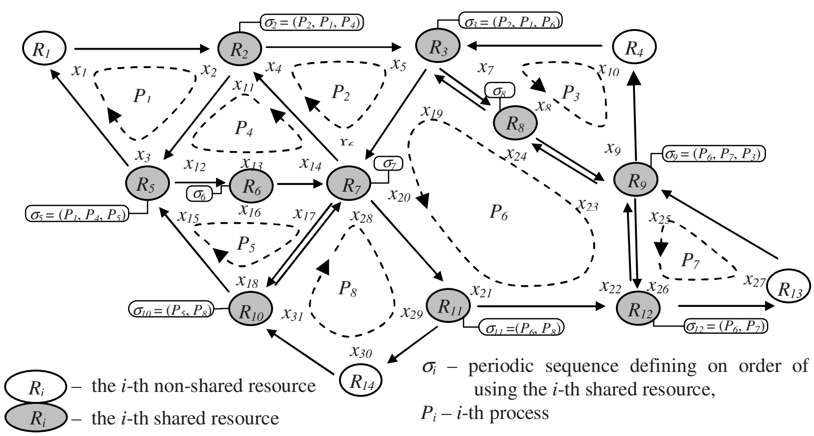


Fig. 1. Graphical representation of CCPS

The processes cooperation is determined by the following constraints [10]:

- the processes share the common resources in the mutual exclusion mode,
- commencement of a successive process operation happens immediately after completing of the current operation provided that there is a possibility of making use of the successive resource requested by the given process,
- during waiting for a busy resource, the process does not release the resource allocated for execution of the previous operation,
- the processes executed cyclically are not pre-emptive, i.e. the resource may not be taken of the process while it is using it,
- in one cycle, a process may pass via any resource along its transportation route once only.

The CCPS consists of q processes: $P = \{P_1,P_2,\ldots,P_q\}$determining the travelling route of AGVs, and k resources $R = \{R_1,R_2,\ldots,R_k\}$ corresponding to the path sectors of FMS considered. A cyclic process $P_i = (p_{i,1}, p_{i,2},...,p_{i,n^i})$ is a sequence, the components of which define the numbers of the resources used for execution of individual process operations, where: $p_{i,j} \in R$ – denotes the resource used by the i-th process in the j-th operation; after completion of the operation with a share of the resource p_{i,n^i}, the operation with the share of the resource $p_{i,1}$ is executed again. The sequence defines the transportation route, i.e., the ordering in which the given process is executed. P_i – may begin at an arbitrary resource of the sequence presented.

The time representation of the i-th cyclic process is constituted by the sequence $T_i = (t_{i,1}, t_{i,2}, ..., t_{i,n^i})$, where $t_{i,j}$ denotes the time of execution of the j-th operation by the i-the process. The sequence $\Theta = (\sigma_j, \sigma_k, ..., \sigma_z)$ of the priority dispatching rules, i.e., accesses of the processes to the shared resources of the transport subsystem, where $\sigma_i = (s_{i,j}, s_{i,k}, ..., s_{i,l})$ – is the sequence, the components of which determine the service order in which processes are served by the i-th resource, $s_{i,k}$ – the k-th process executed on the i-th resource. Each sequence σ_i is periodic and gives access to the i-th resource to every process using it as to avoid processes' starvation.

An initial state $S_0 = (R_i, R_j, ..., R_k)$ is the sequence, components of which $R_j = crd_iS_0$ denote the i-th process beginning its realization on the resource R_j; crd_iS_0 – denotes the i-th coordinate of the vector S_0.

The sequence of operations of the size equal to the number of all operations executed in the system is defined as $p = (P_1, P_2, ..., P_r) = (p_{1,1}, ..., p_{1,n1}, p_{2,1}, ..., p_{2,n2}, ..., p_{r,1}, ..., p_{r,nr})$, where $p_{i,j}$ – denotes the j-th operation of the i-th processes. The operation execution time sequence $T = (T_1, T_2, ..., T_r) = (t_{1,1}, ..., t_{1,n1}, t_{2,1}, ..., t_{2,n2}, ..., t_{r,1}, ..., t_{r,nr})$, where $t_{i,j}$ denotes the execution time of the j-th operation by the i-th process.

The state sequence $x = (x_1, x_2, ..., x_r)$, where: x_i corresponds the operation represented in the sequence p by the i-th coordinate, the value x_i denotes the instant that the operation is begun in the first cycle.

Given a system of class CCPS the structure and process parameters of which are given in the form of vectors defining the vehicle routes P_i and service times T_i in subsequent stations. Does it exist a pair (initial system state S_0, priority rule set Θ) ensuring that the assumed transport processes are executed, with the cycle time not exceeding the arbitrarily given value H? An answer to the problem formulated covers, therefore, the response to the question: Does there exist the set of sufficient conditions ensuring the cyclic (i.e. deadlock-free) execution of the concurrent processes?

3 Logic-Algebraic Method

The system can be described be seen in terms of the knowledge base representation:

$$KB = <C,W,Y; R>,$$

where: $R = \{(c,w,y): F(c,w,y) = 1\}$ – is a relation being the set of all triples (c,w,y), for which the facts F describing the system hold; $F(c,w,y) = (F_1(c,w,y), F_2(c,w,y), ..., F_K(c,w,y))$ – is the composition of the logic fact values being the functions of the variables c, w, y; $F(c,w,y) = 1$ – means the all facts $F_i(c,w,y)$, $i = 1,2,\ldots,K$, holds: $E(F_i(c,w,y)) = 1$, ($E(A)$ – means the Boolean value of the logic formulae A); $c = (c_1, c_2, ..., c_{kc})$ – the sequence of the input variables; $y = (y_1, y_2, ..., y_{ky})$ – the sequence of the output variables; $w = (w_1, w_2, ..., w_{kc})$ – the sequence of the auxiliary variables; $c \in C$, $y \in Y$, $w \in W$, C,Y,W – the sets defining the domains of the variables c, y, w.

3.1 Knowledge Base

The *KB* describing an arbitrary system is presented in the form of the sets C, W, Y, that define domains of variables c, y, w, describing some system properties. C is the set of input variables describing the input properties of the system. Y – is the set of output variables describing the output properties of the system. W consists the auxiliary variables. The *KB* is represented by facts $F(c,w,y)$, i.e., propositions reflecting, on the logic level, the relationship between individual variables c,w,y. The triples c,w,y, for which all facts $F(c,w,y)$ are true, are presented in the form of the relation R. So, the representation of the *KB* can be defined as follows [3]:

$$KB = <S0, \Sigma, X; R> \tag{1}$$

where: $S0$ – the set of all possible initial states S_0, Σ - the set of all possible access rules Θ for the shared resources (input variables), X – the set of all possible forms of the state vector x (output variables), $R = \{(S_0,\Theta, x): F(S_0,\Theta, x) = 1\}$ – the relation defining the values S_0,Θ, x, for which the facts $F(S_0,\Theta, x)$ are true. The set R covers facts $F(S_0,\Theta, x)$ that describe the system properties in dependence of the initial state S_0, the rules of access to shared resources Θ and of the starting times of operations x.

3.2 Knowledge Generation

The *KB* considered can be treated as specification of general assumptions the transportation system has to follow while taking into consideration the time and resources constraints. In other words, the knowledge is specified by a set of facts $F(S_0,\Theta, x)$. The assumptions regarding the transportation system concern of [3]:

- Constraints regarding an order in which processes have to be executed.
- Constraints limiting processes servicing by local resources: the moment x_j that the process P_i may start its execution at the resource R_k is equal to the time required by this process to complete its previous operation $(x_{j-1} + t_{j-1})$:

$$x_j = x_{j-1} + t_{j-1} \tag{2}$$

- Constraints regarding processes servicing by shared resources: the instant x_j that the operation of the process P_i is begun at the shared resource R_k is determined by the maximum within the completion time $(x_{j-1}+t_{j-1})$ of the process P_i on the subsequent resource R_{k-1}, and the instance the operation corresponding to x_p begins its execution on the resource R_{k+1} has been served P_o previously executed on R_k just before P_i.

$$x_j = \max\{x_p, x_{j-1}+t_{j-1}\} \tag{3}$$

The constraints provided (2), (3), can be seen as a part of a transport system specification in the context of its parameters x, S_0, Θ, [3]. Besides of the above mentioned assumptions the other one guaranteeing processes deadlock-freeness can be introduced. It should be reminded that processes P_1, P_2, …, P_n are in deadlock in the case their calls for recourses (busy by processes) form a closed circle of resources request. So, the formulas (2) and (3), can be unified in the following recurrent formulae, and called a state equation (4):

$$^{p}x_v^{s(i+1)}(x) = \max\{x_j+t_j, \max\{x_{j+1}+t_{j+1}, ..., \max\{x_{nz-1}+t_{nz-1}, \max\{x_{nz}+t_{nz}, ... \\ \ldots,\max\{x_{j-1} + t_{j-1}, x_j\}...\}\}\}\} \tag{4}$$

where: x_q – is the moment the process P_j starts at the state S_i, $^{p}x_v^{s(i+1)}(x)$ – denotes the moment the p-th process starts its execution at the next state S_{i+1}.

For the given state S_i the moment the process P_i starts its execution at the next state S_{i+1} can be calculated due to the formulae (4). It can be shown that if the state equation belongs to a class of the identity equations, then the system considered is deadlock-free one [3]. So, the state equation (4) can be employed for calculation of an operation starting moment at any further state. This can be realized through the so called general state equation (5):

$$^{p}X_v^{sw}(x)={}^{p}x_v^{sw}(\,{}^{p}x_v^{s(w-1)}(\,{}^{p}x_v^{s(w-2)}(\,\ldots\,{}^{p}x_v^{s0}(x)\,\ldots\,))) \tag{5}$$

where: $^{p}X_v^{sw}(x)$ –the moment the p-th process starts its operation at the state S_w, the last one in the cycle.

3.3 Decision Problem

Response to the question stated in the Section 2.2 results in pairs (S_0, Θ) determined in terms of R_x guaranteeing the cycle time (determined in terms of R_y) do not exceed a given value H. $R_x = \{(S_0,\Theta)\colon F_c(S_0,\Theta) = 1\}$ – is the set of the values S_0, Θ, for which the input system property $F_c(S_0,\Theta)$ holds, while $R_y = \{x\colon F_y(x) = 1\}$ – is the set of values of x for which the output property $F_y(x)$ holds, where: $F_c(S_0,\Theta)$ is a set of the logic propositions that describe the input system properties in dependence on the initial state S_0 and the priority rules Θ, while $F_y(x)$ is a set of the logic sentences that describe the output properties of the system in dependence of the variable x values. In the case considered, the output property is as follows: $F_y(x)$: $(x_1+t_1 \le H) \wedge (x_2+t_2 \le H) \wedge ... \wedge (x_{rn}+t_{rn} \le H)$ ($F_y(x)$ – means: "the cycle time do not exceed a given value H").

The decision problem consists in determination of a set R_x for which the property F_c holds (6):

$$F_c(S_0,\Theta) \Rightarrow F_y(x) \tag{6}$$

R_x - is a set of pairs (S_0, Θ) for which the system operate in cycles not lasting longer than H. The relation R_x can be obtained from the sets S_{x1}, S_{x2} due to logic-algebraic method [5]:

$$R_x = S_{x1} \setminus S_{x2} \tag{7}$$

where: $S_{x1} = \{(S_0,\Theta)$: $F(S_0,\Theta, x) = 1, F_y(x) = 1\}$;$S_{x2} = \{(S_0,\Theta)$: $F(S_0,\Theta, x) = 1, F_y(x) = 0\}$.

The set S_{x1} can be seen as a set consisting of S_0, Θ, whose follow the facts: $F(S_0,\Theta, x)$, $F_y(x)$. In turn, the set S_{x2} can be seen as a set consisting of S_0, Θ, whose follow the fact $F(S_0,\Theta,x)$ and do not follow $F_y(x)$. The intersection of sets S_{x1}, S_{x2}, consist values of S_0, Θ, for which the fact $F_y(x)$ does not hold. $R_x = \emptyset$ denotes the lack of answer to the question asked. Searching for R_x is the NP-hard problem.

4 Constraint Satisfaction Problem

Each *KB* (1) can be presented in the form of the constraint satisfaction problem (*CSP*) [2] $CS = ((Q, D), Co)$ is defined as follows: Given finite set of discrete decision variables $Q = \{q_1, q_2, ..., q_n\}$, a family of finite variable domains $D = \{D_i \mid D_i = d_{i,1}, d_{i,2}, ..., d_{i,j}, ..., d_{i,m}\}$, $i = 1..n\}$ and the finite set of constraints $Co = \{Co_i \mid i = 1..L\}$ limiting the decision variable values. The admissible solution, i.e. the one that the values of all variables meet all constraints of the set Co is sought for. In case *CSP,* maps the *KB*, the constraints Co play the role of facts included in $F(S_0,\Theta, x)$ while the role of the variables Q – the values of S_0, Θ, x. Therefore, the considered problem results in (8):

$$CS = (((S_0,\Theta, x), D), \{F(S_0,\Theta, x) = 1\}) \tag{8}$$

where: $D = \{D_{S0}, D_\Theta, D_x\}$, D_{S0} - is the set of system resources; D_Θ - is the set including the processes realized in the system; $D_x = \{0,1,...,H\}$ - is the set of the time values; $F(S_0,\Theta,x) = 1$ - denotes the sequences of facts: $(E(F_1(S_0,\Theta, x)) = 1,..., E(F_K(S_0,\Theta,x)) = 1)$, $E(A)$ – means the Boolean value of the logic formulae A.

The solution of the *CSP* problem formulated in such way is a set of sequence (S_0,Θ) and the starting times of individual operations on the resources x for which all constraints presented in the form of the logic sentences $F(S_0,\Theta,x)$ are true. Solving the decision problem (i.e., determination of the R_x) in terms of *CSP* folds into the following problems:

$$\begin{aligned} CS_{Sx1} &= ((S_0,\Theta, x), D), \{F(S_0,\Theta, x) = 1, F_y(x) = 1\}) \\ CS_{Sx2} &= ((S_0,\Theta, x), D), \{F(S_0,\Theta, x) = 1, F_y(x) = 0\}) \end{aligned} \tag{9}$$

The obtained set $R_x = S_{x1} \setminus S_{x2}$ can be treated than as solution of the considered *CSP*, i.e. determines the response to the assumed query.

5 Example of Schedules Design

Given an AGVs system of the structure shown in the Fig. 1. $T_1, T_2, \ldots, T_8$ encompass service times as specifying processes $P_1, P_2, \ldots, P_8$. The system shared resources $R_1, R_2, ..., R_{13}$ and the following parameters are known:

$P_1 = (R_1, R_2, R_5)$, $P_2 = (R_2, R_3, R_7)$, $P_3 = (R_3, R_8, R_9, R_4)$, $P_4 = (R_2, R_5, R_6, R_7)$, $P_5 = (R_5, R_6, R_7, R_{10})$, $P_6 = (R_3, R_7, R_{11}, R_{12}, R_9, R_8)$, $P_7 = (R_9, R_{12}, R_{13})$, $P_4 = (R_7, R_{11}, R_{14}, R_{10})$, $T_1 = (1,2,3)$, $T_2 = (1,2,3)$, $T_3 = (2,1,3,4)$, $T_4 = (1,2,3,4)$, $T_5 = (1,2,3,4)$, $T_6 = (1,2,3,1,2,3)$, $T_7 = (1,2,3)$, $T_4 = (1,2,3,8)$, $x = (x_1, x_2, \ldots, x_{30})$, $p = (R_1, R_2, \ldots, R_{13})$.

Consider the routine query: Does there exist a pair $(S_0,\{\sigma_2,\ldots,\sigma_{12}\})$ such that guarantee the system cycle will not exceed 13 time units?

Due to [3], the *KB* representation corresponding to the given specification is as follows: $KB = \langle S0, \Sigma, X; R\rangle$, where: $R=\{(S_0, \Theta, x){:}F(S_0,\Theta, x) = 1\}$.

So, due to (9) the *CSP*s are as follow:

$$CS_{Sx1} = ((S_0, \Theta, x), D), \{F(S_0, \Theta, x) = 1, F_y(x) = 1\}),$$
$$CS_{Sx2} = ((S_0, \Theta, x), D), \{F(S_0, \Theta, x) = 1, F_y(x) = 0\}),$$

where: $D = \{D_{S0}, D_\Sigma, D_x\}, D_\Sigma=\{D_{\sigma 2}, ..., D_{\sigma 12}\}, D_{S0} =\{R_1, \ldots, R_{14}\}, D_{\sigma 2} = \{P_1, P_2, P_4\}$, $\ldots, D_{\sigma 12}= \{P_6, P_7\}, D_x = \{1, ..., 40\}$.

The fact $F_1(S_0,\sigma_{12},x)$ defining the starting time of the operation x_2, assuming the mutual exclusion of the processes P_6, P_7, on the resource R_{12}, is of the form: $F_1(S_0,\sigma_2,x)$: $\neg(crd_6S_0 = R_{12}) \wedge \neg(crd_6S_0 = R_9) \wedge \ldots \wedge (crd_1\sigma_{12} =P_7) \Rightarrow (x_{22} = \max(x_{25}, x_{21}+t_{21}))$, i.e., if in the state S_0 the process P_6 does not use the resource R_{12}, and in the state S_0, the process P_7 does not use the resource R_9, and – the P_7 process uses the resource R_{12} as the first one, consequently $x_{22} = \max(x_{25}, x_{21}+t_{21})$. The fact defining the output property $F_y(x)$: $(x_1+ t_1 \leq 13) \wedge (x_2+t_2 \leq 13) \wedge \ldots \wedge (x_{30}+t_{30} \leq 13)$ corresponds to the assumed condition: "*the cycle will not exceed 13 time units*".

The sets S_{x1}, S_{x2} as solutions to CS_{Sx1}, CS_{Sx2}, (implemented in OzMozart) results in $R_x = S_{x1} \backslash S_{x2}$ containing 318 alternative sufficient conditions that must be fulfilled in order to guarantee the system cycle do not exceed 13 time units. One of them is as follows: $\{S_0 = (R_1, R_3, R_4, R_7, R_5, R_{12}, R_{13}, R_{10}), \{\sigma_2 = (P_4, P_1, P_2), \sigma_3 = (P_2, P_6, P_3), \sigma_5 = (P_5, P_4, P_1), \sigma_6 = (P_5, P_4), \sigma_7 = (P_4, P_8, P_5, P_2, P_6), \sigma_8 = (P_6, P_6), \sigma_9 = (P_6, P_7, P_3), \sigma_{10} = (P_8, P_5), \sigma_{11} = (P_8, P_6), \sigma_{12} = (P_6, P_7)\}\}$.

6 Imprecise Data Based Scheduling

The similar approach can be applied to the AGVs scheduling in the case the operation times T_f (by analogy to the sequence T) are imprecise [6], [13]:

$$T_f = (t_{f1,1}, \ldots, t_{f1,n1}, t_{f2,1}, \ldots, t_{f2,n2}, \ldots, t_{fr,1}, \ldots, t_{fr,nr})$$

where: $t_{fi,j}$ – is the fuzzy variable determining the execution time of the j-th operation in P_i, $t_{fi,j} =\{(\mu_{I,j}(t), t)\}$, $t\in T_{fi,j}$, $T_{fi,j}$ – the discrete time domain of the membership function $\mu_{i,j}(t)$, $\mu_{i,j}(t)$– is the membership function, assigning to each element x its value $t_{fi,j}$, $\mu_{i,j}(t) \in \{0, 1]$. A discrete membership function represented by sets [9]:

$$t_{fi,j} = \left\{ \frac{\mu_{i,j}(t_{i,j,1})}{t_{i,j,1}}, \frac{\mu_{i,j}(t_{i,j,2})}{t_{i,j,2}}, \ldots, \frac{\mu_{i,j}(t_{i,j,k_{i,j}})}{t_{i,j,k_{i,j}}} \right\} \tag{10}$$

where: $t_{i,j,1}, t_{i,j,2}, \ldots, t_{i,j,k_{i,j}} \in T_{fi,j}$, $T_{fi,j} \subset N$.

In general case the sequence T corresponds to the ordered set of fuzzy variables $T_t = \{T_{f1,1},\ldots,T_{f1,n1},T_{f2,1},\ldots,T_{f2,n2},\ldots,T_{fr,1},\ldots,T_{fr,nr}\}$. In other words, to each variable $t_{fi,j}$ corresponds the domain of discrete variables $T_{fi,j} = \{t_{i,j,1}, t_{i,j,2},\ldots,t_{i,j,ki,j}\}$, and the set $M_{i,j} = \{\mu_{i,j}(t_{i,j,1}), \mu_{i,j}(t_{i,j,2}),\ldots,\mu_{i,j}(t_{i,j,ki,j})\}$, $M = \{M_{1,1},\ldots,M_{1,n1},M_{2,1},\ldots,M_{2,n2},\ldots, M_{r1},\ldots,M_{r,nr}\}$ associating the membership value $\mu_{i,j}(t_{i,j,a})$ to each value of the discrete variable $t_{i,j,a}$.

Let us assume (S_0, Θ), the instants x, etc., are distinct ones. Consider the questions: Does there exist a pair (S_0, Θ) ensuring the AGVs cyclic motion is deadlock-free? What is the cycle time H of the AGVs system in the reached cyclic steady-state?

6.1 Knowledge Base Representation

Taking into account assumed fuzziness of operations execution the relevant knowledge base is defined as follows:

$$KB = <S0, \Sigma, X, T_t, M; R> \tag{11}$$

where: $R = \{(S_0, \Theta, x, d_f, m: F(S_0, \Theta, x, d_f, m) = 1\}$; $d_f = \{d_{f1,1},\ldots,d_{f1,n1},d_{f2,1},\ldots, d_{f2,n2},\ldots,d_{fr,1},\ldots,d_{fr,nr}\}$; $d_{fi,j}\in T_{fi,j}$, $T_{fi,j} = \{t_{i,j,1},t_{i,j,2},\ldots,t_{i,j,ki,j}\}$; $T_t = \{T_{f1,1},\ldots,T_{f1,n1},T_{f2,1}, \ldots,T_{f2,n2},\ldots,T_{fr,1},\ldots,T_{fr,nr}\}$; $m = \{m_{1,1},\ldots,m_{1,n1}, m_{2,1}, \ldots, m_{2,n2}, \ldots, m_{r,1}, \ldots, m_{r,nr}\}$; $m_{i,j} \in M_{i,j}$, $M_{i,j} = \{\mu_{i,j}(t_{i,j,1}), \mu_{i,j}(t_{i,j,2}),\ldots,\mu_{i,j}(t_{i,j,ki,j})\}$, $M = \{M_{1,1},\ldots,M_{1,n1},M_{2,1},\ldots,M_{2,n2},\ldots,M_{r,1},\ldots, M_{r,nr}\}$ – the sequence of subsequences of the membership function corresponding to the elements of the sequence T.

R_x being a set of pairs (S_0, Θ) is determined in quite similar way as in the case of distinct values. The set of alternative sufficient conditions R_x (12) follows form the sets (13).

$$RA_x = A_{x1} \setminus A_{x2} \tag{12}$$

$$\begin{aligned} A_{x1} &= \{(S_0, \Theta, d_f): F(S_0, \Theta, x, d_f, m) =1, F_y(x) = 1\}; \\ A_{x2} &= \{(S_0, \Theta, d_f): F(S_0, \Theta, x, d_f, m) =1, F_y(x) = 0\}; \end{aligned} \tag{13}$$

Since $T_k = \{crd_3\, g: g \in R_x\}$ and $T_r=\{\{d_{f1,1}, \ldots, d_{f1,n1}, d_{f2,1}, \ldots, d_{f2,n2},\ldots, d_{fr,1} ,\ldots, d_{fr,nr} \}:d_{fi,j} \in T_{fi,j}\in T_t\}$, hence $(T_k=T_r)\Rightarrow(R_x=\{(S_0,\Theta): S_0,\Theta\in RA_x\})$ and $(T_k\neq Tr) \Rightarrow (R_x= \varnothing)$.

So, similarly to the distinct variables the sets A_{x1}, A_{x2} are determined due to (9):

$$\begin{aligned} CS'_{Ax2} &= ((S_0, \Theta, x, d_f), D),\{F(S_0, \Theta, x, d_f, m)=1,F_y(x) = 0\}) \\ CS'_{Ax1} &= ((S_0, \Theta, x, d_f), D),\{F(S_0,\Theta, x, d_f, m)=1,F_y(x) = 1\}). \end{aligned} \tag{14}$$

So, the solution of the above problems provides the sets A_{x1}, A_{x2} , and finally leads to the set RA_x. R_x can be calculated due to the following formulae (15). $(S_0,\Theta)\in R_x$ implies there exists the deadlock-free x, for which cycle time follows form (15):

$$h = \max\{x_1+d_{f1,1} ,\max\{x_2+ d_{f\,1,2} ,\max\{\ldots, \max\{x_r+ d_{f\,r,nr}\}\}\ldots\}\}\}. \tag{15}$$

The set $\{d_{f1,1},\ldots,d_{f1,n1},d_{f2,1},\ldots,d_{f2,n2},\ldots,d_{fr,1},\ldots,d_{fr,nr}\}$ provides any combination of distinct moments for fuzzy sets T_t. Since $d_{fi,j}$ can be selected in different ways, hence to each pair (S_0, Θ) many different values of h may correspond. Since h implies $\mu_h(h)$ hence to a given h the different values of $\mu_h(h)$ may correspond (i.e., the same cycle of a given value h can be obtained for different values $d_{fi,j}$). In other words it is assumed that to a given cycle h two values of the membership function $\mu_{hmin}(h)$, $\mu_{hmax}(h)$ determining the lowest and highest rate of h, $\mu_{hmin}(h) \leq \mu_{hmax}(h)$. Consequently, the set of triples $(\mu_{hmin}(h), \mu_{hmax}(h), h)$ can be treated as a fuzzy set determining the cycle time corresponding to (S_0, Θ) for the given fuzzy operation times T_f:

$$H = \{(\mu_{hmin}(h), \mu_{hmax}(h), h)\} \quad (16)$$

The $\mu_{hmin}(h)$, $\mu_{hmax}(h)$ are calculated due to the formulae (17), (18), based on the max and min operators:

$$\mu_{h\min}(h) = \min_{\{d_{f1,1},...,d_{f1,n1},...,d_{fr,nr}\}\in T_h} \{\max\{\mu_{df1,1}(d_{f1,1}), \max\{\mu_{df1,2}(d_{f1,2}),...,\max\{\mu_{dfr,nr}(d_{dfr,nr})\}...\}\}\} \quad (17)$$

$$\mu_{h\max}(h) = \max_{\{d_{f1,1},...,d_{f1,n1},...,d_{fr,nr}\}\in T_h} \{\max\{\mu_{df1,1}(d_{f1,1}), \max\{\mu_{df1,2}(d_{f1,2}),...,\max\{\mu_{dfr,nr}(d_{dfr,nr})\}...\}\}\} \quad (18)$$

T_h – family of sets $\{d_{f1,1}, ..., d_{f1,n1}, ..., d_{fr,nr}\}$ guaranteeing the cycle time equals to h.

6.2 Scheduling

Consider the AGVs system modeled in terms of CCSP as shown in Fig. 2a.

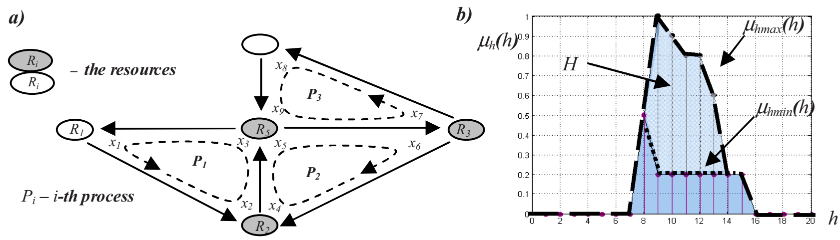


Fig. 2. a) Graphical representation of exemplary CCPS, b) set of H for CCPS

In the system considered the following three transportation processes are executed P_1, P_2, P_3: $P_1 = (R_1, R_2, R_5)$, $P_2 = (R_2, R_5, R_3)$, $P_3 = (R_3, R_4, R_5)$, $x = (x_1, x_2,..., x_9)$, The operation times of particular operations considered as imprecise are treated as fuzzy variables: $T_f = (t_{f1,1}, t_{f1,2}, t_{f1,3}, t_{f2,1}, t_{f2,2}, t_{f2,3}, t_{f3,1}, t_{f3,2}, t_{f3,3})$, where:

$t_{f1,1} = \{0,5/2; 1/3; 0,8/4\}$; $t_{f1,2} = \{1/2; 0.8/3; 0,6/4\}$; $t_{f1,3} = \{0,5/3; 1/4; 0.8/5\}$;
$t_{f2,1} = \{0,8/2; 1/3; 0,8/4\}$; $t_{f2,2} = \{0,5/1; 1/2; 0,8/3\}$; $t_{f2,3} = \{0,5/2; 1/3; 0,8/4\}$;
$t_{f3,1} = \{1/1; 0,9/2; 0,2/3\}$; $t_{f3,2} = \{0,5/3; 1/4; 0,8/5\}$; $t_{f3,3} = \{1/2\ 0,9/4\ 0,2/3\}$.

Given $S_0 = (R_i, R_j, R_k)$, where: $R_i \in P_1$, $R_j \in P_2$, $R_k \in P_2$. Given the set of dispatching rules: $\Theta = (\sigma_2, \sigma_3, \sigma_5)$, where: $\sigma_2 = (s_{2,1}, s_{2,2})$, $\sigma_3 = (s_{3,1}, s_{3,2})$, $\sigma_5 = (s_{5,1}, s_{5,2}, s_{5,3})$.

The question considered: Does the AGVs system considered can be deadlock-free? and if so: What is the cycle of a steady-state? Due to solution of the relevant problems: CS'_{Sx1}, CS'_{Sx2} the set of alternative sufficient conditions is determined. The condition (19) illustrates the one among the obtained five variants.

$$(S_0=(R_2, R_5, R_3), \{\sigma_5=(P_2, P_1, P_3), \sigma_2=(P_1, P_2), \sigma_3= (P_3, P_2)\}). \quad (19)$$

Because of the size of the set obtained (containing 13500 solutions) the calculated fuzzy variable H is represented in a graphic form, see Figure 2b. Fields limited by the bold lines determine the space of the variable H values and membership fuzzy variables that can be obtained for the given imprecise operation times. From Fig. 2b it follows that for the given sufficient condition (19) the cyclic steady state should be something about 9 units of time.

7 Conclusions

The most important advantage of proposed (so called reverse) approach is the possibility of prototyping of the collision-free and deadlock-free AGVs systems operation assuming imprecise operation times. This advantage can be seen as a possibility to response to questions like: What values and of what variables guarantee the AGVs system will operated due to assumed values of performance indexes? What are decision variable domains sufficient for required values of the system performance indexes?

The results obtained can be also seen as a contribution to multi-agent system distributed control, i.e. to the cases when autonomously acting agents have to cooperate under limited resource constraints.

References

1. Banaszak, Z., Zaremba, M., Muszyński, W.: CP-based decision making for SME. In: Horacek, P., Simandl, M. (eds.) 16th IFAC World Congress, P. Zitek, DVD, Prague, Czech Republic (Preprint, 2005)
2. Barták, R.: Incomplete Depth-First Search Techniques. In: Figwer, J. (ed.) A Short Survey, Proceedings of the 6th Workshop on Constraint Programming for Decision and Control, pp. 7–14 (2004)
3. Bocewicz, G., Wójcik, R., Banaszak, Z.: Scheduling of AGV systems under access restrictions to shared resources of the ESP (logic-algebraic model). In: Tchoń, K., Wroclaw, W.K.Ł. (eds.) Proc. of National Conf. on Robotics, pp. 149–163 (2006) (in Polish)
4. Bubnicki, Z.: Introduction to Expert Systems. PWN, Warsaw (1990)
5. Bubnicki, Z.: Learning processes and logic-algebraic method for the systems with knowledge representation. Systems analysis and management. Systems Research Inst. PAS (1999)
6. Dubois, D., Fargier, H., Fortemps, P.: Fuzzy scheduling: Modeling flexible constraints vs. coping with incomplete knowledge. European Journal of Operational Research 147, 231–252 (2003)
7. Lawley, M.A., Reveliotis, S.A., Ferreira, P.M.: A correct and scalable deadlock avoidance policy for flexible manufacturing systems. IEEE Trans. on Robotics and Automation 14(5), 796–809 (1998)
8. Lee, T., Song, J.: Petri net modeling and scheduling of periodic job shops with blocking. In: Proc. of the Workshop on Manufacturing and Petri nets, Osaka, Japan, June 25, 1996, pp. 197–214 (1996)
9. Piegat, A.: Fuzzy modeling and control, Exit, Warsaw (1999)
10. Polak, M., Majdzik, P., Banaszak, Z.A., Wójcik, R.: The performance evaluation tool for automated prototyping of concurrent cyclic processes. Fundamenta Inf. 60(1-4), 269–289 (2004)
11. Ramamritham, K.: Allocation and scheduling of precedence related periodic tasks. IEEE Trans. on Parallel and Distr. Systs. 4(6), 412–420 (1995)
12. Xue, X., Lu, J., Wang, Y., Shen, Q.: Towards an Agent-Based Negotiation Platform for Cooperative Decision-Making in Construction Supply Chain. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, Springer, Heidelberg (2007)
13. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)

On the Multi-threading Approach of Efficient Multi-agent Methodology for Modelling Cellular Communications Bandwidth Management

P.M. Papazoglou[1,3], D.A. Karras[2], and R.C. Papademetriou[3]

[1] Lamia Institute of Technology, Greece
papaz@teilam.gr
[2] Chalkis Institute of Technology, Greece, Automation Dept., Psachna, Evoia, Hellas (Greece) P.C. 34400
dakarras@teihal.gr, dakarras@ieee.org
[3] University of Portsmouth, UK, ECE Department, Anglesea Road, Portsmouth, United Kingdom, PO1 3DJ

Abstract. There are very few studies using the multi agent concept as an alternative and efficient approach for bandwidth management. The goal of the paper is to propose a novel modelling methodology of wireless network services exploiting multi-agent technology and investigating in depth critical implementation issues of multi-threading techniques focused on network agents. Moreover, it is claimed that an advanced negotiation scheme between network service agents can lead to better model network performance. Thus, the multi agent concept seems a suitable solution for modelling and implementing cellular services simulation. The proposed agent based modelling as well as the proposed implementation methodology of the wireless services are herein presented in three levels; (a) multi agent modelling of the wireless services, (b) developing an advanced negotiation scheme based on current status and curve statistics of network performance and (c) investigating critical multi threading based implementation issues of the suggested multi agent modelling. Finally, simulation results show the effectiveness of the proposed multi agent approach as well as of the proposed negotiation scheme based on blocking and dropping probability curve statistics.

1 Introduction

Multi-Agent Systems and channel allocation in Wireless communication Systems

Several strategies have been proposed in the literature for efficient channel allocation. The most known channel allocation strategies are categorized as Fixed Channel Allocation (FCA) [1-5], Dynamic Channel Allocation (DCA) [1,6-9] and Hybrid Channel Allocation (HCA) [1,10] respectively. The main difference between these approaches is how the available channels are allocated to network cells and how this allocation takes place (statically or dynamically). A set of channels are permanently allocated to each cell based on a pre-estimated traffic intensity in the case of FCA. The FCA strategy can not be adapted to changing network traffic conditions. To solve

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 431–443, 2008.

this problem, the DCA strategy is used. According to DCA, the entire set of available channels is accessible to all the cells, and the channels are allocated on a call-by-call basis in a dynamic manner. To achieve the highest adaptability to dynamically changing network conditions, the HCA combines the features of both FCA and DCA techniques.

On the other hand, the Multi-Agent concept is a very promising approach for building specialized models for efficient channel allocation. Several agent features such as adaptability [11,12], autonomy [13-15], collaboration, interactivity, etc make the multi-agent concept very attractive but is useful only when these features can be practically applied. Software agents can be used in communication management [16]. Cell participation supports the call admission in [17] and a cooperative negotiation supports real-time load balancing of a mobile cellular network in [18]. Collaborative software agents give autonomy to base stations, allow negotiation of network resources and improve resource allocation according to [19].

2 The Proposed Multi-agent Model Implementation through Multi-threading Technology

2.1 Principles of the Suggested Multi-agent Simulation Model

The modelled cellular network consists of N hexagonal areas called cells (fig. 1). Each cell has in its center a base station which manages the available channels for servicing the mobile users of the local area.

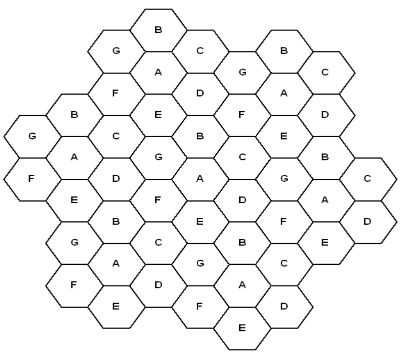


Fig. 1. The network structure

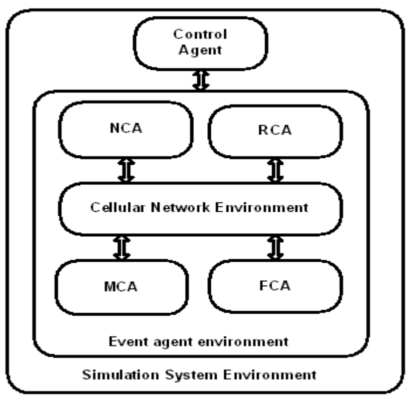


Fig. 2. Multi-Layered / Multi-Agent Architecture

Due to the restricted bandwidth, the same sets of frequencies are assigned in multiple cells that belong to different areas called clusters in order to minimize the total interference between co-channel users. The four network services for mobile users supported in the proposed simulation model are:

- New call arrival (NC). When a new call admission is occurred and a number of channels must be allocated.

- Call termination (FC). An ongoing call is terminated when the call holding time is expired or the total amount of the requested data is transferred.
- Reallocation check (RC). Each connected mobile user is checked for the signal quality (CNR between base station and mobile user and interference from other co-channel users). When the final CNIR (Carrier to Noise plus Interference ratio) is bellow threshold, a channel reallocation is made.
- MU movement (MC). In the case of a user movement, channel reallocation takes place when is needed. The corresponding service moves any connected mobile user through a move pattern.

Each network service is modelled as a distinct agent that its behaviour is defined by specific rules. All the agents interact with one another within the cellular network environment. In a large scale wireless network, a distinct set of the four agents is distributed in each network cluster where a set of different frequencies is used. The basic multi-agent simulation model is structured as a multi-layered model where each layer represents a different functionality. Figure 2, shows the multi-layered model.

2.2 The Proposed Multi-agent Negotiation Strategy

The main goal of the proposed system is to keep balanced the network performance in terms of blocking and dropping probability. The NCA is responsible for new call admissions and so for the resulted blocking probability. On the other hand, the RCA is responsible for channel reallocation and so for the dropping probability. Negotiation between NCA and RCA must be established in order to give call priority when either blocking or dropping probability is at critical level. This negotiation is cooperative or competitive based on the current network performance.

The negotiation and the final decision for the call priorities can be based on current and previous network performance in terms of statistical behaviour of blocking and dropping probability. When current performance is used, the comparison is made between current ***t*** and previous ***t-1*** simulation steps. When previous performance comparison is used, the negotiation algorithm takes in consideration two overlapping and moving windows among the previous simulation statistical data. In other words, ***n*** points of previous blocking/dropping probability curve are statistically analyzed for taking the final decisions about next call priorities. In this approach standard deviation and mean value progress of the selected statistical metrics (e.g. mean value of blocking pr., std of dropping pr., periodicity of blocking pr., etc) are examined. Using standard deviation (std.), the desired stability for a specific metric can be achieved. On the other hand, the mean value can be used in order to limit statistical metrics below a predefined threshold.

The negotiation starts with a dialog between NCA and RCA and is terminated with acceptance or rejection of the corresponding requests between the agents. Just before the negotiation, the NCA and RCA status is examined respectively. Tables 1 and 2 show the corresponding status of the two agents based on current blocking and dropping probability respectively.

Based on the above status, each agent takes decision for its status and sends a request to the other agent. If the current status of the first agent is good (1) and the

Table 1. NC Status description

NCA Status	Description
1	Blocking Pr.(t)<Blocking Pr.(t-1)
0	Blocking Pr.(t)=Blocking Pr.(t-1)
-1	Blocking Pr.(t)>Blocking Pr.(t-1)

Table 2. RC Status description

RCA Status	Description
1	Dropping Pr.(t)< Dropping Pr.(t-1)
0	Dropping Pr.(t)= Dropping Pr.(t-1)
-1	Dropping Pr.(t)> Dropping Pr.(t-1)

status of the second agent is not good (-1), then the second agent requests from the first agent to decrease its priority. Thus, the request is accepted from the first agent and so, the negotiation is cooperative. On the other hand, if both status are not good (-1,-1) then, the negotiation is competitive due to the fact that the agents act only for their own goal. Figure 3 shows the two overlapping and moving windows across the statistical curve.

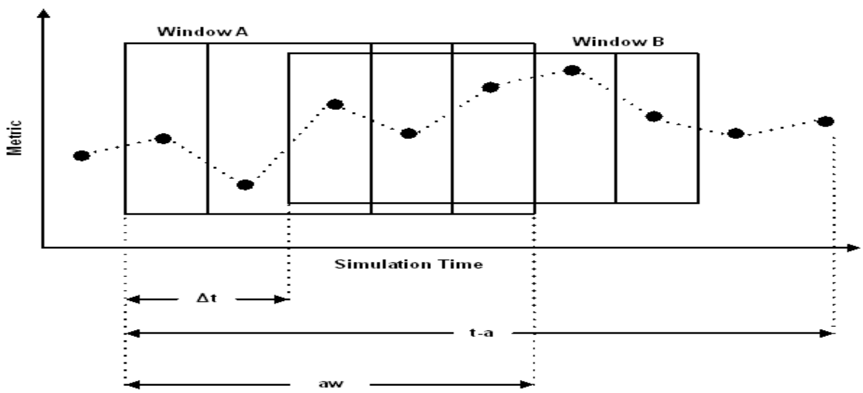


Fig. 3. Overlapping and moving windows across the statistical curve

Window A starts at ***t-a*** and ends at ***t-a+aw*** and window B starts at ***(t-a)+Δt*** and ends at ***(t-a+aw)+Δt***. The corresponding ratios are calculated between the two moving windows. When the stability of the blocking/dropping curve is examined, the NCA and RCA status are calculated as follows (tables 3 and 4):

Table 3. NC Status description

NCA Status	Description
1	std B blocking < std A blocking
0	std B blocking = std A blocking
-1	std B blocking > std A blocking

Table 4. RC Status description

RCA Status	Description
1	std B dropping < std A dropping
0	Std B dropping = std A dropping
-1	Std B dropping > std A dropping

Based on the mentioned tables, each priority for the next simulation step is calculated as follows:

$$NewP = \left\lceil NP \cdot \left[\left(1 + \left(mratio - 1\right)\right) \right] \right\rceil$$

where, ***NP*** represents the normal priority and ***mratio*** the ratio of the selected metric between the two overlapping and moving windows. When ***NewP*** exceeds the range [1,10], the resulted priority is pushed to minimum or maximum allowed limit.

2.3 Multi-threading Based Critical Model Implementation Issues

Creating network agents as Threads
Using multi-threading technology [20-22], the agent independency with private behavioural rules can be effectively supported. Multi-Threading methodology is a very efficient tool for a plethora of scientific tasks such as simulations of real phenomena where many processes occur at the same time (simultaneously), offering a more reliable modelling and simulation for the real system under investigation. Simulation of Cellular Systems bandwidth management presents such characteristics.
The above mentioned network agents can be created as threads within a Java application (fig.4). Any application constitutes at least one thread. The main method of a thread is the ***run()*** method. When the method ***start()*** is activated for a thread then, the corresponding ***run()*** method is executed within the thread body. Figure 4 illustrates the four basic threads created within a Java application.

Figure 5 shows the main structure of the RCA. Each thread is implemented as a different class with a private ***run()*** method which is activated through the ***Threadname.start()*** method within the main thread of the application. After the ***start()*** invocation, the corresponding thread is active during the simulation time and so the rest of the created threads (agents) are also active. A special purpose flag called signal that is controlled by the main clock informs each thread that is time for internal code activation. The flag ***s2*** has the initial value 0 in order to satisfy the ***if*** condition for the main code activation inside thread. After the thread execution completion (for current simulation step), the main clock is informed through flag ***s2*** which takes the value 1.

When a deterministic thread scheduler is used (JVM scheduler inactive), the flag *signal* is removed and the thread is controlled with *Threadname.suspend()* and *Threadname.resume()* methods for a specific time slice. Using deterministic thread scheduling, the network events are fully controlled and the concurrency is efficiently supported.

Agent dialog and decision
When the network agents basic action (channel reallocation, new call admission, etc) is completed, each agent (e.g. NCA/RCA) sends a message (request) to the other agent based on its status (see tables 1 and 2). After the agents action completion the negotiation modules are activated concurrently. Figure 6 illustrates the basic structure of the RCA negotiation module.

Initially, the RCA-NM checks for any incoming message from NCA. If an incoming message exists, RCA-NM checks for its status that has been determined within the RCA body. If RCA status is not critical, then its self priority is decreased in order to help the NCA (cooperative negotiation). Otherwise, if the RCA status is critical, then the request for priority decrement that is coming from NCA is rejected while its self priority has been already increased (competitive negotiation) within the

RCA body. Figure 7 shows a sample of the negotiation dialog between RCA and NCA. In this dialog, the RCA sends a message (request) to NCA. The response from NCA is based on its current status.

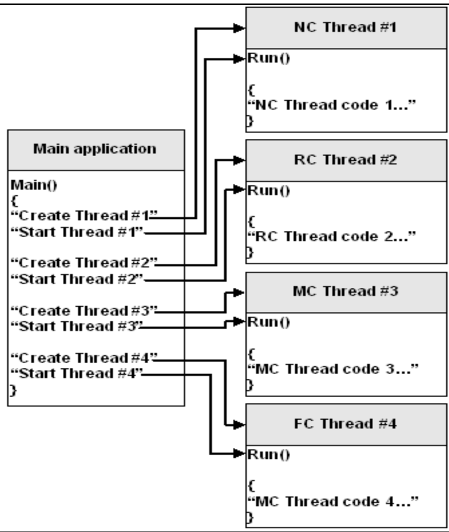


Fig. 4. Creating threads

```
public class reallocation extends Thread
{
/* RCA */
        public void run()
        {
          while (timenow < timeend) /* WITHIN
SIMULATION TIME*/
          {
            /* CHECK ACTIVATION BY
CONTROL/CLOCK AGENT*/
            if ((signal==2) && (s2==0))
            {
                        /* main thread code */
                        s2=1;
               }//if
          }//while
        }//run()
}//class
```

Fig. 5. General structure of the RCA

```
public class new_rc_negotiation
{
  public void run()
     {/*CHECK FOR ANY RECEIVED
MESSAGE FROM NCA*/
         if (RC_RECEIVE==1)
         {
     /*IF SELF STATUS IS NOT CRITICAL*/
           if (RC_status>-1)
           {
             cpri=RCpriority; /* GET SELF
PRIORITY */
    /*IF SELF PRIORITY IS NOT MINIMUM*/
             if (cpri>1)
             { /* DECREASE SELF PRIORITY TO
HELP NCA */
               rcall.setPriority(cpri-1);
RCpriority=cpri-1;
   }//if
           }//if
         }//if
       }//run()
     }//class
```

Fig. 6. Basic structure of the RCA negotiation module

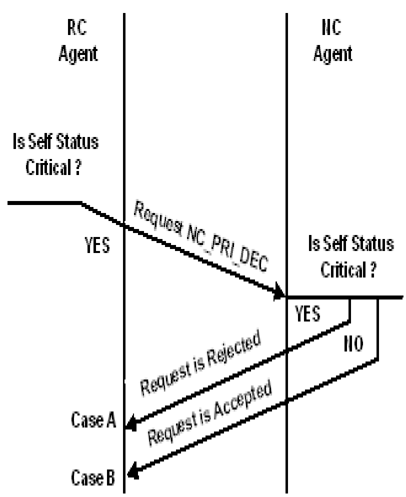


Fig. 7. Sample of the Negotiation dialog between RCA and NCA

Simulation model architecture
The main goals of the implemented architectures are the investigation of (a) the effectiveness of multi-agent concept based on multi-threading technology for modelling and simulation the experimental wireless network compared to a single multi-threaded approach and (b) the features and difficulties of the multi-threading implementation during the development phases.

Single multi-threaded implementation architecture (does not support agent negotiation)
Figure 8 shows the implementation architecture that is based on the four main threads. In this implementation there are three additional threads which are:

(a) the controller which contains the clock and synchronizes the main and supplementary threads, (b) the initialization thread that prepares the simulation environment for the next simulation step and (c) the finalization thread that completes each finished simulation step. NC and FC threads use two different distributions for new call generation and mobile user movement respectively. After thread activation, several procedures take place such as signal measurements which include CNR calculation between base station and mobile user and the final CNIR which is based also on the interference caused by other co-channel users. These measurements are necessary in order to manipulate correctly the reserved and available channels. A different instance of the necessary signal measurement procedures exists within each thread in order to give higher grade of concurrency and independency. On the other hand, a user registry (UR) exists where the most critical data of mobile users such as current position, given channels, requested capacity, etc are stored. The user registry can be accessed by any active thread and so a synchronization mechanism must be developed and applied. To avoid unwanted synchronization problems, a local user registry has been built for each thread. Thus, when a thread tries to access the main user registry, the corresponding data are stored firstly in the local user registry and they are published to the other threads. If two or more threads try to access the main user registry for changing data for the same user, then this access is completed through the mutual excluded area, otherwise the access is completed directly. Due to the fact that the access of user registry for the same mobile user is quite rare, this approach eliminates any possibility for deadlock where two or more threads are waiting each other for accessing the common data area.

Multi-agent model implementation based on multi-threading technology (it does support agent negotiation)
The above implementation architecture has been extended for supporting multi-agent behaviour including the multi-agent negotiation capabilities. Thus, an extended multi-layered architectural model (fig. 9) which includes additional negotiation modules for each agent has been developed. The additional negotiation modules constitute the implementation of the above presented negotiation rules.

Synchronization issues
As mentioned before, the simulation system uses a UR for keeping a detailed record of each connected MU. Due to that fact, the UR constitutes a shared resource area. When concurrent events take place, two or more threads try to access the UR at the

"same time". An active thread can be pre-empted by the scheduler when an access activity in a shared resource is not completed. While this thread is now pre-empted, another thread tries also to access the shared resource. Due to the time slice given to that thread by the scheduler the access activity is completed. After the re-activation of the first thread, the semi-completed access activity of the first phase is now complete. If the above two threads try to access the same record in the UR, then the resulted data are incorrect depending on the thread switching. For the above reasons, the UR of the simulation system must be accessed through synchronized method. This synchronization will prevent the shared resource area from simultaneous access by two or more threads. When a thread has locked an object (e.g. access method for UR), and is waiting for another thread to finish, while that other thread is waiting for the first thread to release that same object before it can finish then a deadlock occurs.

In the classical thread synchronization approach signal calculation procedures and the UR access methods (Write/Read) are completely synchronized (fig. 10)

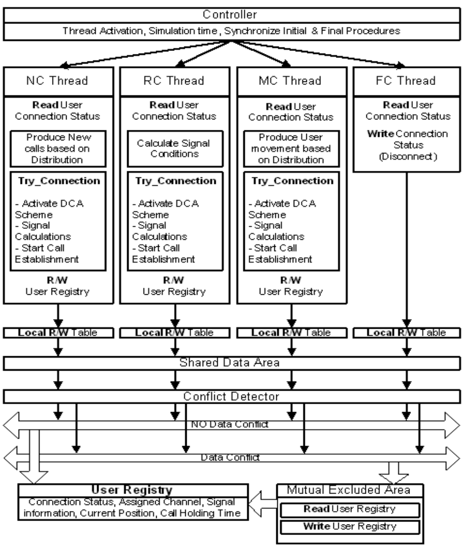


Fig. 8. Single multi-threaded architecture

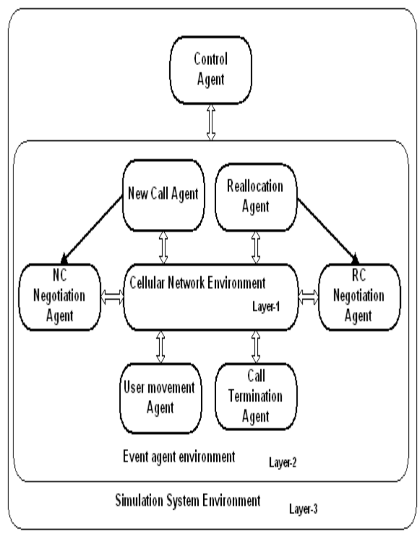


Fig. 9. Extended multi-layered model for supporting multi-agent negotiation

```
public synchronized float READ_UserRegistry(…) {…}
public synchronized void WRITE_UserRegistry(…) {…}
public synchronized int try_connection(…) {…}
```

Fig. 10. Classical synchronization approach

Another simple approach is the use of two only synchronized methods for accessing the user registry (Read/Write). Due to that fact, all the thread load competition for accessing common resources is concentrated on the two methods. This synchronization scheme causes more deadlocks than the initial approach in figure 10 due to the fact that the active threads try to access always the same

```
/* accessing UR service from NC thread */
    synchronized(synch)                {a=synch.READ_UserRegistry(…);              …
synch.WRITE_UserRegistry(…);}
/* accessing UR service from RC thread */
    synchronized(synch) {… synch.WRITE_UserRegistry(…);}
    public float READ_UserRegistry(…) {…}
    public void WRITE_UserRegistry(…) {…}
    public int try_connection(…) {a=synch.READ_UserRegistry(…);
    …synch.WRITE_UserRegistry(…);}
```

Fig. 11. Multiple synchronized blocks

synchronized methods. To minimize the deadlock possibility, multiple synchronized blocks can be used instead of complete method synchronization. Figure 11, shows how the multiple synchronized blocks are used. This approach can solve many deadlock problems but can not eliminate them. Thus, the proposed implementation architecture is the most efficient for deadlock avoidance due to the fact that uses conditional synchronization.

Thread timing behaviour
In [23] a centralized cooperative multi-agent negotiation scheme applied to a multi-agent layered architecture for designing and simulating resource allocation in cellular communication systems has been presented. Simulation results in [23] have shown that controlling thread timing behaviour might be a negotiation mechanism providing fairly better results than traditional modelling approaches but on the other hand, that behaviour could not be controlled completely. Thus, in the present paper the proposed negotiation scheme is based on selected performance metrics instead of thread timing features as a much more effective negotiation mechanism improving significantly service performance results.

2.4 Scaling Up

In large scale network where the coverage area is divided in clusters (fig. 12), a set of agents must be assigned for each cluster. When the network consists of ***N*** cells distributed in clusters of ***i*** cells each, the resulted total number of cluster is ***N/i*** and the total required agents are ***4*(N/i)***.

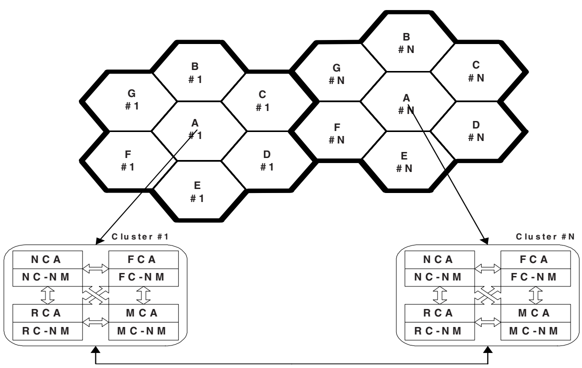


Fig. 12. Scaling up

3 Evaluation of the Model

The two most widely known evaluation metrics for the performance of a wireless network are the blocking and dropping probability. The blocking probability represents the number of blocked calls compared to the total new call requests. A call is blocked when the wireless network can not allocate the requested channels. On the other hand, the dropping probability constitutes another critical performance metric for the wireless network. This probability represents all the dropped calls when the network can not allocate any channels to ongoing calls. Through Monte Carlo simulation executions, the statistical behaviour of the simulation model can be evaluated. In this study, two statistical characteristics of the simulation model are investigating based on the two mentioned probabilities. The first characteristic is the model stability which can be evaluated through the standard deviation progress of the blocking and dropping probability and the second is the periodicity of the model that is calculated between the simulated operational days of the simulation model.

Assume that the blocking probability within the 24 day-hours is represented by the sets setA and setB for two days respectively. Each set has probability members 1 through n (A1 to An and B1 to Bn). The elements of the two sets are subtracted one by one (|A1-B1|, |A2-B2|,…,|A(n-1)-B(n-1)|, |An-Bn|). Thus, a set with the blocking probability differences BPDk for every hour within the pair of days is formulated:

$$BPD_k[i] = \left| seta[i] - setb[i] \right| \quad (1)$$

where k is the selected pair and i is the i-element of each set. The difference seta[i]-setb[i] is calculated only if the seta[i] and setb[i] values are not outliers otherwise the data pairs are rejected. The mean Mk is calculated for all the possible pairs :

$$M_k[] = \frac{\sum_{i=1}^{n} BPD_k[i]}{n} \quad (2)$$

Finally, the periodicity measurement by calculating the mean value among the mean Mk is performed. When the difference between elements within pairs of different days tends to zero, the pairs tend to be identical and so the periodicity factor tends to zero for the given pair. When higher periodicity is resulted (smaller periodicity factor), then the behaviour of simulation model is more stable among the daily model in simulation time.

4 Simulation Results

The experimental results have been generated using Monte Carlo executions. The experiments conducted are divided in two categories. First, we investigate the performance of MT EI Model (Multithreaded Event Interleaving model that does not support agent negotiation) versus the MT MAS EI Model (Multithreaded Multi-agent System Event Interleaving model that does support agent negotiation)- ***(Single multi-threaded model (MT EI Model–Multithreaded Event Interleaving model that does not support agent negotiation) versus multi-threaded and multi-agent model (MT***

MAS EI Model - Multithreaded Multi-agent System Event Interleaving model that does support agent negotiation)). More specifically, the mean or STD (standard deviation) values of the blocking and dropping probabilities resulted by applying the above specified models are examined. In the second series of experiments and results the kind of negotiation model is investigated. The DCA algorithm involved in all experiments is the classic DCA [1].

Negotiation based on statistical curve calculations

In the second type of negotiation, the standard deviation metric has been used. Even with small moving window size, the resulted stability of the simulation model is

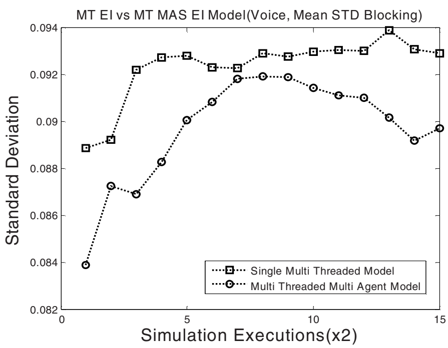


Fig. 13. Single Multi Threaded Event Interleaving Versus Multi Threaded Multi Agent Event Interleaving model (Voice, Mean STD Blocking Probability)

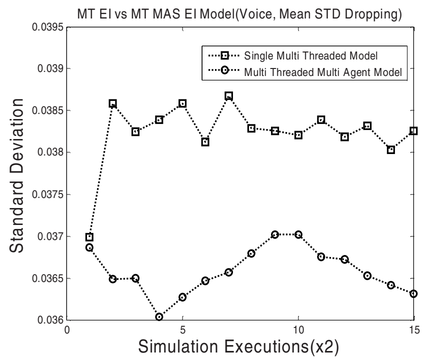


Fig. 14. Single Multi Threaded Event Interleaving Versus Multi Threaded Multi Agent Event Interleaving model (Voice, Mean STD Dropping Probability)

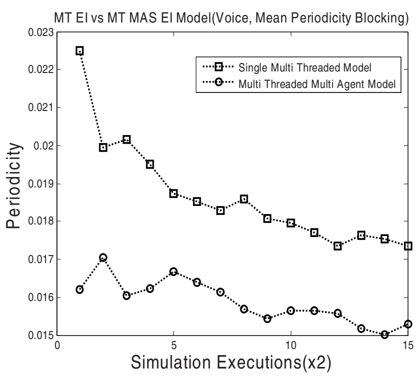


Fig. 15. Single Multi Threaded Event Interleaving Versus Multi Threaded Multi Agent Event Interleaving model (Voice, Mean Periodicity Blocking Probability)

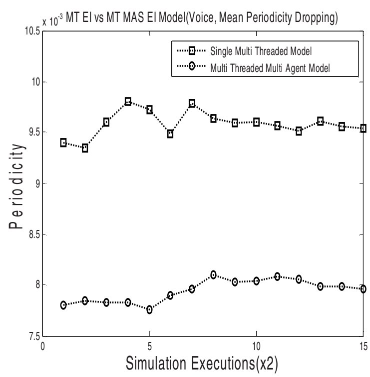


Fig. 16. Single Multi Threaded Event Interleaving Vs Multi Threaded Multi Agent Event Interleaving model (Voice, Mean Periodicity Dropping Probability)

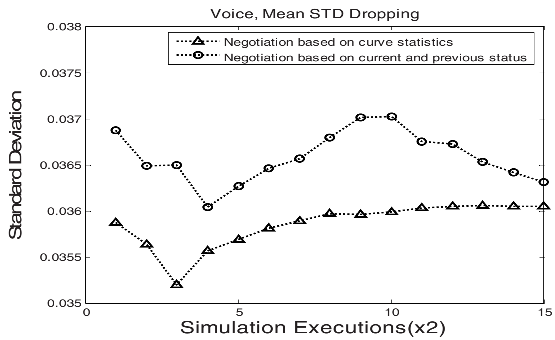


Fig. 17. Standard deviation of dropping probability based on the two types of negotiation

better compared to the single ratio of current and previous state. Thus, more accurate prediction of the future model behaviour can be made. Figure 17 shows the model performance based on the two types of negotiation. When curve statistics are used, the windows width are two points and Δt=1.

5 Conclusions and Future Work

This paper investigates an integrated approach towards efficiently modelling and implementing simulation of wireless communications bandwidth management, through applying a Multi-agent approach based on a novel negotiation scheme for user services. Simulation modelling and implementation methodologies involving a novel multi-agent system are herein presented. Moreover, two negotiation techniques for achieving dynamic adaptation of the model to current user services are investigated. When the negotiation scheme based on curve statistics is applied, model performance is increased due to better predictability of network behaviour. Implementation methodology of the multi agent model is shown to be a critical choice concerning design and performance. The proposed implementation is based on Java Multi-threading introducing various drawbacks to be faced effectively.

References

[1] Zhang, M., Yum, T.S.: Comparisons of Channel Assignment Strategies in Cellular Mobile Telephone Systems. IEEE Tran. Vehicular Technology 38(4), 211–215 (1989)

[2] Lai, W.K., Coghill, G.C.: Channel Assignment through Evolutionary Optimization. IEEE Tran. Vehicular Technology 45(1), 91–96 (1996)

[3] MacDonald, V.H.: The cellular Concepts. The Bell System Tech. Journal 58, 15–42 (1979)

[4] Elnoubi, S.M., Singh, R., et al.: A New Frequency Channel Assignment Algorithm in High Capacity Mobile Communication Systems. IEEE Tran. Vehicular Tech. VT-21(3), 125–131 (1982)

[5] Xu, Z., Mirchandani, P.B.: Virtually Fixed Channel Assignment for Cellular Radio-Telephone Systems: A Model and Evaluation. In: IEEE Int. Conf. Communications, ICC 1992, vol. 2, pp. 1037–1041 (1992)

[6] Cimini, L.J., Foschini, G.J.: Distributed Algorithms for Dynamic Channel Allocation in Microcellular Systems. In: IEEE Vehicular Technology Conference, pp. 641–644 (1992)
[7] Cox, D.C., Reudink, D.O.: Increasing Channel Occupancy in Large Scale Mobile Radio Systems: Dynamic Channel Reassignment. IEEE Tran. Vehicular Tech. VT-22, 218–222 (1973)
[8] Del Re, E., Fantacci, R., et al.: A Dynamic Channel Allocation Technique based on Hopfield Neural Networks. IEEE Tran. Vehicular Technology VT-45(1), 26–32 (1996)
[9] Sivarajan, K.N., McEliece, R.J., Ketchum, J.W.: Dynamic Channel Assignment in Cellular Radio. In: IEEE 40th Vehicular Technology Conference, pp. 631–637 (1990)
[10] Kahwa, T.J., Georgans, N.D.: A Hybrid Channel Assignment Schemes in Large-Scale, Cellular Structured Mobile Communication Systems. IEEE Tran. Comm. 26, 432–438 (1978)
[11] Splunter, S., Wijngaards, N., et al.: Structuring Agents for Adaptation. In: Alonso, E., Kudenko, D., Kazakov, D. (eds.) AAMAS 2000 and AAMAS 2002. LNCS (LNAI), vol. 2636, pp. 174–186. Springer, Heidelberg (2003)
[12] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Prentice-Hall, Englewood Cliffs (2002)
[13] Huhns, M., Singh, M. (eds.): Agents and Multiagent Systems: Themes, Approaches, and Challenges. Readings in Agents, ch. 1, pp. 1–23. Morgan Kaufmann Publishers, San Francisco (1998)
[14] Norman, T., Long, D.: Goal Creation in Motivated Agents. In: Wooldridge, M.J., Jennings, N.R. (eds.) ECAI 1994 and ATAL 1994. LNCS, vol. 890, Springer, Heidelberg (1995)
[15] Ekdahl, B.: How Autonomous is an Autonomous Agent? In: SCI 2001, July 22-25, 2001, Orlando (2001)
[16] Hayzelden, A., et al.: S/W Agents for Future Communication Systems. Springer, Heidelberg (1999)
[17] Iraqi, Y., Boutaba, R.: A Multi-agent System for Resource Management. In: Ambler, A.P., Calo, S.B., Kar, G. (eds.) DSOM 2000. LNCS, vol. 1960, pp. 218–229. Springer, Heidelberg (2000)
[18] Bigham, J., Du, L.: Cooperative Negotiation in a MultiAgent System for Real Time Load Balancing of a Mobile Cellular Network. In: AAMAS 2003, July 14–18 (2003)
[19] Bodanese, E.L.: A Distributed Channel Allocation Scheme for Cellular Networks using Intelligent Software Agents, PhD thesis, University of London (2000)
[20] Oaks, S., et al.: Java Threads, 3rd edn. O'Reilly, Sebastopol (2004)
[21] Kramer, J.M.: Concurrency: State Models & Java Programs, 2nd edn. John Wiley & Sons, Chichester (2006)
[22] Lindsey, C.S., et al.: An Introduction to Scientific and Technical Computing with java. Cambridge University Press, Cambridge (2005)
[23] Papazoglou, P.M., Karras, D.A., et al.: A Multi-agent Architecture for Designing and Simulating Large Scale Wireless Systems Resource Allocation. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 405–415. Springer, Heidelberg (2007)

Design of the Architecture for Agent of the Personality Security in the Ubiquitous Environment*

Eun-Ser Lee[1] and Sang Ho Lee[2]

[1] Computer Engineering, Andong National University
emp777@daum.net
[2] College of Information Science School of Computing, Soongsil University
shlee@comp.ssu.ac.kr

Abstract. Many programs are exposed a threat in the ubiquitous system. In the result, we are attacked and discharge personal information from the many factors. When executing the various application programs of a computer generally or connecting with a specific site through the Internet A place with many examples into which carry out a user and the user-identification child (ID) and the password are made to enter in order to check a user's identity, If in agreement as compared with the identifier which a user inputs, ID which had the password registered, and a password, the authority to perform an applicable program, and the authority to use the various contents of an applicable site will be granted to the user. In this paper is provided the design of the user identification method in the ubiquitous environment. Also, We are make use of grid computing between host and targets. Therefore, This system is provide for the flexibility in the heterogeneous system.

1 Introduction

In 1991, Mark Weiser, then chief technology officer for Xerox's Palo Alto research Center, described a vision for 21st computing that countered the ubiquity of personal computers. "The most profound technologies are those that disappear, "he wrote." They weave themselves into the fabric of everyday life until they are indistinguishable from it"[14]. Ubiquitous computing is the method of enhancing computer use by making many computers available throughout the physical environment, but making them effectively invisible to the user. Since we started this work at Xerox PAPC in 1988, a number of researchers around the world have begun to work in the ubiquitous computing framework[15].

We still view computers primarily as machines that run programs in a virtual environment that we enter to perform a task and leave when we finish. Pervasive computing presumes an altogether different vision.

* In this paper of data was provided by PassHUB CO.,LTD. This work was supported by the Korea Research Foundation Grant funded by Government (MOEHRD) (KRF-2006-0050J03803).

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 444–453, 2008.

When executing the various application programs of a computer generally or connecting with a specific site through the Internet A place with many examples into which carry out a user and the user-identification child (ID) and the password are made to enter in order to check a user's identity, If in agreement as compared with the identifier which a user inputs, ID which had the password registered, and a password, the authority to perform an applicable program, and the authority to use the various contents of an applicable site will be granted to the user[1]. However, it sets to a bank, a card issuer, etc. who the correlation of the aforementioned Prior art of a password is fixed with a user-identification child, and manage many identifiers. The identifier which is in agreement with an applicable password regardless of the count limit of an input error if many identifiers to a fixed password are changed and it continues inputting even if it restricts the password input error to a fixed identifier can be discovered. ** -- Since it has the regulation to which anyone can also learn how to enter an identifier and a password, it tries to steal, without being noticed by the user by a user's side, or a camera is installed that it seems that it is easy, Even if it is not him, appearance of the password to an identifier is investigated and carried out. if the password which looks at that a user enters an identifier and a password directly, checks the program which records the contents to input, carries out usurpation and tapping on a communication path, or is kept is stolen -- a user -- There was a problem that a user's authority could be plagiarized now[2].

This research is intend representing of the ubiquitous system for the security of the personal information. Also, in this paper explains what is method and evaluation about the personal information security in ubiquitous computing system.

2 Related Work

2.1 Ubiquitous System

Ubiquitous computing, or ubicomp, systems designers embed devices in various physical objects and places. Frequently mobile, these device (such as those we carry with us and embed in cars) are typically wirelessly networked. Some 30 years of research have gone into creating distributed computing systems, and we've invested nearly 20years of experience in mobile computing[16].

A ubicomp system involves some integration between computing nodes and the physical world. For example, a smart coffee cup, such as a Media cup, serves as a coffee cup in the usual way but also contains sensing, processing, and networking elements that let it communicate its state(full or empty, held or put down). So, the cup can give colleagues a hint about the state of the cup's owner. Or consider a smart meeting room that senses the presence of users in meetings, records their actions, and provides services as they sit at a table or talk at whiteboard. The room contains digital furniture such as chairs with sensors, whiteboards that record what's written on them, and projectors that you can activate from anywhere in the room using a PDA(personal digital assistant)[3][4].

2.2 Security

Looked at from a higher level, users require security for their resources and, in many cases, privacy for themselves. Mobile computing has led to an under standing of vulnerabilities such as the openness of wireless networks. But physical integration and spontaneous interoperation raise new challenges, requiring new models of trust and authentication as well as new technologies[3][4][10].

In ubiquitous system, trust is an issues because of spontaneous interoperation. Components might belong to disparate individuals or organizations and have no relevant a priori knowledge of one another or a trusted third party. Fortunately, physical integration can work to our advantage – at least with an appropriate placement o the semantic Rubicon. Humans can make judgments about their environments' trustworthiness, and the physical world offers mechanisms for bootstrapping security based on that trust[7][8][9]. For example, users might exchange cryptographic keys with their environment or each other over a physically constrained channel such as short-range infrared, once they have established trust[5][6].

2.3 Grid Computing

Recently grid computing has received a lot of attention under the influence of the far quicker development in application software than computer hardware. Grid computing can be understood as the expanded parallel computing technology for wider usages that has been used in multi process system or cluster system and etc[17].

Even though using the past parallel computing technology, grid computing has expanded the limit of application in respect of equipments and management. In respect of equipment, it enables from mobile equipments such as PDAs to the most expensive super computers to operate as one virtual system. It is possible because of grid middleware that is a core in grid computing. Grid middleware achieves the cooperation of various equipments by using open, general-purpose grid standard. The increased availability of equipments means the expansion of the limit of problem solving. For instance, the volume of available computing resources is rapidly getting lager as wide-spread personal computers in the world are connected to grid system[18].

In respect of management, Grid computing introduces the concept of virtual structure. Virtual structure is what the available equipments are being grouped in order to solve a certain problem. The permits from an individual or institution holding equipments are necessary for grouping. Contracts are formed between each individual and institution. The usage of equipments in virtual structure is strictly controlled by such a contract. The concept of virtual structure is inevitably necessary in grid computing because it often can not get the sufficient computing resources from the equipments of an individual or an institution[19].

Available resources for the share could be various according to the requirement of application. Most of all, we can mention the resource of processing unit. Processing unit can be used naturally as application code is carried out in shared computers. Memory resources are as same as processing unit. Besides, whatever computer components that can operate independently, even special equipments connected to computer, could be shared and used. Not only hardware but also software such as program or data could be used as grid resources. Grid standard enables user to share various kinds of resources by defining in person in standardized and general-purpose way. The parallel computing technology can be applicable or inapplicable according to the characteristics of application itself. Such characteristics are exposed in grid computing as well. However, the paralleling is more likely possible because most applications needing grid computing require vast resources and time[20].

Grid computing can be used in energy, financial services, manufacturing, bioinformatics, telecom, government, education and so on. Such areas have been benefited from grid computing technology using high computing capacity and broad data in order to solve various difficult problems restricted by time and resources. New area has been continuously explored for the usage of grid computing. It has been expected for the appearance of new killer applications.

3 Main Theory

3.1 Overview

In this chapter is consist of the brain key theory and concept, architecture. First of all, we are introduce the Brain key theory and operation. Brain key theory is already published in the KES 2007 conference. In this paper is focused of the operation and configuration in the ubiquitous system and evaluation. And next is provide the configuration of the Brain key in the ubiquitous system.

Brain key system is needed to the new technologies. That is embedded system. Therefore, system is divide into host and target.

First of all, we are introduce to the Brain key algorithm and method. In this theory was already published in the KES 2007. Additional architecture is provide to the latter of paper.

3.2 Introduction of the Brain Key

This research by inputting the value calculated by the variable value offered from the function and password control system which were fixed in advance between the user and the password control system It is a thing about the user symptom which led the indirect password input which enables it to receive a user check. The password control system which controls a password is constituted so that it cannot embezzle, if the password entered even if it supervised the password which a user enters in more detail is led, and it is related with having enabled it to check a user through this.

When executing the various application programs of a computer generally or connecting with a specific site through the Internet A place with many examples into which carry out a user and the user-identification child (ID) and the password are made to enter in order to check a user's identity, If in agreement as compared with the identifier which a user inputs, ID which had the password registered, and a password, the authority to perform an applicable program, and the authority to use the various contents of an applicable site will be granted to the user.

However, it sets to a bank, a card issuer, etc. who the correlation of the aforementioned Prior art of a password is fixed with a user-identification child, and manage many identifiers. The identifier which is in agreement with an applicable password regardless of the count limit of an input error if many identifiers to a fixed password are changed and it continues inputting even if it restricts the password input error to a fixed identifier can be discovered. ** -- Since it has the regulation to which anyone can also learn how to enter an identifier and a password, it tries to steal, without being noticed by the user by a user's side, or a camera is installed that it seems that it is easy, Even if it is not him, appearance of the password to an identifier is investigated and carried out. if the password which looks at that a user enters an identifier and a password directly, checks the program which records the contents to input, carries out usurpation and tapping on a communication path, or is kept is stolen -- a user -- There was a problem that a user's authority could be plagiarized now[21].

3.3 Design of the Method

The password storage area 1 where Fig.1 is an example of the general password structure of a system which carried out the introductory configuration of the concept over this paper, and a password [as opposed to an identifier (ID) in the password control system of this research] is saved, The password input approach storage area 2 where the formula which made the password to the same identifier as a basic variable, and a function are saved, The reference variable storage area 3 where the variable referred to in case a user inputs is saved, and the reference-sign management equipment 4 which manages a reference sign, The output unit 5 which displays this, and the channel storage area 6 which defined the class of exercise to an identifier, The input unit 7 into which a user inputs a result, and the reaction-time storage area 8 where the reaction time to input is saved, The central processing unit 9 which carries out the process of an applicable channel which the user demanded and carries out control management after checking whether the password registered by comparing and analyzing the result inputted as the inputted reaction time and the formula are known, It consists of an extraordinary storage area 10 required at the time of execution, an output port 11 connected with the exterior, and a communication wire 12 which exchanges the signal of these mutual.

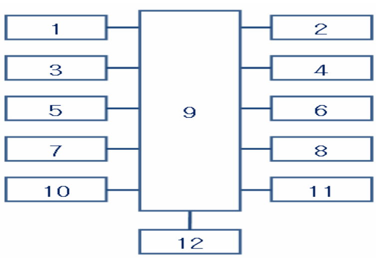


Fig. 1. Structure of a password control system

In case the phase 21 and user who input the identifier for which a user is going to use the password control system of constituted this research like the flow chart of the password control system of Fig.2 input the user-validation approach which led, thus, a password, a reference sign, a reference variable, By the approach of determining and expressing a constant, the count approach made by the channel, analysis, and sequence It judges whether the phase 23 re-calculated according to the approach of classifying an effective thing and an effective channel for the result inputted by the phase 22 of inputting the result which must be inputted, and the user, and calculating the response time by which it was fixed, and the inputted response time are within the response time which the user set. Do in the phase 25 which carries out the process of a channel which regarded it as what checked the user to an identifier when the phase 24 which analyzes with the result which the system calculated and is compared and the result suited, and was demanded together. The phase 26 which carries out the process processed when the result which the user inputted is not in agreement with the result which the system calculated and it is not a user is included.

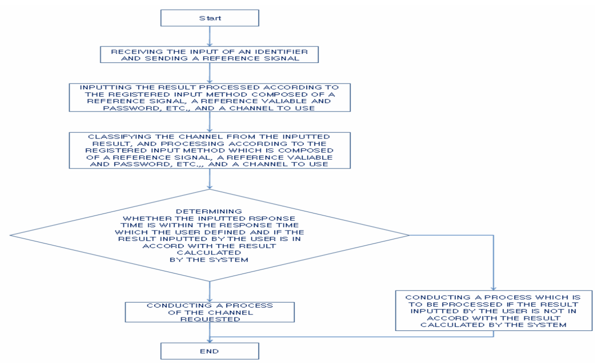


Fig. 2. Method for identification of a user

The user symptom which led the indirect password input of such this invention is as follows when the example which it may be applied to various fields using the approach of checking the user knowing the user validation to an identifier and authentication of industry, in object for **, and was applied to phone banking (Phone Banking) in it is explained in detail. Since the configuration of a phone banking system can be regarded as having added the automatic-answering system (ARS) with the configuration of the password control system of drawing 1 for the user interface, it will be explained with reference to drawing 1 as it is. In phone banking, an automatic-answering system (ARS) changes according to the class of terminal connected with a password control system in the interface part for communicating with a user and voice with the output unit 5 and input unit 7 of a password control system. Since it is unrelated to the summary of this invention, a detailed technique and a detailed notation decide to omit. The user who is going to receive phone banking service pushes the account number used after connecting with an automatic-answering system (ARS) by telephone. (Phase of receiving the input of an identifier by Fig.2)

Carrying out extraordinary preservation of the time of day which sorted out the figure of arbitration for the reference sign in the reference-sign management equipment 4 of a password control system, and sent one or the figure beyond it to the automatic-answering system (ARS) with the contents of a reference sign one by one in the extraordinary storage area 10 of delivery and a password control system at the target, an automatic-answering system (ARS) tells a reference sign with voice. (Phase of sending a reference sign by Fig.2) [21]

3.4 Architecture of the Brain Key

We are provide the concept diagram about the brain key in the ubiquitous environment. This diagram is consist of four layers that is service, application, grid infrastructure and network.

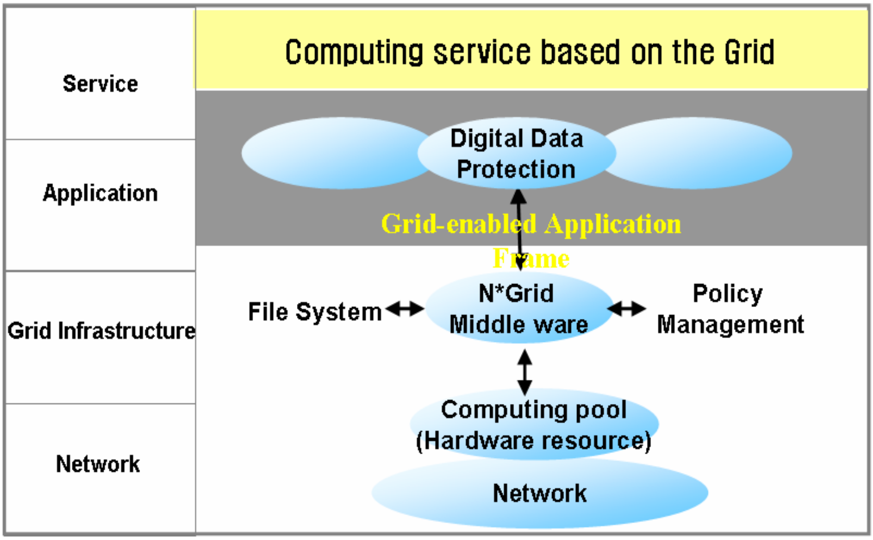


Fig. 3. Concept diagram for operation

Each of part is define like the following.

Part	Role
Service	This layer is provide services among the user needed
Application	This layer is provide flexibility between system and tools
Grid infrastructure	This layer is provide infrastructure for the usability of hardware resource
Network	This layer is provide connection between server to peers

And the detail system is like the following.

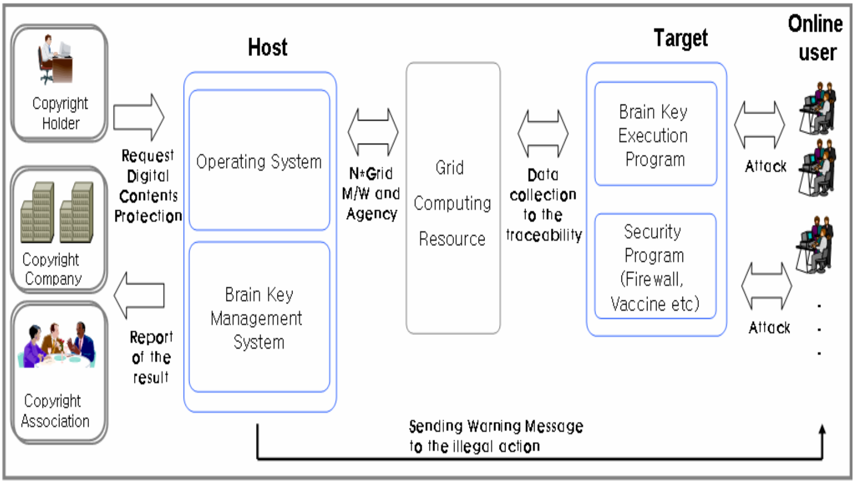


Fig. 4. Host and target system for the ubiquitous environment

Host and target system is connected by the grid computing system. Grid computing is can be use the hardware resource. Therefore, system is improving that flexibility and usability. Host and target system is split up operating of the program. That type is useful of the ubiquitous system and environment.

4 Conclusion

In this paper provide architecture of the personality security in the ubiquitous environment. This architecture is apply to the various digital system and tools. The concept of this paper is what may be applied to various fields on industry. Single devices, such as a digital door lock, therefore, of course It may be utilized as a keyword of various encryption, such as the control managerial system of a side gate, a communication link, a file, a digital signature key, a smart card, and a card reader machine, and a decryption. The various user validations and electronic financial dealings which are utilized in the Internet electronic commerce, Small wireless radios, both directions TV, etc., such as a credit card enquiry machine, an ATM terminal and a cellular phone, and PDA. The place which is a thing broadly applicable to the field which exchanges a mutual signal, if this is a person with usual knowledge, it can constitute the password control system suitable for it within limits which are not missed to the summary of this invention easily.

Therefore, various systems and types is require the architecture for the flexibility and security and usability of the support.

For the further study is evaluate the architecture of the personal security. And we are supplements architecture after applying of the real world. Also, we are analysis of the architecture efficiency against security of the defense and flexibility of rate. Therefore, we are put to use the reliable method of the evaluation.

References

[1] Pressman, R.S.: Software Engineering. Mcgraw-hill international edition, New York (1997)

[2] Luger, G.F.: Artificial intelligence. Addison-Wesley, Reading (2001)

[3] Beigl, M., Gellersen, H.W., Schmidt, A.: MediaCups: Experience with Design and Use of Computer-Augmented Everyday Object. Computer Networks (May 2001)

[4] Ponnekanti, S.R., et al.: ICrafter: A Service Framework for Ubiquitous Computing Environment. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) UbiComp 2001. LNCS, vol. 2201, Springer, Heidelberg (2001)

[5] Feeney, L., Ahlgren, B., Westerlund, A.: Spontaneous Networking: An Application oriented Approach to Ad Hoc Networking. IEEE Comm. Magazine (June 2001)

[6] Lee, E.-S., Hwang, S.-M.: Definition of Security Requirement Items and Its Process to Security and Progress Management, LNCIS, vol. 344 (August 2006)

[7] Lee, E.-S., Hwang, S.-M.: Design Implementation of Web Security Access Control System for Semantic Web Ontoloty. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3481, Springer, Heidelberg (2005)

[8] Lee, E.-S., Lee, M.: Development System Security Process of ISO/IEC TR 15504 and Security Considerations for Software Process Improvement, LNCIS, vol. 344 (August 2006)

[9] Lee, E.-S., Lee, S.-H.: Design progress management for Security Requirements in Ubiqiiuous computing using COQUALMO. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, Springer, Heidelberg (2006)

[10] Cho, N.-d., Lee, E.-S.: Design and Implementation of Semantic Web Search System using Ontology and Anchor Text. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, Springer, Heidelberg (2006)

[11] ISO/IEC 15408-1:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model

[12] ISO. ISO/IEC 15408-2:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 2: Security functional requirements

[13] ISO. ISO/IEC 15408-3:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 3: Security assurance requirements

[14] Saba, D., Mukberjee, A.: Pervasive computing a paradigm for the 21st century. IEEE Computer Society Press, Los Alamitos (2003)

[15] Weiser, M.: Some comuter science issues in ubiquitous comuting. IEEE Computer Society Press, Los Alamitos (1993)

[16] Kindgerg, T.: System software for ubiquitous computing. IEEE Computer (2002)

[17] Foster, I., Kesselman, C.: The Grid: blueprint for a new computing infrastructure. Morgan Kaufmann Publishers Inc., San Francisco (1998)

[18] Abbas, A.: Grid Computing: Practical Guide To Technology & Applications. Charles River Media (2003)

[19] Foster, I., Kesselman, C., Tuecke, C.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications 15(3), 200–222 (2001)

[20] Foster, I.: What is the Grid? A Three Point Checklist. Grid Today 1(6) (July 22, 2002), http://www.gridtoday.com/02/0722/100136.html

[21] Lee, E.-S., Kim, H.-K., Lee, S.H.: Design of the Intelligence User Identification Method for the Improved Password Input Method and the Personality Security. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 581–588. Springer, Heidelberg (2007)

The Multi Agent System Solutions for Wireless Sensor Network Applications

Khin Haymar Saw Hla, YoungSik Choi, and Jong Sou Park

Department of Computer Engineering, Korea Aerospace University, Korea
{haymar,choimail,jspark}@kau.ac.kr

Abstract. This research proposes to employ multi-agent system solutions to the applications concerning with the wireless sensor network environment. It realizes the multi-agent community by combined use of intelligent sensors and middle agent architecture. Intelligent sensor nodes exploit as autonomous agents which monitor the events at the environment. Directory facilitator or middle agent is used as a yellow pages directory for the agent look ups and treat the ad hoc sensor network as an agent community. For retrieving the data from the target nodes, mobile software agent is used for energy saving manner. Simulation results have approved that using multi-agent system solution is highly suitable to access and distribute sensory data efficiently and timely manner.

1 Introduction

Wireless sensor networks have emerged as a new monitoring and control solution for a variety of applications. Due to the small dimensions, sensor nodes have strong hardware and software restrictions in terms of processing power, memory capability, power supply and communication throughput. The energy is the most critical restriction, given that it is typically not rechargeable. Many applications are using this technology in many areas such as habitual monitoring, environmental monitoring, home monitoring and military explorations, etc. This research focuses on the event monitoring applications in which, the system needs to detect dynamic environment and adapt to the situation and present information that is relevant to it. Due to the energy restriction, sensor information should not be broadcast over all network nodes. Transmitting in a sensor network should be performed to nodes that can be of benefit by that specific information. Since decisions should be made by the network reactively regarding to the changes of an environment, employing intelligent system approaches are preferable. Intelligent sensors can interact with their environment reactively, and therefore they adapt well to such continuous changes. When system presumes to response the user's requests, results information should be collected within the areas which have the active events. In order to reduce the energy consumption in data collection, reduction in redundancy of transmitted data is also important. In the wireless sensor network, the source nodes are closed to one another at the environment and generate a lot of sensory data traffic with redundancy. Therefore, only the aggregated data should be used to reduce the energy consumption in the transmissions. There is a need for

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 454–463, 2008.

mechanism for advertising, finding, fusing the information of the active areas. To address these issues, middle agents are proposed. Middle agents are entities that assist in locating the ultimate information providers with the ultimate requesters in the dynamic environment. Many different types of middle agents are useful in large, distributed, open multi-agent problem solving systems. These include matchmaker or yellow page agents that process advertisements, blackboard agents that collect requests and brokers that process both. The behaviors of each type of middle-agent have certain performance characteristics such as privacy, robustness, and adaptive-ness qualities that are related to characteristics of the external environment and to the agents themselves [3].

This paper proposes to employ multi-agent system solutions to the event monitoring applications concerning with the wireless sensor network environment. It realizes the multi-agent community by combined use of intelligent sensors and middle agent architecture. This agent community consists of intelligent sensor nodes, which exploit as autonomous agents, data centric sensor nodes, middle agent or directory facilitator and mobile agents. For detecting dynamic environment changes, the autonomous agents are able to perceive their environment and respond in a timely manner to the changes that occurred in it. The middle agent enables an agent to find another agent by functionality, capability and availability, without knowing who or where the provider agent might be. For retrieving the data from the target nodes, mobile software agents are used for energy saving manner. Mobile agents enable to migrate one host to another to gather information and perform task and reducing network bandwidth of messages required to process data.

The rest of the paper is organized as follows: in section 2, we discuss the existing work on agent oriented wireless sensor networks. In section 3 we define the definitions and terms that are used for. In section 4, we present the proposed architecture and agent role behaviors for the system. In section 5 we discuss the performance tradeoffs over the current system and existing systems. Finally, in section 6, we conclude and lay out some future work.

2 Related Work

Many researches have proposed to use the adapting agents to work on distributed and complex sensor networks. To address the problems of heterogeneity, current research endeavors have shown the utility of multi-agent systems approaches on the wireless sensor networks. Tynan et al. [9] proposed multi-agent systems architecture to interconnect the wide range of heterogeneous devices that may posses various levels of resources. Karlsson et al. [6] proposed the agent oriented programming paradigm for development of intelligent sensor networks. The case study was performed by implementing test bed using JADE. The application of the specific implementing is an Unattended Ground Sensor Network (UGSN) for surveillance of moving targets. They proposed to use autonomous sensor nodes, which exchange information, reason and collaborate with each other and formed fully deployed intelligent sensor network. All these research works have shown that the agent oriented software paradigm has approved to be highly suitable in the development of complex and distributed sensor network systems.

Data integration is one of the issues in the distributed wireless sensor networks. Mobile agent approaches are increasingly being applied in data integration. In the mobile agent model, data stay at the local site, while the processing task is moved to the data sites [10]. By transmitting computation engine instead of data, network bandwidth requirement is reduced [1]. Qi. et al. [7] proposed to use the mobile agent based distributed sensor network (MADSN) for the distributed sensor integration tasks. They deployed the mobile agent paradigm to be aware of reactively to the continuously changing network conditions to guarantee successful performance of the application tasks. MADSN is based on the assumption that the operation of mobile agent is only carried out within one hop in a clustering-based architecture. Chen et al. [2] proposed mobile agent based directed diffusion (MADD) to distributed sensor network for scalable and energy efficient data aggregation. They considered mobile agents (MA) in multi-hop environments and adopted directed diffusion (DD) to dispatch MA. In directed diffusion, a sensing task is disseminated throughout the sensor network as an interest queries for named data, that is user interests are diffused through the sensor network. Sink node floods a query towards the sensors of interest and intermediate nodes set up gradients to send data along the routes toward the sink node [4]. However, the current mobile agent based directed diffusion framework is designed to retrieve the data directly from the sensor network whenever there are requests from the users. For retrieving requests only from the active area, in that case, some enhancement for the framework is needed.

3 Definitions and Terms

In this section, we first define the definitions and terms that are used for. In particular, this system relies on two types of sensor nodes, intelligent sensors and data centric sensors and five types of agents such as autonomous agent, mobile agent, directory facilitator agent, requester agent and provider agent. Knowledge about requester agent preferences and provider agent capabilities are also examined.

3.1 Definitions

Definition 1: An intelligent sensor is defined as a sensor that monitor to the environment and when it detects an event in the environment, it will broadcast to the base station.

Definition 2: A data centric sensor is defined as a sensor, which senses the environment and collect the sensed data locally.

Definition 3: An autonomous agent is a system situated within a part of an environment that senses that environment and acts on it overtime, in pursuit of its own agenda and so as to effect what it senses in the future.

Definition 4: A mobile agent is a special kind of software which can execute autonomously. Once dispatched, migrates from node to node performing data processing autonomously, while software can typically only execute when being called upon by other routines.

Definition 5: Directory facilitator is an entity to which other agents advertise their capabilities and it can coordinate agent's activities and can satisfy requests on behalf of their subordinated agents.

Definition 6: Requester agent is an entity that requests the information from the system on behalf of the user's needs.

Definition 7: Provider agent is a kind of agent that provides the information within an active environment through their superior agents.

3.2 Terms Used

Two types of information are mainly used, preferences and capabilities. A preference is knowledge about what types of information have utility for a requester and capability is knowledge about what types of requests can be served by a provider [3]. A specific request is an instance of an agent's preferences, and specific reply in service of request is an instance of an agent's capabilities. An agent can have a mental state with respect to a particular specification of a preference or capabilities. An announcement is a capability specification that is committed to servicing for the active event.

4 The System Overview

The study focuses on the heterogeneous wireless sensor network environment as depicted in figure1. It consists of some intelligent sensor nodes, which are used for the event monitoring and multiple data centric sensor nodes, which are used to store data locally. Directory facilitator or middle agent is used to act as a yellow pages directory for the agent look ups and treat the ad hoc sensor network as an agent community. The intelligent sensors act as autonomous agents, which are able to perceive their environment and respond to the changes autonomously.

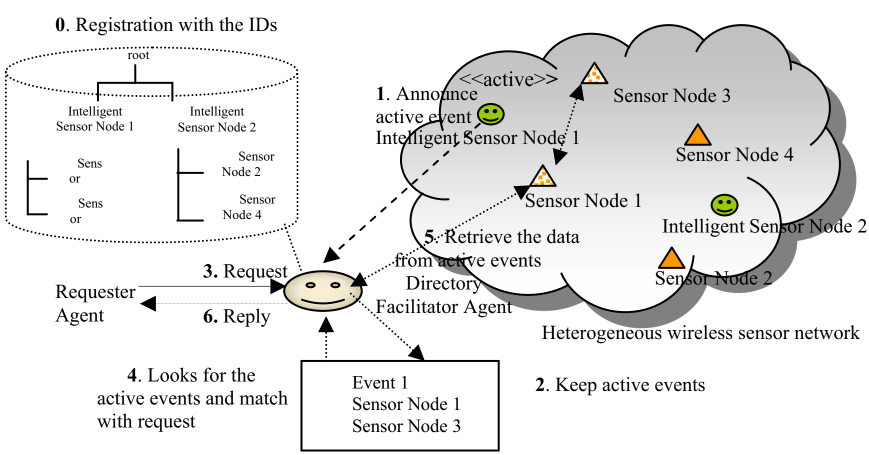


Fig. 1. Overview of the proposed system

When an event occurs in the environment, the intelligent sensors announce to the directory facilitator. Directory facilitator keeps the identities of the intelligent sensors and their associated data centric sensors in its database. When directory facilitator receives the announcements from the intelligent sensors, it keeps the announcements as active services. When requester requests the data, directory facilitator looks for the matched active events from its database and retrieves the identities of data centric nodes which are associated to intelligent sensor and performs the data collection task by using mobile agents. Then directory facilitator keeps these data in the database. When there are same requests from the users, directory facilitator responses the request by retrieving the data from the database.

4.1 Agent Role Behaviors

This system has three basic agent roles: that of the requester, the provider that is intelligent sensor nodes and their associated data centric sensor nodes of the wireless sensor network and the directory facilitator, which is middle agent of the system.

4.1.1 Role of Requester

Requesters are the users of the applications. Let $R=\{r_1,r_2,\ldots,r_n\}$ be a set of users and let $Q=\{q_1,q_2,\ldots q_n\}$ be a set of queries that requested by the user r_i, where $[i\in 1,n]$. There exists some queries q_x, q_y, where $q_x=q_y$ or $q_x \neq q_y$, that is, there can be same or different queries from the different users.

$$\forall R \exists q_x, q_y \in Q \text{ which is either } q_x=q_y \text{ or } q_x \neq q_y$$

4.1.2 Role of Provider

In this system, providers are intelligent sensor nodes and their related data centric sensor nodes. First providers must register to the directory facilitator for their existence. They announce for the active events, which is committed to the service capabilities.

Registration

First, intelligent sensor nodes and their associated data centric nodes must register to the directory facilitator (DF). Let $D=\{d_1,d_2,\ldots,d_n\}$be a set of data centric sensor nodes and $G=\{g_1,g_2,\ldots,g_n\}$ be a set of intelligent sensor nodes in the sensor network environment. Each data centric sensor node is associated with a unique identifier (ID). A node G is said to contain D if and only if $D \subseteq G$.

Announcement

When an event occurs at the environment, intelligent sensor nodes detect the movements and announce to the DF. Let $S=\{s_1,s_2,\ldots,s_n\}$ be the set of environment states of the wireless sensor network environment denoted by ψ.

The announcement is realized by intelligent sensor nodes g_i, at ψ where $[i \in 1,n]$, through a set of behaviors. A behavior is a pair (c,a) where $c\subseteq P$ is a set of percepts called condition and $a\in A$ is an announcement. A behavior(c,a) will fire when the environment state $s\in S$ at ψ.

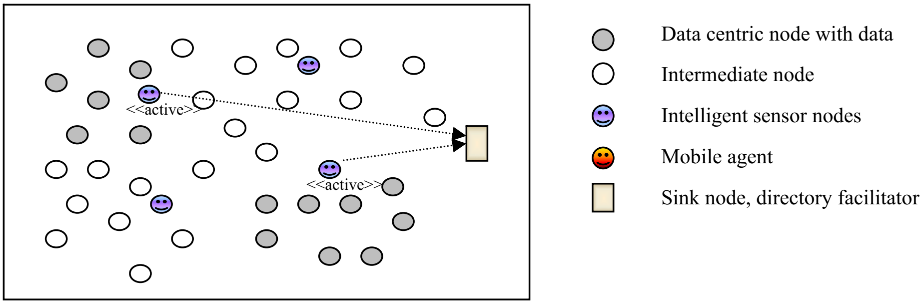


Fig. 2. When an event occurs at the environment, intelligent sensor nodes announce to the directory facilitator

4.1.3 Role of Directory Facilitator

Role of the directory facilitator (DF) is to keep the provider's identities (IDs) and their capability announcements at its local database. When requester asks for a preference, DF matches with a provider's capabilities and responses with a set of matched events to the requester.

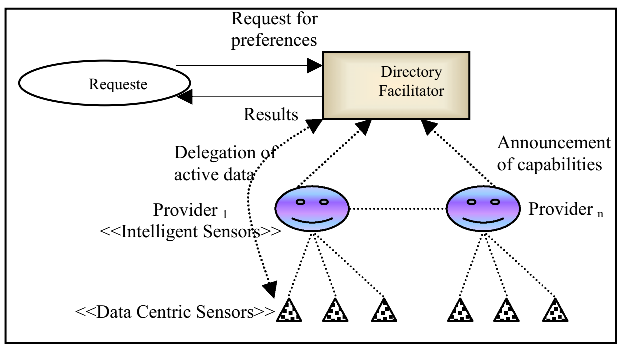


Fig. 3. Role of directory facilitator

Delegation of active data

DF keeps the IDs of the intelligent sensor nodes and their associated data centric sensor nodes value pairs at its database. When the DF receives the announcements from the providers, it searches the IDs of the data centric sensor nodes which are associated with the IDs of the intelligent sensor nodes that are active. For every pair of IDs (g_k,d_i), where $[i,k \in 1,n]$, $d_i \in D$ $g_k \in G$ and $D \subseteq G$ the DF retrieves the associated IDs that meets the requester's IDs.

Matching preferences

When requester submits a request to the DF, it looks for the database and matches against for the request according to the active events. Events with the similar capabilities are clustered into the same community. Request from several providers may need to some complex integration and then serves the integrated results to the requesters.

Then DF finds out the mobile agent (MA) threads to get the sensed data from the active nodes. Multiple MAs are used to collect the sensed data.

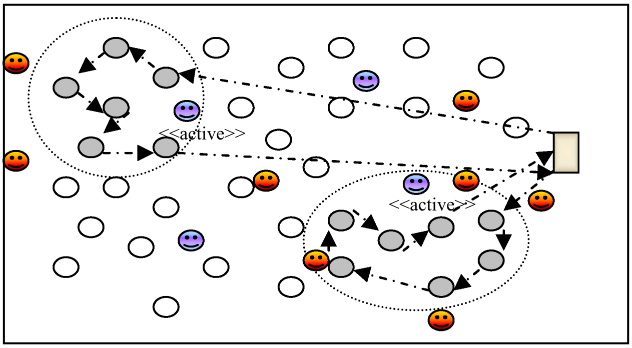


Fig. 4. Retrieve the data from the target nodes associated with the intelligent nodes by MAs

Data integration

The degree of sensed data reduction among sensors is related to the distance between sensors. It is very likely for closely located sensors to generate redundant sensed data. Therefore, the data aggregation is necessary in order to reduce transmission. The data aggregation task is performed by the MA threads. MAs collect the data hop by hop within the active range and aggregate these data according to the specified threshold. For data aggregation, we calculate the size of data result accumulated by the MA using similar method in MADD [2]. A sequence of data result can be fused with the aggregation factor p where $(0 \leq p \leq 1)$. Value of p depends on the type of applications. Let r, where $[r \in 0,1]$ be the reduction ratio by the MA assisted local processing. Let $s^i{}_{data}$ be the size of raw data at source i, and let R_i be the size of reduced data. Then R_i is represented as,

$$R_i = s^i{}_{data}(1-r) \tag{1}$$

Let $s^i{}_{ma}$ be the amount of accumulated data result after the MA leaves source i, where R_i is the amount of data that will be aggregated by p. Then $s^i{}_{ma}$ is represented as,

$$s^i{}_{ma} = s^{i-1}{}_{ma} + (1-p)R_i \tag{2}$$

5 Performance Tradeoffs

5.1 Performance Analysis

This section presents an analysis due to the energy consumptions for the active events, which are matched with the user's request in WSN environment. We account for the energy consumption of MA packet migration for the active nodes for these events.

Let E denote the energy consumption for MA migration for an event. Let E_f be the energy consumption of MA migrating from the sink to the data centric sensor node, E_t be the average energy consumption of MA traverse through the active data centric nodes and E_b be the average energy consumption of MA migrating from the data centric nodes to sink node for an event. Let s_{data} be the size of sensed data including header packet. Let s_p be the size of processing code. Then E_t is represented as

$$E_t = \sum_{i=1}^{N} (s_{data} .m_p + s^i_{ma} + s_p).(m_t + m_r) + c \tag{3}$$

Where N is the number of data centric nodes within an active environment, m_p is the energy consumption for processing bit; m_t and m_r are the energy consumption for receiving and transmitting a bit respectively. c is used for the fix energy cost to transmit a packet.

Let s^N_{ma} be the size of an MA packet including header packet after the MA visits the active data centric sources and H is the number of hops along the path between active data centric nodes. Then E_b can be represented as

$$E_b = (s^N{}_{ma}.(m_t + m_r) + c).H \tag{4}$$

Then we can calculate E as

$$E = (E_f + E_t + E_b) \tag{5}$$

Let $n_{matchevent}$ be the number of active and matched events occurs in WSN environment and E_{total} be the total energy consumption for these events. Finally, E_{total} can be calculated as

$$E_{total} = \sum_{i=1}^{n_{matchevent}} (E_i) \tag{6}$$

5.2 Simulation Results

We perform simulation based on the energy consumption by using the parameter values on the table 1 and 2. A sensor network with 110 sensor nodes is considered. Run with 10 intelligent sensor nodes and 100 data centric sensor nodes, in which each intelligent sensor node has 10 associated data centric sensor nodes. When there is a request from the user, the DF retrieves the data only from the active area which matches the requester's request. Therefore, trade offs between queries and matched events are considered during simulations as depicted in figure 5. Since it doesn't need to look for the interested data from the whole sensor network whenever there is a query, the energy consumption E for the overall request is considerably reduced. The simulation result shows that the maximum energy consumptions are not more than the total energy consumptions for the number of events. When the number of user queries is more than matched events within WSN environment, the energy consumptions for the overall queries can be considerably reduced.

Table 1. Energy Consumption Parameters

Normalized initial energy of sensor node (W·sec) 4500		
Incremental cost (μW·s/bytes)	m_t	1.9
	m_r	0.5
	m_p	0.39
Fixed cost (μW·s)	c	843

Table 2. Basic Specification

Network size	500m× 500m
Topology configuration mode	Randomized
Intelligent Sensor node	10
Data centric Sensor node	100
Number of source nodes (N)	Default:10
Size of sensed data (s_{data})	Default: 1KB
Size of processing code (s_p)	Default: 2MB
Raw data reduction ratio (r)	Default: 0.8
Aggregation ratio (ρ)	Default: 0.2
Number of Queries	20
Number of events	10

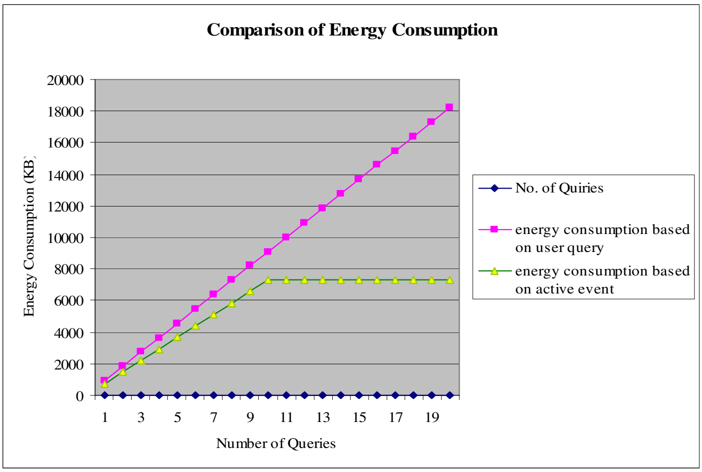


Fig. 5. Comparison of energy consumption over query based and middle agent based approaches

If there are the same queries from the different users, this approach can improve search speed and search efficiency by keeping the capabilities of active events at DF's database. Simulation results have approved that using multi-agent system solution is highly suitable to access and distribute sensory data efficiently and timely manner.

6 Conclusions and Future Work

This approach allows building the wireless sensor network environment as an agent community that can work in distributed and heterogeneous WSN. It describes the construction and effective used of middle agents in the WSN environment, whilst considering user preferences and agent capabilities. Major advantages are that we can considerably reduce the energy consumption by using intelligent sensor nodes, and improve search speed and search efficiency when there is large number of users or queries from the distributed and ad hoc WSN as the middle agent community is correctly formed. In our future work, we plan to study how self-organized community of middle agent effects for distributed and heterogeneous WSN environment.

Acknowledgements

This research was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-210-D00005).

References

[1] Chen, M., Kwon, T., Choi, Y.: Data Dissemination based on Mobile Agent in Wireless Sensor Neteworks. In: Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary (2006)
[2] Chen, M., Kwon, T., Choi, Y., Leung, V.C.M.: Mobile Agent-Based Directed Diffusion in Wireless Sensor Network. EURASIP Journal on Advances in Signal Processing (2007)
[3] Decker, K., Sycara, K., Williamson, M.: Middle Agents for the Internet. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence (1997)
[4] Intanagonwiwat, C., Govindan, R., Estrin, D., Heidemann, J., Silva, F.: Directed Diffusion for Wireless Sensor Networking. IEEE/ACM Transactions on Networking 11(1) (2003)
[5] Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks. Technical Report 00-732
[6] Karlsson, B., Backstorm, O., Kulerza, W., Axelsson, L.: Intelligent Sensor Networks- an Agent –Oriented Approach. In: Proceedings of the Workshop on Real World Wireless Sensor Networks (2005)
[7] Qi, H., Xu, Y., Wang, X.: Mobile-agent Based Collaborative Signal and Information Processing in Sensor Networks. Proceedings of the IEEE 91(8), 1172–1183 (2003)
[8] Tseng, Y.-C., Kuo, S.-P., Lee, H.-W., Huang, C.-F.: Location Tracking in Wireless Sensor Network by Mobile Agents and its Data Fusion Strategies. Computer Journal 47(4), 448–460 (2004)
[9] Tynan, R., Ruzzelli, A.G.: A Methodology for Deployment of Multi-Agent systems on Wireless Sensor Networks. In: Proceedings of the 17th International Conference on Software Engineering and Knowledge Engineering (2005)
[10] Aglets, http://www.trl.ibm.co.jp/aglets

A Study of Implanted and Wearable Body Sensor Networks

Sana Ullah[1], Henry Higgin[2], M. Arif Siddiqui[1], and Kyung Sup Kwak[1]

[1] Graduate School of Telecommunication Engineering
253 Yonghyun-Dong, Nam-Gu, 402-751, Inha University Incheon South Korea
sanajcs@hotmail.com, arif.sid@hotmail.com, kskwak@inha.ac.kr
[2] Zarlink Semiconductor Company
Caldicot NP26 5YW United Kingdom
henry.higgins@zarlink.com

Abstract. Recent advances in intelligent sensors, microelectronics and integrated circuit, system-on-chip design and low power wireless communication introduced the development of miniaturised and autonomous sensor nodes. These tiny sensor nodes can be deployed to develop a proactive Body Sensor Network (BSN). The rapid advancement in ultra low-power RF (radio frequency) technology enables invasive and non-invasive devices to communicate with a remote station. This communication revolutionizes healthcare system by enabling long term health monitoring of a patient and providing real time feedback to the medical experts. In this paper, we present In-body and On-body communication networks with a special focus on the methodologies of wireless communication between implanted medical devices with external monitoring equipment and recent technological growth in both areas. We also discuss about open issues and challenges in a BSN.

Keywords: Body Sensor Network, Wireless Body Area Network, Implanted Communication, Wearable Computing.

1 Introduction

The leading cause of death in US is heart disease, i.e. about 652,486 and 150,074 people died due to cardiovascular and cerebrovascular diseases [1]. In South Korea, 17% people die due to cerebrovascular diseases [2]. The health care expenditure in US is expected to reach 2.9 trillion by 2009 and 4 trillion by 2015, or 20% of Gross Domestic Product (GDP) [3]. Cardiovascular disease is the leading cause of death and it accounts for approximately 30% of all deaths worldwide [4]. In UK, it is 39% of all deaths [5]. In Europe, 90% of people die due to arrhythmogenic event [6]. Irregular heart beat causes such deaths and can be monitored before heart attack. Holter monitors are used to collect cardio rhythm disturbances for offline processing without real time feedback. Transient abnormalities are sometimes hard to capture. For instance, many cardiac diseases are associated with episodic rather than continuous abnormalities such as transient surges in blood pressure, paroxysmal arrhythmias or

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 464–473, 2008.

induced episodes of myocardial ischemia and their time cannot be predicted [6]. The accurate prediction of these episodes improves the quality of life.

Body Sensor Network (BSN) is a key technology to prevent the occurrence of myocardial infarction, monitoring episodic events or any other abnormal condition and can be used for long term monitoring of patients. The term BSN is first coined by Prof Guang-Zhong Yang of imperial college London. A BSN consists of miniaturised, low power and noninvasive or invasive wireless biosensors, which are seamlessly placed on or implanted in human body in order to provide an adaptable and smart healthcare system. This seamless integration of small and intelligent wireless sensors is used to monitor the patient's vital signs, provide real time feedback and can be a part of diagnostic procedure, maintenance of chronic condition, supervised recovery from a surgical procedure and to monitor effect of drugs therapy [7]. Each tiny biosensor is capable of processing its own task and communicates with a network coordinator or PDA. The network coordinator sends patient information to a remote server for further analysis. Episodic examination of a patient captures a snapshot of recovery process and skips other potential complications [8]. A BSN focuses on early detection of life threatening abnormalities and maintenance of chronic condition [9]. Long term monitoring of patient activities under natural physiological states improves quality of life by allowing patients to engage in normal daily life activities, rather than staying at home or hospital [10]. Moreover, implants for therapeutic and diagnostic purposes are also becoming more common. They can be used to restore control over paralyzed limbs, enable bladder and bowel muscle control, maintain regular heart rhythm, and many other functions. These implants significantly improve the quality of life of many patients. Though BSN research is in inception, but a number of on going research has enabled the innovation of several prototypes for unobtrusive pervasive healthcare system. However, no standard exists for a BSN due to considerable number of issues and challenges such as interoperability, privacy and security, low power communication, biosensor design, baseline power consumption, communication link between the implanted device and external monitoring and control equipment. The scope of a BSN spans around three domains: Off-body communication, On-body communication and In-body communication. Off-body communication is the communication from the base station to the transceiver on human side. On-body communication is the communication within on-body networks and wearable system. In-body communication is the communication between invasive or implantable devices with a base station.

The rest of the paper is divided into four categories. Section 2 contains a detailed discussion on invasive or in-body communication with a special focus on the methodologies of wireless communication between implanted medical devices with external monitoring equipment. This includes discussion on inductive coupling, in-body RF communication and antenna design. Section 3 presents a brief discussion on non-invasive or on-body communication and the recent technological growth in this area. Section 4 contains discussion on multi-agent technology for a BSN. Section 5 focuses on open issues and challenges in a BSN. Finally, we present conclusion to our work.

2 In-Body Communication

Advancement in implant technology and RF communication has enabled the communication of invasive or implanted device with a remote base station and can monitor every aspect of a patient. These new implant technologies require a communication link between the in-body device and external monitoring and control equipment. Zarlink semiconductor has introduced the world's first wireless chip, which supports a very high data rate RF link for communication with an implantable device [11]. The ZL70101 ultra-low power transceiver chip satisfies the power and size requirements for implanted communication systems and operates in 402-405 MHz Medical Implantable Communications Service (MICS) band [12].

There are several ways to communicate with a human body implant, including methods that use electromagnetic induction (similar to radio frequency identification, or RFID) and the others that use radio frequency (RF). Both are wireless and their use will depend on applications.

2.1 Inductive Coupling

Several applications still use electromagnetic coupling to provide a communication link to implanted devices, with an external coil held very close to the patient that couples to a coil implanted just below the skin surface. The implant is powered by the coupled magnetic field and requires no battery for communication. As well as providing power, this alternating field is also be used to transfer data into the implant. Data is transferred from the implanted device by altering the impedance of the implanted loop that is detected by the external coil and electronics. This type of communication is commonly used to identify animals that have been injected with an electronic tag. Electromagnetic induction is used when continuous, long-term communication is required.

The base band for electromagnetic communication is typically 13.56 MHz or 28 MHz, with other frequencies also available. Its use is subject to regulation for maximum Specific Absorption Rate (SAR). Inductive coupling achieves the best power transfer when using large transmit and receive coils, meaning it's impractical when space is an issue or devices are implanted deep within the patient. This technique does not support a very high data rate and cannot initiate a communication session from inside of the body. Please read Finkenzeller et.al [13] for further details.

2.2 In-Body RF-Communication

Compared with inductive coupling, RF communication dramatically increases bandwidth and enables a two-way data link to be established. The Medical Implant Communication Service (MICS) band of 403 MHz to 405 MHz is gaining worldwide acceptance for in-body use [14]. This band has a power limit of 25 μW in air and is split into 300 kHz wide channels. The human body is a medium that poses numerous wireless transmission challenges. The body is composed of varied components that are not predictable and will change as the patient ages, gains or loses weight, or even changes posture. There are formulas for designing free-air communications but it's very difficult to calculate performance for an in-body communication system. To compound the design challenge, the location of the implanted device is also variable. During surgery the implant is placed in the best position to perform its primary function, with little consideration for wireless performance.

Typical dielectric constant (ε_r), conductivity (ρ) and characteristic impedance $Z_0(\Omega)$ properties of muscle and fat are shown in Table 1.

Table 1. Body Electrical Properties

Frequency	Muscle			Fat		
	(ε_r)	ρ (S.m^{-1})	$Z_0(\Omega)$	(ε_r)	ρ	$Z_0(\Omega)$
100	66.2	0.73	31.6	12.7	0.07	92.4
400	58	0.82	43.7	11.6	0.08	108
900	56	0.97	48.2	11.3	0.11	111

The dielectric constant has an effect on the wavelength of a signal. In air the wavelength can be found from Equation 1 where $\varepsilon_r = 1$. However in a different medium the wavelength is reduced as in Equation 2.

$$\lambda = 300\frac{10^6}{f} \tag{1}$$

where λ is the wavelength in air in meters and f is frequency in Hz.

$$\lambda_{medium} = \frac{\lambda}{\sqrt{\varepsilon_r}} \tag{2}$$

where λ_{medium} is the wavelength in medium.

At 403 MHz the wavelength in air is 744 mm, but in muscle with $\varepsilon_r = 50$ the $\lambda_{medium} = 105$ mm. This is of considerable help in designing implanted antennas where physical size is an important consideration. The conductivity of muscle is 0.82Sm^{-1} – this is more than air, which is almost zero. The effect of this is similar to surrounding the implant with seawater that will attenuate the signal as it passes through. This results in reduced penetration. The characteristic impedance (Z_o) is relevant when it changes, such as at the fat-muscle boundary. This will cause part of the signal to be reflected by a term known as reflection co-coefficient Γ, found from Equation 3.

$$\Gamma = \frac{Z_o - Z_r}{Z_o + Z_r} \tag{3}$$

where Z_o is the impedance of free space (377 Ω), and Z_r is the impedance of medium in Ω. This results in a signal being reflected of magnitude Γ of incident signal power. So for muscle-fat boundary Γ = 80% of the signal is reflected. As an implant does not have an earth (ground), the case or other wires will also radiate. This means that signals will be radiated from the antenna and other structures associated with the implant. More details are available in Yang et.al [15].

2.3 Antenna Design

An in-body antenna needs to tuneable, using an intelligent transceiver and routine. This will enable the antenna coupling circuit to be optimised and the best signal strength obtained. Often size constraints dictate the choice of a non-resonant antenna. A non-resonant antenna will have lower gain and therefore be less sensitive on receiving and radiate less of the power generated by the transmitter. This makes design of the antenna coupling circuit even more important.

A patch antenna can be used when the implant is flat and there is no room to deploy a short wire. Patch antennas comprise a flat substrate coated on both sides with conductor. The substrate is typically alumina or a similar body-compatible material, with platinum or platinum/iridium coating both surfaces. The upper surface is the active face and is connected to the transceiver. The connection to the transceiver needs to pass through the case where the hermetic seal is maintained, requiring a feed-through. The feed-through must have no filter capacitors present; these are common on other devices. A patch antenna will be electrically larger than its physical size because it will be immersed in a high ε_r medium. It can be made to appear even larger electrically if the substrate is of high ε_r. The off-resonance antennas have low radiation resistance, typically in the order of a few Ohms for a patch. A loop antenna is an option where it can be deployed attached to the implant case. The loop antenna operates mostly with the magnetic field, whereas the patch operates mostly with the electric field. The loop antenna delivers comparable performance to that of a dipole, but with a considerably smaller size. Also the magnetic permeability of muscle or fat is very similar to that of air, unlike the dielectric constant that varies considerably. This property enables an antenna to be built and used with much less need for retuning. A loop antenna can be mounted on the case in a biocompatible structure. Equations 4 and 5 relate to small and large loops, other equations exist for multi-turn loop designs.

$$R_{rad} = 31200\left(A/\lambda^2\right)^2 \qquad A \le \lambda^2/100 \tag{4}$$

where R_{rad} is radiation resistance and A is the loop area and λ the wavelength in medium.

$$R_{rad} = 3270\left(A/\lambda^2\right)^2 \qquad A > \lambda^2/100 \tag{5}$$

More details of antenna design can be found from Kraus [16] Fujimoto [17], Lee [18], and Krall [19]. The performance of an implanted communication system within a body is difficult to predict or simulate. Approximation to a human body can be made with a body phantom liquid as described in the book edited by Yang [15]. Unlike applications in air, there are no reliable equations to use and therefore only limited simulation models. That said, simulation can provide a guide to the propagation from a body but should not be used to guarantee performance.

3 On-Body Communication

The rapid growth in intelligent sensors, microelectronics and integrated circuit, system-on-chip design, and low power wireless communication has introduced the development of miniaturized and non-invasive sensor nodes. These non-invasive sensor nodes can be placed on human body to create an on-body communication network, which can be used for ambulatory health monitoring of a patient. Unlike in-body communication where the devices are implanted in human body, in on-body communication network, the tiny sensors are placed on the body with out implantation, which provides long term health monitoring and prevents the occurrence of life threatening events. The information is gathered into a central intelligent node or PDA, which also provides an interface to the patient as well as communicates with a remote server. A BSN usually consists of three levels [20]. The first level is called sensor level, which consists of miniaturised low power sensors such as ECG (electrocardiogram), SpO2 (oxygen saturation sensor), EMG (electromyography) and EEG (electroencephalography). The second level called PDA or central intelligent node collects patient information and communicates with the remote station. The third level consists of a remote base station, which keeps patient medical records and provides diagnostic recommendations [20].The GPRS system is used to track the patient's location. A number of on-going projects such as CodeBlue [21], MobiHealth [22] and Connect [23] have facilitated research in on-body communication networks. A system architecture of wireless body area network is presented in [20], where existing Telos platform with an integrated wireless ZigBee compliant radio with on-board antenna is modified by adding ISMP (Intelligent signal processing module) component. This architecture performs real time analysis of sensors data, provides feedback to the user and forwards the user's information to a telemedicine server. A project called Ubiquitous Monitoring Environment for Wearable and Implantable Sensors (UbiMon) aims to develop a smart and affordable health care system and is designed by using six components: the sensors, the remote sensing units, the local processing units, the central server, the patient database, and the workstation [24]. A BSN node for on-body network is developed during this project. The BSN node provides a versatile environment for pervasive healthcare applications and requires 0.01mA in active mode.The BSN node uses IEEE 802.15.4 (Zigbee) wireless link as a low power communication protocol. However, the narrowband implementation doesn't satisfy the energy consumption budget of the sensor nodes and hence, an alternative solution is required. The emerging UWB technology is considered to be the best alternative solution, which could reduce the baseline power consumption of sensor nodes. A pulse-based UWB scheme for on-body communication networks [25],UWB channel measurement with antennas placed on human body [26] and UWB antennas for a BSN [27] have urged researchers to consider UWB technology for communication within on-body networks.

4 Multi-agent Technology for BSN

In case of critical condition, the patient's data should be transferred to a remote server for diagnosis and prescription. This requires the development of smart multi-agent system for healthcare services. In most of the projects such as Mobihealth [22], Politechnico[28] and Tele-medicare [29], the patient's medical information are extracted from PDA and forwarded to a central server in hospital using subsequent multi-agent operations. A multi-agent architecture proposed in [30] uses ontology based mobile agent for real time diagnosis. A multi-agent based healthcare system (MAHS) is presented in [31], which is mainly divided into three areas: BSN, Surrogate System and hospital subsystem. The surrogate system connects BSN and hospital subsystem. This multi-agent system is divided into five main agents: Patient Monitoring Agent (PMA), Gate Agent (GA), Supervisor Agent (SA), Manager Agent (MA) and Doctor Agent (DA). The combined operation of these agents provides patient monitoring, real time feedback to the patient and emergency management. PMA collects data from miniaturised sensors and forwards to surrogate system via SA. GA verifies patient's authentication for his requests. SA controls the surrogate system. MA controls the hospital subsystem. DA provides diagnosis and prescription based on the collected data to PMA. All these multi-agent systems for pervasive health care services require further investigation. The management of the huge amount of patient's data and determining patient's condition based on collected data is a challenging issue and requires advance data mining techniques.

5 Open Issues and Challenges

A proactive BSN system requires the resolution of many technical issues and challenges such as biosensor design, power scavenging issue, low power RF data paths, scalability, fault tolerance, low power communication protocol, mobility, interoperability, security and privacy. In on-body communication networks, biocompatibility is the most important issue. The biosensor is often placed on human body and its reliability is relied on the interface between the sensor and tissue or blood [32]. A number of biosensors are developed such as ECG sensor based on Telos platform [33], SpO2 sensor and ECG sensor based on a BSN node [34], DNA sensor [26], 3D accelerometers and gyroscope [23] and piezoresistive shear stress sensor [35]. Another important factor is battery life. The solution of some technical issues such as sensor design, RF design and low power MAC protocol contributes to extend the battery life. Lithium based batteries can operate at 1400-3600J/cc and provide long period of operation i.e. from few months to years [36]. A recently developed Sony product "Bio Battery" which generates electricity from sugar can be a promising candidate to solve the power scavenging issue [37]. IMEC developed a thermal micro power generator, which converts thermal energy to electrical energy [38]. The radio interface is also a major challenge and its power consumption in a BSN must be reduced below the energy scavenging limit (100 micro Watt) [25].

The current sensors nodes are mostly based on RF circuit design. Reducing the power consumption of RF transceiver plays a significant role in increasing the lifetime of a sensor. UWB technology is the best solution to increase the operating period of sensors. However the Power Spectral Density (PSD) must be calibrated inside the Federal Communication Commission (FCC) mask for indoor applications. The tiny biosensors wirelessly transmit the collected information to the central intelligent node. Design of a low power and secure communication protocol for a BSN is the most important issue. HTTP protocol is designed to transfer data to remote base station [39]. Chipcon CC2420 uses IEEE 802.15.4 (ZigBee) wireless link for transmitting physiological data between sensors. A cross layer protocol (MAC/Network layer) called Wireless Autonomous Spanning Tree Protocol (WASP) is presented where a spanning tree is set up autonomously and the network traffic is controlled by broadcasting scheme messages over the tree [40]. An extended version of WASP protocol called Cascading Information Retrieval by Controlling Access with Distributed Slot Assignment (CICADA) is presented, which guarantees low delay and good resilience to mobility [41].

6 Conclusions and Future Prospects

A Body Sensor Network (BSN) consists of miniaturised, invasive and non-invasive, low power autonomous sensor nodes, which are seamlessly placed on or implanted in human body in order to provide an adaptable and smart healthcare system. A successful BSN system requires the resolution of many technical issues and challenges, which includes but not limited to interoperability, QoS, privacy and security, low power RF data paths, power scavenging issue, biosensor design, scalability and mobility. Moreover, in implant communication, the implant transceiver needs to be sensitive on receive with the ability to tune the antenna for best response. In this paper, we briefly discussed In-body and On-body communication networks. We talked about the methodologies of wireless communication between implanted medical devices with external monitoring equipment. Moreover, we presented a comprehensive discussion on on-body communication networks with a special focus on the recent technological trend in this area. Technical issues and challenges in BSN have also been discussed. Future applications include smart health care services, remote diagnostic and telemedicine, wearable technology to monitor vital signs, smart nursing homes, emergency communication and patient's data maintenance. The broadband signaling scheme such as UWB is a promising candidate to satisfy power consumption budget of sensor nodes and is under investigation in different research institutes. To enable uplink communication from sink to nodes, the WASP and CICADA need to be improved.

References

1. http://thecommunityguide.org/nchs/fastats/lcod.htm
2. http://www.who.int/whosis/mort/profiles/mortwprokorrepofkorea.pdf
3. Borger, C., et al.: Health Spending Projections Through 2015: Changes on the Horizon. In: Health Affairs Web Exclusive W61 (February 22, 2006)
4. http://www.innovationmagazine.com/innovation/volumes/v7n3/feature2.shtml
5. Barroso, A., Benson, J., et al.: The DSYS25 sensor platform. In: Proceedings of the ACM sensys 2004 (2004)

6. Lo, B., Yang, G.Z.: Key Technical Challenges and Current Implementations of Body Sensor Networks. In: IEE Proceedings of the 2nd International Workshop on Body Sensor Networks (BSN 2005), pp. 1–5 (April 2005)
7. Otto, C., Milenkovic, A., Sanders, C., Jovanov, E.: System Architecture of a Wireless Body Area Sensor Network for Ubiquitous Health Monitoring. Journal of Mobile Multimedia 1(4), 307–326 (2006)
8. Lo, B., Atallah, L., Aziz, O., ElHelw, M., Darzi, A., Yang, G.Z.: Real time pervasive monitoring for post operative care. In: BSN 2007, Aachen, Germany (2007)
9. Lo, B., Yang, G.Z.: Architecture for Body Sensor Networks. In: IEE Proceedings of the Perspective in Pervasive Computing, October 25, 2005, pp. 23–28 (2005)
10. Milenkovic, A., Otto, C., Jovanov, E.: Wireless Sensor Networks for Personal Health Monitoring: Issues and an Implementation. Computer Communications 29, 2521–2533 (2006)
11. Sivard, A., Bradlet, P., et al.: The challenge of designing in-body communications, Embedded Systems Desing 10/26/04, 05:16 PM EDT
12. http://products.zarlink.com/productprofiles/ZL70101.htm
13. Finkenzeller, K.: RFID Handbook, 2nd edn. Wiley International, Chichester (2003)
14. MICS Band. Australian Communications Agency Paper SP 6/03 (October 2003)
15. Yang, et al.: Body Sensor Networks. Springer, Heidelberg (2006)
16. Kraus, J.D.: Antennas, 2nd edn. McGraw Hill, New York (1988)
17. Lee, F.L., Wei, C.: Microstrip and Printed Antennas, 1st edn. Wiley, Chichester (1997)
18. Fujimoto, K., Henderson, A., Hirasawa, K., James, J.R.: Small Antennas, 1987 Research Studies Press, 16 Coach House Cloisters, 10 Hitchin Street, Baldok, Hertfordshire, SG7 6AE, UK (1987)
19. Krall, A.D., McCorkle, J.M., Scarzello, J.F., Syeles, A.M.: The Omni Microstrip Antenna: A New Small Antenna. IEEE Trans. Antennas and Propagation AP27, 850–853 (1979)
20. Jovanov, E., Milenkovic, A., Otto, C., de Groen, P.: A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. Journal of NeuroEngineering and Rehabilitation 2(6) (March 1, 2005)
21. http://www.eecs.harvard.edu/mdw/proj/codeblue/
22. http://www.mobihealth.org/
23. http://www.cs.wustl.edu/jain/cse574-06/ftp/medicalwireless/index.html
24. http://www.ubimon.net
25. Ryckaert, J., Fort, A., Gyselinckx, B.: Ultra-wideband communication for wireless body area networks. In: International Workshop on UWB Technologies - IWUWBT (2005)
26. http://www.imec.be/
27. Zasowski, T., Althaus, F., Stager, M., Wittneben, A., Troster, G.: Uwb for non invasive wireless body area networks:channel measurements and results. In: IEEE Conference on Ultra Wideband Systems and Technologies, UWBST 2003, Reston, Virginia, USA (November 2003)
28. Winnem, O.M., Walderhaug, S.: TeleMediCare Project, SINTEF Telecom and Informatics (2002)
29. Ouchi, K., Suzuki, T., Doi, M.: LifeMinder: A Wearable Healthcare Support System Using User's Context. In: Proceedings of the 22nd ICDCSW 2002 (2003)
30. Kim, S.: Ubiquitous Healthcare: The OnkoNet Mobile Agents Architecture, NODe, pp. 265–277 (2003)

31. Kang, E., Im, Y., Kim, U.: Remote Control Multi-Agent System for u-Healthcare Service. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 636–644. Springer, Heidelberg (2007)
32. Lo, B., Yang, G.Z.: Key Technical Challenges and Current Implementations of Body Sensor Networks. In: IEE Proceedings of the 2nd International Workshop on Body Sensor Networks (BSN 2005), April 2005, pp. 1–5 (2005)
33. http://www.ece.uah.edu/jovanov/whrms/
34. http://vip.doc.ic.ac.uk/bsn/m185.html
35. Hsieh, M.C., Fang, Y.K., Ju, M.S., Chen, G.S., Ho, J.J., Yang, C.H., Wu, P.M., Wu, G.S., Chen, T.Y.: A BioMEMS Review: MEMS Technology for Physiologically Integrated Devices. J. Microelectromech. Syst. 10, 121–127
36. Yeatman, E.M.: Advances In Power Sources For Wireless Sensor Nodes. In: Proceedings of the 1st International Workshop on Wearable and Implantable Body Sensor Networks, April 6-7 (2004)
37. http://www.sony.net/SonyInfo/News/Press/200708/07-074E/index.html
38. http://www.imec.be/wwwinter/research/en/human/ambulatoryeeg.pdf
39. Dokovsky, N., van Halteren, A., Widya, I.: BANip: enabling remote healthcare monitoring with Body Area Networks. In: Guelfi, N., Astesiano, E., Reggio, G. (eds.) FIDJI 2003. LNCS, vol. 2952, pp. 62–72. Springer, Heidelberg (2004)
40. Braem, B., Latre, B., Moerman, I., Blondia, C., Demeester, P.: The Wireless Autonomous Spanning tree Protocol for multihop wireless body area networks. In: Proceedings of the First International Workshop on Personalized Networks, ICST, San Jose, California, USA (2006)
41. Latre, B., Braem, B., Moerman, I., Blondia, C., Reusens, E., Joseph, W., Demeester, P.: A Low-delay Protocol for Multihop Wireless Body Area Networks, Ghent University IMEC

Complexity of Road Networks as Agents' Environments

– Analyzing Relevance between Agent Evaluations and Their Environments –

Kazunori Iwata[1], Nobuhiro Ito[2], Youhei Kaneda[3], and Naohiro Ishii[2]

[1] Dept. of Bussiness Administration, Aichi University
[2] Dept. of Applied Infromation Science, Aichi Institute of Technology
[3] Dept. of Compute Science and Engineering, Nagoya Institute of Technology
kazunori@vega.aichi-u.ac.jp, n-ito@aitech.ac.jp,
agent-staff@phaser.elcom.nitech.ac.jp, ishii@aitech.ac.jp

Abstract. This paper is a first attempt to analyze the relevance between evaluations of agent systems and environments of the agents. We define a road network and complexity indexes for maps in order to analyze them as environments for agents. In addition, we perform evaluation experiments, in which an agent moves on a map like a vehicle. We investigate the relevance between the analysis results and the experimental results. Finally, we show that the analysis of maps is useful in evaluating agent systems.

1 Introduction

In this paper, we analyze the relevance between the evaluation of agent systems and the environments of agents. We focus especially on multi-agent simulations with maps being the environments of the agents. In the research area of agents, it is a challenge to find a method with which to evaluate the behavior of agents or multi-agent systems, as the evaluation depends on the environments in which agents perform [1,2]. Hence, it is necessary to clarify the relevance of the evaluation of multi-agent systems and the environments of agents in order to find a method to evaluate multi-agent systems keeping their environments in mind. Moreover, to clarify the relevance, the environments should be analyzed and quantified. Hence, this paper is a first attempt to analyze the relevance between the evaluation of agent systems and the environments of the agents. In this paper, we define a road network and complexity indexes for maps in order to analyze these as environments for agents. In addition, we perform evaluation experiments, in which an agent moves on a map like a vehicle. We investigate the relevance between the analysis results and the experimental results. Finally, we consider the results of the investigations.

2 Evaluation of Multi-agent Systems

Multi-agent systems refer to the systems in which a number of agents interact with one another to solve complicated issues [3]. In this paper, "agents" refer to

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 474–484, 2008.

actors that are able to recognize their circumstances by interacting with their environments and can then solve many kinds of problems autonomously.

In the research area of agents, it is a challenge to find a method to evaluate the behavior of agents or multi-agent systems, because there is no explicit evaluation method for agents that transcends the development of agents' theory and the implementation and practical application of agents. However, it is difficult to specify evaluations for agents due to interdependence of agents and their environments. To illustrate the difficulty of evaluations, consider the following:

(i) If agents *A1* and *A2* work in an environment *E*, it is easy to evaluate which agent is better suited to the environment. However, the result of the evaluation will not be applicable to other environments besides *E*.
(ii) If agent *A1* can work better than agent *A2* in environment *E1* and agent *A2* can work better than agent *A1* in environment *E2*, it is difficult to decide which agent performs better.

These examples illustrate why it is necessary to clarify the relevance of agents and their environments in the evaluation of agents and multi-agent systems. In order to clarify the relevance, a detailed analysis of the environment is required. In this paper, we focus on an agent which moves on maps. In other words, we select an autonomous robot agent, because there are many kinds of multi-agent systems. Hence, we concentrate on map data as the environment and analyze it to clarify the relevance. The map data includes several features, such as roads, railroads, rivers, buildings, geographical features and so on. We note especially road information and analyze it, because there is a strong relation between roads and the evaluation of agents, which move on maps.

3 Maps and a Road Network

3.1 Map Data and Road Information

We use 1/25,000 map data of all the areas of Japan. This data includes roads, railroads, rivers, shorelines of lakes, coastlines, administrative districts, ground control points, place names, public institutions and altitude. These maps are released by the Geographical Survey Institute[4]. The data is described in G-XML, which is the Japanese Industrial Standards' format. We use the information about roads, which consists of nodes of road and road edges. Nodes of road are intersections, blind alleys and junctions between roads, and have the attributes of "ID", "longitude" and "latitude". Road edges are roads, each of which connects the interval between two nodes of road. Their data attributes are "ID", "type", "width of road", and "IDs of two corner points" which are nodes of road.

3.2 A Road Network

In this paper, we regard the information about roads as a road network and define a road network using the attributes given by the information. We denote a road network as G, as a weighted digraph (V, E). "V" indicates a point set regarded as a node set in a graph and defined in Definition 1. "E" indicates a road set regarded as an edge set in a graph and defined in Definition 2.

Definition 1. ***Point Set***

$$V = \{v | v \text{ has ``longitude'' and ``latitude''}\}, \tag{1}$$

where v is a node created by a node of road.

Definition 2. ***Road Set***

$$E = \{e | e \text{ has } v_{head},\ v_{tail},\ \text{''length'' and ''width''}, v_{head} \in V, v_{tail} \in V\}, \tag{2}$$

where e is a directed edge from v_{head} to v_{tail} and created by a road edge, "length" means the Euclidean distance from v_{head} to v_{tail} and "width" equals "width of road".

Definition 3. ***Road Network***
A road network G is a weighted digraph (V, E) with two weighting functions l and w. These functions are defined as follows:

$$l : E \to \mathbf{R}, \tag{3}$$

$$w : E \to \mathbf{R}, \tag{4}$$

where l derives a real-valued attribute "length" from an edge e and w derives a real-valued attribute "width" from an edge e.

A road network G is in the X-Y Cartesian coordinate system where the x- and y-axes represent latitude and longitude, respectively. Each road e in G is a segment of line connecting a v_i and v_j ($v_i, v_j \in V$ and $v_i \neq v_j$). In addition, G has the following properties:

(i) Strongly-connected digraph.
(ii) No loop back edge.
(iii) No multiple edges between any pair of nodes.

4 Analysis of Road Networks

4.1 Definitions for Analysis

In a road network $G = (V, E)$, the distance of a path $p =< v_0, v_1, \ldots, v_k >$ is obtained by calculating the sum of the length weights for all elements in the path [5,6,7]. It is denoted by $l(p)$ and calculated as follows:

$$l(p) = \sum_{i=1}^{k} l(e_{v_{i-1},v_i}), \tag{5}$$

where e_{v_{i-1},v_i} is a directed edge from $v_{head} = v_{i-1}$ to $v_{tail} = v_i$.

Definition 4. ***Distance of Shortest Path***
Let $p_{u,v}$ denote a path from u to v ($u \neq v$), then the set of shortest paths $sp_{u,v}$ is calculated as follows:

$$SP_{u,v} = \underset{p}{\operatorname{argmin}}\, l(p) \in \{p|\forall p_{u,v}\}. \tag{6}$$

Hence, the distance of the shortest path $\delta(u,v)$ is as follows:

$$\delta(u,v) = l(sp_{u,v}), \tag{7}$$

where $sp_{u,v} \in SP_{u,v}$.

Definition 5. ***Available Roads Rate, Available Length Rate in Roads, Available Area Rate in Roads***
In a road network $G = (V, E)$, the available roads rate (ARR), available length rate in roads (ALR) and available area rate in roads (AAR) are calculated using Eqs. (8), (9) and (10), respectively:

$$ARR = \frac{|E_{WR}|}{|E|}, \tag{8}$$

$$ALR = \frac{\sum_{e_2 \in E_{WR}} l(e_2)}{\sum_{e_1 \in E} l(e_1)}, \tag{9}$$

$$AAR = \frac{\sum_{e_2 \in E_{WR}} l(e_2) \times w(e_2)}{\sum_{e_1 \in E} l(e_1) \times w(e_1)}, \tag{10}$$

where let $E_{WR} = \{e|w(e) \geq 5.5m\}$

Let s be an entrance node and t be an exit node in $G = (V, E)$ $(s, t \in V)$, then the flow of G is a real function $f : E \rightarrow \mathbf{R}$ with the following three properties for all nodes u and v[8]:

(i) **Capacity constraints:** $e_{u,v} \in E$, $f(e_{u,v}) \leq w(e_{u,v})$. The flow along an edge cannot exceed its capacity.
(ii) **Skew symmetry:** $f(e_{u,v}) = -f(e_{v,u})$. The net flow from u to v must be the opposite of the net flow from v to u.
(iii) **Flow conservation:** $\sum_{w \in V} f(e_{u,w}) = 0$, $_{unless\ u=s\ or\ u=t}$. The flow to a node is zero, except for the source, which "produces" flow, and the sink, which "consumes" flow.

The maximum flow $|f_{max}|$ of G is defined in Definition 6.

Definition 6. ***Maximum Flow***

$$|f_{max}| = \max\left\{ \sum_{e_{s,u} \in E} f(e_{s,u}) \middle| \sum_{e_{s,u} \in E} f(e_{s,u}) = \sum_{e_{v,t} \in E} f(e_{v,t}) \right\} \tag{11}$$

The ratio of intersections RI is calculated in Definition 7.

Definition 7. ***Ratio of Intersections***

$$RI = \frac{|V_I|}{|V|}, \tag{12}$$
$$V_I = \{v_I \in V | (id_{v_I} \geq 3) \wedge (od_{v_I} \geq 3)\},$$

where id_{v_I} is the indegree of v_I and od_{v_I} is the outdegree of v_I. The node set V_I is the set of all intersections in G.

The degree of two edges, $e_{x,v}$ and $e_{v,z}$, is denoted as $a_{x \to v \to z}$ ($0° < a_{x \to v \to z} \leq 180°, x, v, z \in V, x \neq v \neq z$). The ratio of arranged area (RAA), which indicates similarity between a graph G and a square grid, is given by Definition 8.

Definition 8. ***Ratio of Arranged Area***

$$RAA = \frac{|A_{AR}|}{|A|}, \tag{13}$$

where $A = \{a_{x \to v \to z}\}$ and $A_{AR} = \{a_{x \to v \to z} | (75° \leq a_{x \to v \to z} \leq 105°) \vee (165° \leq a_{x \to v \to z} \leq 180°)\}$

The ratio of straight connected edges and the ratio of straight connected edges to distance are denoted by RSE and $RSED$, respectively.

Definition 9. ***Ratio of Straight Edge and Ratio of Straight Edge to Distance***

$$RSE = \frac{|V_{S(v_0,v_n)}|}{|V_{M(v_0,v_n)}|}, \tag{14}$$

$$RSED = \frac{|V_{S(v_0,v_n)}|}{\delta(v_0, v_n)}, \tag{15}$$

where $V_{M(v_0,v_n)} = \{v_1, \ldots, v_{n-1}\}$ is the middle point of the shortest path and $V_{S(v_0,v_n)} = \{v_i | v_i \in V_M(v_0, v_n), a_{v_{i-1} \to v_i \to v_{i+1}} \geq 170°, 1 \leq i \leq n-1\}$ denotes the middle point set of the almost straight path ($a_{v_{i-1} \to v_i \to v_{i+1}} \geq 170°$).

The local road-connectivity is defined in Definition 10.

Definition 10. ***Local Road-Connectivity***
The local road-connectivity of two edges $m, n \in V$ is the size of the smallest edge cut disconnecting m from n.

RPR in Definition 11 is the ratio of the number of usage of an edge e to the total number of shortest paths.

Definition 11. ***Ratio of Passing through a Road***

$$RPR = \frac{u}{u_{max}}, \tag{16}$$

where $u = \sum_{i,j \in V} |SP_{i,j}|$ and u is the total number of times an edge e is used by all shortest paths.

4.2 Complexity Measures for Road Networks

Having analyzed road networks, we define five complexity measures for road networks using Definitions 4∼11 with other indexes.

Definition 12. ***Complexity Measures for Road Networks***

(a) ***Complexity of length****: This indicates how many miles separate two points using the average length of all edges($e \in E$) and the average "distance of shortest paths" in Definition 4 for all pairs of nodes.*
(b) ***Complexity of width****: This indicates how wide a road network is using ARR, ALR, AAR in Definition 5, the average width of all edges($e \in E$) and the average "maximum flow" in Definition 6 for all pairs of nodes.*
(c) ***Complexity of density****: This indicates the density of a road network using RI in Definition 7.*
(d) ***Complexity of distortion****: This indicates how much distortion there is in a road network using RAA in Definition 8 and the average RSE and RSED for all pairs of nodes in Definition 9.*
(e) ***Complexity of weakness****: This indicates how weak a road network is using the average local road-connectivity in Definition 10 for all pairs of nodes and the average RPR for all edges($e \in E$) in Definition 10.*

5 Evaluation Experiments

5.1 Evaluation Simulations

We implemented a vehicle movement simulator and performed evaluation experiments on the complexities using the simulator. The conditions for the simulation are as follows:

(i) Using 1/25,000 maps of different 28 areas: we divide the map of Japan into 10 regions based on the 8th Area Classification. We then select the following three types of maps for each region:
  (a) a cabinet-order designated city (two areas have no cabinet-order designated city),
  (b) a city, and
  (c) a town or a village.
(ii) Using a vehicle agent: the agent in the simulation is modeled on a vehicle, that keeps within the speed limit on straight roads and slows down when turning corners depending on the angle of the corner.
(iii) Using two different agents, which have different search path algorithms:
  (a) vehicle agent A moves along the shortest path, and
  (b) vehicle agent B moves along a path which minimizes migration time by considering slowdowns at corners.

We established two indexes for the simulation which indicates how efficiently each agent moves. These are

(a) **Migration Time**: which is the time taken by an agent to move from the place of departure to the destination.
(b) **Mean Speed**: which gives the average speed of an agent while moving from the place of departure to the destination.

5.2 Relevance of Complexity Measures and Simulation Results

As space is limited, we omit the details of the analysis results for road networks and the simulation results, and concentrate on the relevance of the complexity measures and the simulation results.

The correlation coefficients between the complexity measures and the migration time or mean speed of vehicle agent A and vehicle agent B are given in Tables 1 and 3, respectively. The strength of the correlation coefficients for vehicle agent A and vehicle agent B are shown in Tables 2 and 4, respectively, where the underlined data in italic represent different estimations.

Table 1. Correlation Coefficients between Complexity Measures and Migration Time or Mean Speed of Vehicle Agent A

	Correlation Coefficients between Complexity Measures and	
	Migration Time	Mean Speed
Average length of all edges	0.7367	0.5997
Average distance of shortest paths	0.9368	0.5616
Average width of all edges	-0.3968	-0.0898
ARR	-0.2567	0.5227
ALR	-0.3868	0.4383
AAR	-0.3325	0.4431
Average maxmum flow	-0.5719	-0.2234
RI	-0.8386	-0.2641
RAA	-0.5250	-0.1196
RSE	-0.4271	0.0626
$RSED$	-0.5240	-0.4610
RPR	0.5151	0.2495
Average local road-connectivity	-0.7158	-0.2617

5.3 Consideration of the Relevance

The correlation coefficients between the migration time of vehicle agents A and B and the average width of all edges, ARR, ALR and AAR show weak correlations, whereas those between the other measures and the migration time show medium to strong correlations. We now consider the strong correlations.

(a) The correlations for average length of all edges and the average distance of the shortest paths show that the longer the road is on the map, the more time is required for the movement of a vehicle agent on it.

Table 2. Strength of Correlation Coefficients

	Migration Time	Mean Speed
Average length of all edges	**Strong**	*Medium*
Average distance of shortest paths	**Strong**	*Medium*
Average width of all edges	*Weak*	Little
ARR	Weak	Medium
ALR	Weak	Medium
AAR	Weak	Medium
Average maximum flow	Medium	*Weak*
RI	**Strong**	*Weak*
RAA	Medium	Little
RSE	Medium	*Little*
RSED	Medium	*Medium*
RPR	Medium	*Weak*
Average local road-connectivity	**Strong**	*Weak*

Table 3. Correlation Coefficients between Complexity Measures and Migration Time or Mean Speed of Vehicle Agent B

	Correlation Coefficients between Complexity Measures and	
	Migration Time	Mean Speed
Average length of all edges	0.7571	0.3722
Average distance of shortest paths	0.9424	0.3649
Average width of all edges	-0.4390	0.1937
ARR	-0.2645	0.6695
ALR	-0.3879	0.5686
AAR	-0.3420	0.6236
Average maximum flow	-0.6104	-0.0859
RI	-0.8507	-0.0552
RAA	-0.5570	-0.1448
RSE	-0.4505	0.2812
RSED	-0.5582	-0.1778
RPR	0.5119	0.1442
Average local road-connectivity	-0.7323	-0.0538

(b) The correlation for *RI* shows negative correlation. This means that a vehicle agent can move easily on a map with a high rate of intersections.

(c) The correlation for average local road-connectivity shows a negative correlation. This indicates that the more paths that exist to a destination, the less time a vehicle requires to move to that destination.

The correlation coefficients for the mean speed of vehicle agents A and B show medium, weak or little correlations. The correlation coefficients for *ARR*, *ALR* and *AAR* show medium correlations. These results show that the greater the width is of each road on a map, the higher the speed with which agents can

Table 4. Strength of Correlation Coefficients

	Migration Time	Mean Speed
Average length of all edges	**Strong**	*Weak*
Average distance of shortest paths	**Strong**	*Weak*
Average width of all edges	*Medium*	Little
ARR	Weak	Medium
ALR	Weak	Medium
AAR	Weak	Medium
Average maximum flow	Medium	*Little*
RI	**Strong**	*Little*
RAA	Medium	Little
RSE	Medium	*Weak*
RSED	Medium	*Little*
RPR	Medium	*Little*
Average local road-connectivity	**Strong**	*Little*

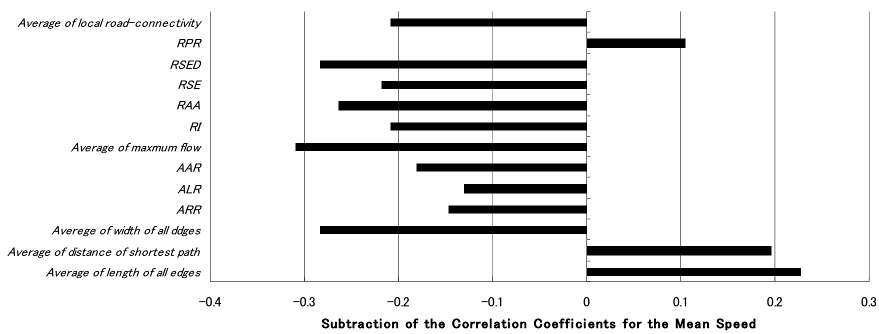


Fig. 1. (Correlation Coefficients for the Mean Speed of Vehicle Agent A) − (Correlation Coefficients for the Mean Speed of Vehicle Agent B)

move. Moreover, we consider the subtraction of the correlation coefficients for the mean speed of vehicle agent B from those for vehicle agent A. The results are shown in Fig. 1.

The results imply that vehicle agent A can move more speedily on maps which have larger values for the average length of all edges and the average distance of the shortest paths. Vehicle agent B can move more speedily on maps which have larger values for *ARR*, *ALR* and *AAR*. In order to evaluate the results, we investigate vehicle agent A by:

(a) Selecting the top 10 maps for each index, namely the average length of all edges and the average distance of the shortest paths.
(b) Calculating the average speed for the selected maps.
(c) Comparing the average speed with that for all the maps.

In order to evaluate the results, we investigate vehicle agent B by: the following:

(a) Selecting the top 10 maps for the indexes ARR, ALR and AAR.
(b) Calculating the average speed for the selected maps.
(c) Comparing the average speed with that for all the maps.

Table 5 shows the results of the comparisons.

Table 5. Comparing Average Speed

		Average Speed	Standard Deviation of Speed
Vehicle Agent A	Selected Maps	24.53	2.579
	All Maps	22.14	2.998
Vehicle Agent B	Selected Maps	26.83	2.113
	All Maps	24.25	3.205

The results imply that we can specify a map using the complexity indexes so that an agent efficiently.

6 Conclusion and Future Work

This paper was the first attempt to analyze the relevance of evaluation of multi-agent systems and the environments of the agents. We have defined a road network and its complexity indexes in this paper and have analyzed the indexes for many areas. In addition, we have shown the relevance between vehicle simulation and the indexes by investigating the results of the simulations and the analysis. The results of the investigations indicate that the complexity indexes are useful in specifying the maps on which an agent moves effectively.

Our future work includes the following:

(a) Investigating other environments for multi-agent systems besides maps.
(b) Analyzing the relevances amongst many kinds of multi-agent systems and their environments.
(c) Suggesting ways to evaluate multi-agent systems taking account of their environments.

References

1. Kinoshita, T., Sugawara, K.: Agent Oriented Computing ∼ Foundation and Application of Agents∼ in Japanese. Soft Research Center (1995)
2. Pynadath, V.D., Tambe, M.: Multiagent teamwork: Analyzing the optimality and complexity key theories and models. In: Proceedings of International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), pp. 873–880 (2002)
3. The Japanese Society for Artificial Intelligence, ed.: Encyclopedia of Artificial Intelligence in Japanese. KYORITSU SHUPPAN CO., LTD (2005)
4. GSI: 1/25,000 map information of Japan (in Japanese) (2003), http://sdf.gsi.go.jp/

5. Asano, T., Iwano, K., Umeo, H., Yamashita, M., Wada, K.: Algorithm Introduction in Japanese, vol. 1. Kidai Kagaku sha Co., Ltd (1995)
6. Asano, T., Iwano, K., Umeo, H., Yamashita, M., Wada, K.: Algorithm Introduction in Japanese, vol. 2. Kidai Kagaku sha Co., Ltd (1995)
7. Harary, F.: Graph Theory: New Ed edn. Westview Press (1994)
8. Wikipedia: Flow network (2007), http://en.wikipedia.org/wiki/Flow_network

Resource Limitations, Transmission Costs and Critical Thresholds in Scale-Free Networks

Chung-Yuan Huang[1,2], Chia-Ying Cheng[3,*,**], and Chuen-Tsai Sun[3]

[1] Department of Computer Science and Information Engineering
[2] Research Center for Emerging Viral Infections
Chang Gung University
259 Wen-Hwa 1st Road, Taoyuan 333, Taiwan, R.O.C.
[3] Department of Computer Science
National Chiao Tung University
1001 Ta Hsueh Road, Hsinchu 300, Taiwan, R.O.C.
[*] gis92566@cis.nctu.edu.tw

Abstract. Whether or not a critical threshold exists when epidemic diseases are spread in complex networks is a problem attracting attention from researchers in several disciplines. In 2001, Pastor-Satorras and Vespignani used a computational simulations approach to show that epidemic diseases which spread through scale-free social networks do not have positive critical thresholds. In other words, even if a disease has almost no chance of being transmitted from one person to another, it can still spread throughout a scale-free network. However, they ignored two key factors that have a large impact on epidemic dynamics: economic resource limitations and transmission costs. Every infection event entails tangible or intangible costs in terms of time, energy, or money to the carrier, recipient, or both. Here we apply an agent-based modeling and network-oriented computer simulation approach to analyze the influences of resource limitations and transmission costs on epidemic dynamics and critical thresholds in scale-free networks. Our results indicate that when those resources and costs are taken into consideration, the epidemic dynamics of scale-free networks are very similar to those of homogeneous networks, including the presence of significant critical thresholds.

Keywords: Critical threshold, economic resource limitations, transmission costs, scale-free networks, power-law degree distribution.

1 Introduction

Whether or not a critical threshold exists when epidemic diseases are spread in complex networks is a problem attracting attention from researchers in a range of disciplines [1, 7, 9-13]. The critical threshold is one of the most important indicators of whether or not an epidemic outbreak has occurred. When considering epidemic dynamics, critical thresholds serve as gates for explaining why some epidemic diseases with small numbers of patients at the initial stage evolve into pandemic diseases,

** Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 485–494, 2008.

while others disappear shortly after the initial stage. This situation resembles "non-equilibrium phase transitions" of concern to epidemiologists. Viable answers to the epidemic/pandemic question can help epidemiologists, public health professionals, and computer scientists to better understand core epidemic disease questions, estimate epidemic dynamics and diffusion situations, and develop effective and public health policies and immunization strategies [3, 9, 13].

Pastor-Satorras and Vespignani report that epidemic diseases which spread within scale-free networks do not have positive critical thresholds. In other words, even if a disease has almost zero chance of being passed from one person to another, it can still be spread within a scale-free network. Researchers who have studied epidemic dynamics and critical thresholds in scale-free networks based on Pastor-Satorras and Vespignani's results have consistently concluded that regardless of transmission capability, every infectious disease, computer virus, and cultural trend has a very high probability of being stably transmitted and surviving in a scale-free network [1, 3, 6]. In short, scale-free networks such as those associated with human sexual contact and the Internet are thought to be capable and vulnerable platforms for spreading epidemic diseases such as HIV/AIDS and computer viruses.

Pastor-Satorras and Vespignani's [9-13] conclusions are sources of great concern among public health professionals and companies concerned with fighting computer viruses. However, they overlook an important factor—most epidemiological models fail to consider individual differences and resource limitations associated with human interactions and contacts. During transmission processes, every infection event entails tangible or intangible costs in terms of time, energy, or money to the carrier, recipient, or both. In this paper we will refer to all economic resources consumed during the transmission process as transmission costs. Our results indicate that when those resources and costs are taken into consideration, the epidemic dynamics of scale-free networks are very similar to those of homogeneous networks, including the presence of significant critical thresholds. We believe these results will be useful to epidemiologists and public health professionals working on epidemic dynamics and critical thresholds in scale-free networks, as well as computer and "smart phone" experts concerned with computer virus transmission.

2 Epidemic Dynamics in Complex Networks

In complex networks, nodes are used to represent entities in biological environments or on the Internet. Links indicate close relationships or certain interaction channels between two entities; those with direct connections are called neighbors [8]. When simulating the transmission dynamics of epidemic diseases in complex networks, epidemiologists usually assume that nodes in complex networks randomly run through an SIS cycle (*S*usceptible $\rightarrow$ *I*nfected $\rightarrow$ *S*usceptible). During each time step, each susceptible node is subject to a probability ν infection rate if it is connected to one or more infected nodes. Infected nodes recover at a probability ε recovery rate and once again become susceptible. Based on the definitions of the infection ν and recovery ε rates, effective spreading rate λ is defined as $\lambda = \nu/\delta$. Without a lack of generality, recovery rate ε can be assigned a value of 1, since it only affects individuals during a period of disease propagation. Pastor-Satorras and Vespignani [9] define

$\rho(t)$ as the density of infected nodes at time t. When time t becomes infinitely large, ρ can be represented as a steady density of infected nodes. Using these definitions, Pastor-Satorras and Vespignani applied dynamic mean-field theory to the SIS model and assumed homogeneous mixing hypothesis according to the topological features of homogeneous networks for obtaining the stable density of infected nodes ρ during long time periods (Eq. 1) as well as the critical threshold λ_c (Eq. 2). According to Equations 1 and 2, a positive and nonzero critical threshold λ_c exists in a homogeneous network based on the SIS model. If the value of the effective spreading rate exceeds the critical threshold ($\lambda \geq \lambda_c$), the infection spreads and gains persistence. If the effective spreading rate is below the critical threshold ($\lambda < \lambda_c$), the infection dies at an exponential speed. In summary, the primary prediction of an SIS epidemiological model in a homogeneous network is the presence of a positive critical threshold, proportional to the inverse of the $<k>$ average number of neighbors of every node, below which epidemics die and endemic states are impossible.

Pastor-Satorras and Vespignani relaxed their homogeneity assumption for homogeneous networks and obtained the critical threshold λ_c in a scale-free network as Equation 3. The results indicate that in scale-free networks with a connectivity exponent of $2 < \gamma \leq 3$ and for which $<k^2> \rightarrow \infty$ is the limit of a network of infinite size ($N \rightarrow \infty$), the critical threshold λ_c is very close to 0 ($\lambda_c \rightarrow 0$).

$$\rho = \begin{cases} 0 & \lambda < \lambda_c \\ \dfrac{\lambda - \lambda_c}{\lambda} & \lambda \geq \lambda_c \end{cases} \tag{1}$$

$$\lambda_c = \frac{1}{<k>} \tag{2}$$

$$\lambda_c = \frac{<k>}{<k^2>} \tag{3}$$

Pastor-Satorras and Vespignani [13] express the total prevalence ρ for the SIS epidemiological model in a BA scale-free network as a function of the effective spreading rate λ, and compare it to the theoretical prediction for a homogeneous network. As shown in Figure 1, the total prevalence ρ of a BA scale-free network reaches 0 in a continuous and smooth way when the effective spreading rate λ decreases; this indicates an absence of any critical threshold ($\lambda_c = 0$) in a BA scale-free network. As long as $\lambda > 0$, epidemic diseases can be stably transmitted in the network and eventually reach a steady state. This explains why scale-free networks are fragile in epidemiological spreading situations. Since social networks and the Internet both have “the rich get richer” properties, computer viruses, biologically infectious diseases, and cultural trends can be stably transmitted even when initial infection cases occur in small, limited areas.

These conclusions explain the epidemic outbreak mechanisms of biologically infectious diseases and computer viruses. According to traditional epidemiological theory, large-scale pandemics only occur when the effective spreading rate exceeds a specific critical threshold. However, Pastor-Satorras and Vespignani claim that infections can proliferate in scale-free networks regardless of their effective spreading rates. This idea represents major threats to public health concerns and computer data.

For finite-size scale-free networks, Pastor-Satorras and Vespignani [12] introduced the concept of maximum connectivity k_c (dependent on N), which has the effect of restoring a boundary in connectivity fluctuations and inducing an effective nonzero critical threshold. According to the definition of the maximum connectivity k_c, $<k^2>$ in Eq. 3 clearly has a finite value in finite-size scale-free networks. However, in this situation the critical threshold vanishes as network size increases.

3 Economic Resource Limitations and Transmission Costs

According to analytical calculations and numerical simulations presented by Pastor-Satorras and Vespignani [9-13], scale-free networks do not have critical thresholds. In other words, they argue that in scale-free networks any infectious disease with a low spreading rate can become an epidemic, which raises serious concerns for epidemiologists, public health professionals, and computer scientists. However, in a world where new infectious diseases, cultural trends, and computer viruses are emerging every day, we know that very few reach epidemic proportions or persist in social networks or the Internet; a majority dies almost immediately following genesis. This situation contradicts Pastor-Satorras and Vespignani's conclusions, and served as our motivation to look at limitations in transmission processes other than complex network topological features that make it difficult for infectious diseases, trends, or computer viruses to persist.

In addition to the topological characteristics of complex networks, individual differences and environmental factors exert considerable influences on transmission dynamics and epidemic disease diffusion. Huang et al. [5] used Watts and Strogatz's small-world networks to investigate the influence of individual differences on epidemic simulations. Specifically, they used a sensitivity analysis to show that when an agent-based modeling and network-oriented computer simulation approach is applied to exploring epidemic transmission dynamics in small-world networks, researchers should focus not only on network topological features, but also on proportions of specific values of individual differences related to infection strength or resistance. Less emphasis should be placed on details of the topological connection structures of small-world networks and the distribution patterns of individual difference values.

We believe that many of the studies published in the past five years have been overly focused on ways that the topological features of social networks and the Internet affect epidemic dynamics and critical thresholds. This has occurred at the expense of two key factors: individual differences and economic resources associated with human interactions and contacts. Pastor-Satorras and Vespignani's work on the topological features of scale-free networks is rightfully considered important in terms of epidemiological modeling; their ideas have inspired many studies on critical thresholds and immunization strategies. However, in their analytical calculations and numerical simulation models they assume that transmission events are cost-free. Such an assumption may be suitable for computer viruses that are spread via emails that contain large numbers of recipient addresses, but it is not accurate when applied to infectious diseases or cultural trends.

Economists view resources as tangible or intangible. Available amounts of resources such as time, energy, and money are usually limited. Furthermore, many resources are non-reproducible. Carriers of infectious diseases who spend resources on specific recipients cannot reuse the same resources on other recipients; conversely, recipients cannot reuse resources spent on individual carriers. Using the definition of *cost* in economics, we will define all resources spent on transmission processes as "transmission costs," and use our agent-based modeling and network-oriented computer simulation approach to (a) prove the appropriateness of introducing the concepts of resource limitations and transmission costs into epidemiological research, and (b) show how doing so exerts a positive impact on public health policies and immunization strategies.

Our task was to find a nonzero, positive critical threshold in a scale-free network under real conditions of economic resource limitations and transmission costs. To this end we applied the state transfer concept of SIS models adopted by Pastor-Satorras and Vespignani [9] as our core simulation model architecture and incorporated parameters to simulate behavioral and transformative results arising from agent interactions. When designing our simulation model we took into consideration the effects of network structure, differences among individuals, economic resource limitations, and transmission costs to create a system in which individuals express ranges of behavioral patterns.

Each agent in a complex network owns a set of properties and behavioral rules used to demonstrate the features and statuses of persons in social networks or computers connected to the Internet. A link between two nodes means that the connected agents have a close relationship or share a specific interaction channel. An infectious disease or computer virus can be transmitted via this link. At each discrete time step, the epidemiological state of each node is determined by its behavioral rules, original epidemiological state, neighbors' epidemiological states, infection rate, and recovery rate. As stated above, $\rho(t)$ is defined as the density of infected nodes present during time step t. When time step t becomes infinitely large, ρ can be represented as a steady density of infected nodes.

A complex network $G(N, M)$ with N nodes and M links is constructed using the algorithm described in Section 3 prior to setting relevant parameters and attributes for the nodes involved in the simulation; discrete time t is set at 0. During a simulation, nodes take turns interacting with neighboring agents for specified time intervals. Note that individual agent interactions do not result in immediate influences and that simultaneous state changes only occur when all agents in a complex network complete their interactions. Accordingly, interaction sequences do not influence interaction processes or results. At the beginning of each discrete time step, the usable economic resources of each agent v_i are reset to $R(v_i)$, meaning that all agents renew and/or receive supplemental resources. For example, for most people their energy level is revived after a night of sleep. In our later experiments, the statistical distribution of individual economic resources can be delta ($r_{Constant}$), uniform, normal, or power-law, as long as the mean value $<r>$ satisfies Eq. 4.

At each discrete time step, each v_j agent chooses and interacts with one neighboring agent from all of its *Neighbors*(v_i). After the interaction process is finished, agents v_i and v_j must have transmission costs $c(v_i)$ and $c(v_j)$ ($0 \le c(v_i) \le R(v_i)$ and $0 \le c(v_j) \le R(v_j)$) deducted from their respective economic resources regardless of the interaction result. If $R(v_i) < c(v_i)$ after the interaction, agent v_i cannot interact with other neighbors because all of its resources have been used up. Otherwise, it repeats the interaction process by choosing another neighboring agent until its resources are exhausted. Assume that infected and contagious agent v_i is adjacent to susceptible and infection-prone agent v_j. When the two agents come into contact, a combination of infection rate $Rate_{Infect}$, agent v_j's resistance level, and a random number r determines whether or not v_j is infected by v_i. If the random number r is lower than the infection rate $Rate_{Infect}$, agent v_j's epidemiological state becomes I (Infected). Simultaneously, infected agents are cured and become susceptible with a recovery rate $Rate_{Reset}$. Without a lack of generality, recovery rate $Rate_{Reset}$ can be assigned as 1, meaning if agent v_j is infected by other agents at discrete time step $t - 1$, it will recover and become susceptible at discrete time step t, since it only takes affect according to the definition of the infection disease propagation time scale.

$$\langle r \rangle = \frac{\sum_{i=1}^{N} R(v_i)}{N} = r_{Constant} \tag{4}$$

4 Experimental Results

We initially used the first simulation experiment to show that an epidemic disease spread within a scale-free social network does have a positive critical threshold if economic resources and transmission costs are taken into consideration—a new conclusion that differs from those reported by previous researchers. The first simulation experiment focused on the universality and generality of the steady density curve and critical threshold when individual economic resources and transmission costs are taken into consideration. Usable economic resources $R(v_i)$ of individual v_i at time t was set at 16 units and transmission cost $c(v_i)$ set at one unit, thus accounting for 6.25% of the individual's total usable economic resources.

We compared the relationship between effective spreading rate λ and steady density ρ for the SIS epidemiological model using three types of complex network platforms: small-world, scale-free without transmission costs, and scale-free with limited individual economic resources and transmission costs. As shown in Figure 2, the eight suites of simulation experiments generated consistent results that did not become contradictory following changes in node and edge numbers. We therefore suggest that the results can be applied to various scale-free networks used to simulate infectious diseases.

The red curves in Figure 2 show that the steady density ρ of the SIS epidemiological model based on scale-free networks reached 0 in a continuous and smooth way when the effective spreading rate λ was decreased, indicating the absence of an critical threshold ($\lambda_c = 0$) in scale-free networks without transmission costs. The blue curves show that infectious diseases do have critical thresholds in small-world networks. If the value of the effective spreading rate exceeds the critical threshold, the infection will spread throughout the network and eventually reach a steady density $\rho(\lambda)$. If $\lambda < \lambda_c$, the infection dies almost immediately. The green curves represent the steady densities $\rho(\lambda)$ of infectious diseases in scale-free networks when individual economic resources and transmission costs are taken into consideration. In addition to being very similar to the blue steady density curves in small-world networks, the green curves have critical thresholds (approximately 0.14). One conclusion drawn from the results of the first simulation experiment is that individual economic resources, transmission costs, and average vertex degrees exert significant influences on epidemic dynamics and critical thresholds in scale-free networks. The same conclusion can be applied to the second and third simulation experiments.

The second simulation experiment was conducted to investigate how the amount of an individual's economic resources $R(v)$ affects the epidemic dynamics and critical thresholds of infectious diseases spread in scale-free networks when the transmission costs are constant. In other words, the second experiment focused on the relationship between the ratio of transmission costs to the total amount of economic resources (hereafter referred to as "the ratio") and critical threshold. To evaluate the influence of the ratio on epidemic dynamics and critical thresholds, we used ten economic resource quantities and assigned the transmission cost $c(v_i)$ of each interaction event as a single unit accounting for 25%, 12.5%, 8.33%, 6.25%, 5%, 4.17%, 3.57%, 3.13%, 2.78% and 2.5% of an individual's economic resources, respectively.

As shown in Figure 3, the critical threshold significantly increased as the ratio grew. For instance, when the resources $R(v_i)$ of individual v_i at time t were designated as 8 units, the critical threshold was approximately 0.22 (Fig. 3, pink curve)—significantly greater than that of a small-world network with the same number of nodes and edges (Fig. 2, blue curve) and same average number of vertex degrees (Fig. 2). The opposite was also true: when the $R(v_i)$ of individual v_i at time t was designated as 40 units, the shape of the steady density curve (Fig. 3, red curve) was very close to that of the scale-free network without transmission costs (Fig. 3, black curve) and the critical threshold was reduced to 0.09. According to Figure 4, a linear correlation existed between the critical threshold and the ratio. Another interesting observation was that the steady density curve grew at a slower rate as the ratio increased (Fig. 3)—that is, the ratio and steady density had a negative linear correlation. One conclusion drawn from the second simulation experiment is that when transmission costs increase or economic resources decrease, the critical thresholds of spreading infectious diseases in scale-free networks grow, while the steady density shrinks according to diffusion rate.

The third simulation experiment was designed to investigate the effects of economic resource distribution on the epidemic dynamics and critical thresholds of infectious diseases spread in scale-free networks—specifically, to determine how different distribution types (delta, uniform, normal, power-law) of economic resources and their statistical distribution parameters (standard deviation in a normal distribution, number of values and range in a uniform distribution) affect the steady density curves and critical thresholds of infectious disease diffusion in scale-free networks marked by limited individual economic resources and transmission costs.

The orange, green, and purple steady density curves in Figures 5 and 8 represent the delta (fixed value of 16), uniform, and normal distributions of individuals' economic resources, respectively; normal distributions are shown in Figures 6 and 9. We found that all had the same critical threshold (≈0.14) and that their steady density curves almost overlapped with each other when the average value of the individuals' economic resources was the same. However, as shown in Figures 7 and 10, if those same economic resources reflected a power-law distribution and no correlation existed between the amount of an individual's economic resources and vertex degrees, the resulting steady density curve (pink) grew more slowly than those of the other three distributions, even though they all had the same critical threshold.

The same experimental results were produced as long as the average value of the individuals' economic resources was the same (Figs. 5 and 8). The steady density curves and critical thresholds were almost the same across different distribution types, regardless of whether the individuals' economic resources obeyed a uniform distribution with a range of 2 or 3 (Figs. 6 and 9, green bars) or a normal distribution with a standard deviation of 2 or 3 (Figs. 6 and 9, purple curves). From the two significantly different groups of steady density curves in Figures 5 and 8 (orange, green and purple versus pink), we conclude that as long as researchers ensure that economic resources do not obey a power-law distribution, they can simply assign each individual's $R(v_i)$ at time t as the average value <r> (Eq. 4) of the statistical distribution derived from the real-world scenario to facilitate their experiments without affecting simulation results.

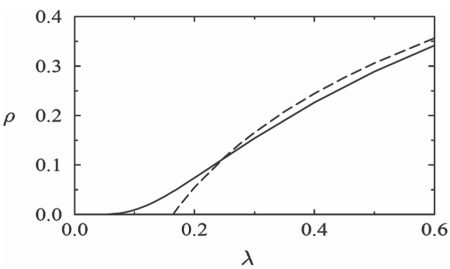


Fig. 1. Prevalence ρ in steady state as a function of effective spreading rate λ. Dashed and solid lines represent BA scale-free and WS small-world networks, respectively [13]

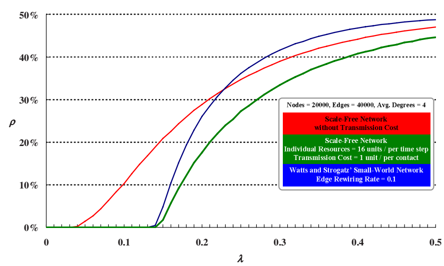


Fig. 2. Relationship between effective spreading rate and steady density

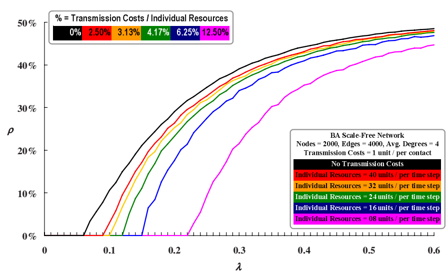


Fig. 3. How the amount of an individual's economic resources affect steady density curves

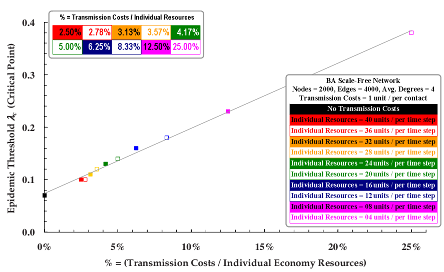
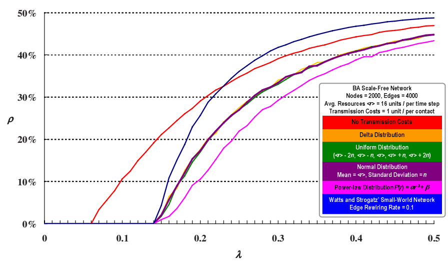


Fig. 4. Relationship between ratio of transmission costs to an individual's economic resources and critical threshold



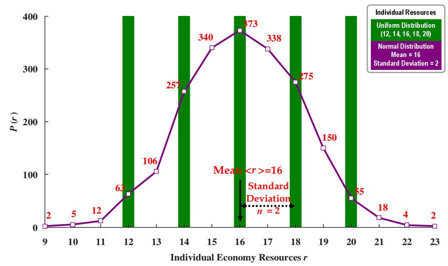
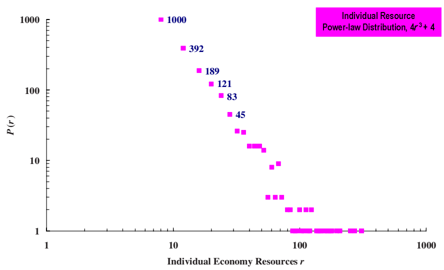


Fig. 5. How different distribution types of individual resources affect steady density curves and critical thresholds

Fig. 6. A uniform and normal distribution of individual resources with average value of 16



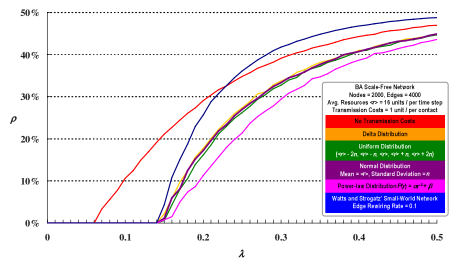
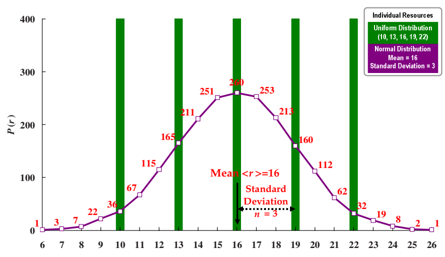


Fig. 7. Individual economic resources in a power-law distribution

Fig. 8. How different types of individual resource distributions affect steady density curves and critical thresholds



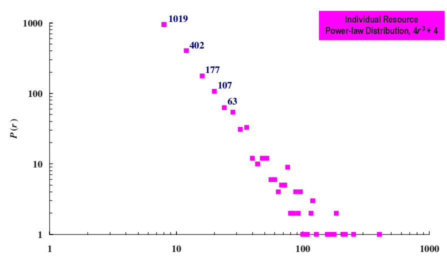


Fig. 9. A uniform and normal distribution of individual resources with average value of 16

Fig. 10. Individual economic resources in a power-law distribution

5 Conclusion

We used agent-oriented modeling and complex networks to construct infectious disease simulation models for the purpose of investigating how economic resources and transmission costs influence epidemic dynamics and thresholds in scale-free networks. Our results indicate that when economic resources and transmission costs are taken into consideration, a critical threshold does in fact exist when infectious diseases are spread within a scale-free network. We suggest that our new conclusions can help epidemiologists and public health professionals understand core questions of disease epidemics, predict epidemic dynamics and diffusion, and develop effective public health policies and immunization strategies.

Acknowledgement

This work was supported in part by National Science Council, Taiwan, Republic of China under grant NSC 96-2221-E-182-041 and in part by Chang Gung Memorial Hospital under Grants CMRPD260021.

References

1. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998)
2. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
3. Dezso, Z., Barabási, A.L.: Halting viruses in scale-free networks. Physical Review E 65, 55103 (2002)
4. Erdös, P., Renyi, A.: On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy of Science 5, 17–60 (1959)
5. Huang, C.Y., Sun, C.T., Lin, H.C.: Influence of local information on social simulations in small-world network models. Journal of Artificial Societies and Social Simulation 8(4) (2005)
6. Liu, Z.H., Lai, Y.C., Ye, N.: Propagation and immunization of infection on general networks with both homogeneous and heterogeneous components. Physical Review E 67, 031911 (2004)
7. May, R.M., Lloyd, A.L.: Infection dynamics on scale-free networks. Physical Review E 64, 066112 (2001)
8. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
9. Pastor-Satorras, R., Vespignani, A.: Immunization of complex networks. Physical Review E 65, 36134 (2001)
10. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Physical Review Letter 86(4), 3200–3203 (2001)
11. Pastor-Satorras, R., Vespignani, A.: Epidemic dynamics and endemic states in complex networks. Physical Review E 63, 066117 (2001)
12. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in finite size scale-free networks. Physical Review E 65, 035108 (2002)
13. Pastor-Satorras, R. , Vespignani, A.: Epidemics and immunization in scale-free networks. In: Bornholdt, S. Schuster, H.G. (eds.), Handbook of Graphs and Networks, Wiley-VCH (2003)

Temporal Difference Learning and Simulated Annealing for Optimal Control: A Case Study

Jinsong Leng, Beulah M. Sathyaraj, and Lakhmi Jain

School of Electrical and Information Engineering,
Knowledge Based Intelligent Engineering Systems Centre,
University of South Australia, Mawson Lakes SA 5095, Australia
Jinsong.Leng@postgrads.unisa.edu.au,
Beulah.Moses@postgrads.unisa.edu.au,
Lakhmi.Jain@unisa.edu.au

Abstract. The trade-off between exploration and exploitation has an important impact on the performance of temporal difference learning. There are several action selection strategies, however, it is unclear which strategy is better. The impact of action selection strategies may depend on the application domains and human factors. This paper presents a modified Sarsa(λ) control algorithm by sampling actions in conjunction with simulated annealing technique. A game of soccer is utilised as the simulation environment, which has a large, dynamic and continuous state space. The empirical results demonstrate that the quality of convergence has been significantly improved by using the simulated annealing approach.

Keywords: temporal difference learning, agent, convergence, simulated annealing.

1 Introduction

An intelligent agent is expected to be capable of adapting and learning in an uncertain environment. The ability of learning can be built up by interacting with the environment in which it is situated. Reinforcement learning is a proper approach for an agent to learn from experience.

Reinforcement learning is the learning of a mapping from situations to actions so as to maximise a scalar reward or reinforcement signal [19], which mainly includes Dynamic Programming (DP), Monte Carlo Method, and Temporal Difference Learning technique (TD). DP [4,8] has been well investigated under the formalism of Markov Decision Processes (MDPs). TD [18] is a form of asynchronous DP, and the value function is estimated by sample episodes. TD method addresses the problem of approximating the optimal action strategy through and while interacting with the environment.

Trial–and–error, delayed rewards, and trade-off between exploration and exploitation are three important features in TD algorithm. As a kind of unsupervised learning, TD has to balance exploration and exploitation during the

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 495–504, 2008.

learning period. To effectively build up the reward structure, the strategy dealing with the trade-off between exploration and exploitation has to be considered. Some strategies have been used for balancing exploration and exploitation, e.g., greedy strategies, randonmised strategies, annealing-like strategies. It is unclear which policy is better and the performance may depend on the task and on human factors. Only few careful comparative studies are available [19].

The simulated annealing (SA) [15] is a technique for solving combinatorial optimisation problems. SA starts with a high temperature and gradually decreases the temperature over time. As a consequence, the agent starts with the high possibility of exploration and turns towards exploitation by incorporating a probability function in accepting or rejecting new solutions. SA approach does not require large computer memory, which can speed up the computation.

In this paper, the simulated annealing technique is utilised for sampling possibility distribution over actions at a given state, so as to improve the quality of performance. A novel algorithm Sarsa(λ) is proposed by combining on-policy learning algorithm with SA method. A comparative study is conducted in a real-time, stochastic, and dynamic testbed called SoccerBots [1]. The simulation results are compared by using ϵ-greedy policy and ϵ-greedy policy with simulated annealing, in conjunction with a linear approximation function called Tile Coding [2]. The experimental results demonstrate that the quality of the performance can be enhanced by using the simulated annealing technique.

The rest of the paper is organised as follows: Section 2 introduces the TD technique and simulated annealing method. The simulation environment and algorithm are detailed in Section 3. The empirical results are analysed and discussed in Section 4. Section 5 presents the related work. Finally, we discuss future work and conclude the paper.

2 Background

2.1 Temporal Difference Learning

TD is a form of model-free approach and is a combination of DP and Monte Carlo Method [18,19]. The idea of TD is to learn how to take action through trial-and-error interactions in a dynamic environment. The value function is accumulated by an incremental learning model and infinite horizon discount model.

The incremental learning model accumulates the value function based on temporal difference errors, as given in Equation (1):

$$V_{t+1} = V_t + \alpha \left[R_{t+1} - V_t\right] \tag{1}$$

where the parameter α is learning rate, $0 \leq \alpha \leq 1$, R_{t+1} is the accumulated reward.

Based on the infinite horizon discount model, the one-step update at a given time t and state s is shown in equation (2).

$$V_{t+1}(s) = V_t(s) + \alpha \left[r_{t+1}(s) + \gamma V_{t+1}(s') - V_t(s)\right] \tag{2}$$

where the parameter γ is the discount rate, $0 \leq \gamma \leq 1$. The rewards in the future are geometrically discounted by the parameter γ, r_{t+1} is the immediate reward.

The use of eligibility traces is another mechanism to improve the convergence of value function [11,18]. The aim of eligibility traces is to assign credit or blame to the eligible states or actions. The ways for evaluating the traces include accumulating traces and replacing traces.

The accumulating traces are given by:

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a), & \text{if } s \neq s_t \\ \gamma\lambda e_{t-1}(s,a) + 1, & \text{if } s = s_t \end{cases} \quad (3)$$

whereas replacing traces use $e_t(s,a) = 1$ for the second update [19].

The trade-off between exploration and exploitation is a problem the learning agent has to deal with. On one hand, the agent must explore the environment explicitly, in case to fall into local optimum. On the other hand, the agent need to exploit the existing knowledge to make the value function converge quickly. However, how can the agent know that the sequence of actions that has been found is the best? It may be helpful to integrate some parametric optimisation techniques with reinforcement learning.

Normally, there are some action selection strategies [19]: e.g., ϵ-greedy policy is to select a random action with probability ϵ and the optimal action with probability 1-ϵ. The drawback is that unlucky sampling may cause the rewards obtained from optimal action to fall into a local optimum. On the another hand, randomised strategy is to select the action according to probability p, which is sampled by Boltzmann distribution. Yet, another strategy is the annealing-like approach, which is used in this paper.

2.2 Simulated Annealing

SA [10] is based on the theory of Markov processes and motivated by the physical process of annealing [15] in solids. The solids are heated to a temperature above the melting point and then allowed to cool. When the solid is cooled to a lower temperature, the solid reaches the stable crystalline structure. If the cooling is done slowly then the crystalline structure is stable. However, if the cooling is done fast then the obtained solid has crystalline imperfection. Therefore, for the system to reach equilibrium, the cooling has to be done slowly. Slower does not mean it can be done forever to obtain stable state. Simulated Annealing is analogous to this process.

The SA algorithm searches the search space similar to the thermodynamic change of energy-to-energy state. First random position in the search space is chosen to be the initial state or current state and another random position is chosen as the next state. The value functions at both the current state and the next state are evaluated. These value functions are compared. If the next state yields a better solution then it is chosen as the best solution. Even otherwise, the next state is chosen with a probability. This probability of acceptance of worse solution is P (accept) which is expressed as given in Equation 4.

$$P(accept) = exp(-c/T) < r \tag{4}$$

Where c = △E is the change in the evaluation function, T is the current temperature and r is a random number between 0 and 1. It is not required to stick on to a particular solution, which could lead to locking up in local optimisation rather than global optimisation. The problem of getting locked to local optimum is overcome in SA by letting the possibility of the choice of accepting the worse solution with a certain probability of P(accept).

The cooling schedule or temperature schedule T of SA is how the temperature is decremented and is the mapping of the temperature to time. It depends on starting temperature, final temperature, decrement in temperature (linear or nonlinear) and number of iterations at each temperature. There are various types of cooling schedules available that are used in practice.

SA Algorithm [16] has been used to solve many combinatorial optimisation problems and some continuous variable problems. The choice of the temperature schedule or the cooling schedule, can be either linear or non-linear.

The algorithm is given in Fig. 1:

```
1.    Current node = MAKE-NODE (INITIAL-RANDOM
        STATE [Problem, Schedule[t]]).
2.    For t = 1 to ∞ do:
3.       If T = 0 then return Current;
4.        Next node = a randomly selected successor of Current;
5.        △E = VALUE[Next] - VALUE[Current]
6.       If △E ≥ 0 then Current = Next;
7.       Else Current = Next only with probability exp(-△E/T) ;
```

Fig. 1. Basic Algorithm of Simulated Annealing

The inner loop of Simulated Annealing algorithm need to be run as long as the search space; where the nodes are present are explored rather than exploring the entire search space. The value of the current and the next solution is evaluated and the difference is △E. When the difference is greater than 0, then the next solution is accepted, else the next solution is accepted with a probability of P(accept) = exp(-$\triangle E/T$), where T is the current temperature.

3 Details of Simulation Environment and Algorithms

TeamBots is a Java-based collection of application programs and Java packages for multiagent mobile robotics research [1]. Each soccer team can have no more than 5 players. Two teams are built and the individual and team strategies for each team are defined.

The scenario is defined to learn a ball interception skill for the soccer player in SoccerBots. For the ball interception, the ball is kicked at a certain angle and speed. The player is away from the ball at a certain distance and angle to ball, in order to intercept the ball with the highest speed and the fewest steps. The intercepting ball problem is that a soccer agent is trained to find the optimal interception direction at each step toward the ball in order to catch an approaching ball in the shortest time.

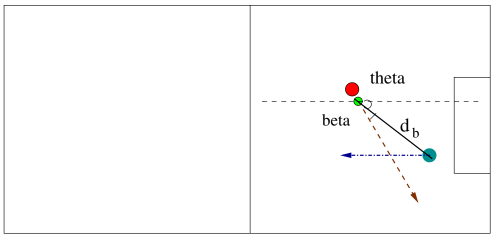


Fig. 2. Ball Interception Problem

In addition, the state space reduction technique is necessary for a large, stochastic, and dynamic environment. In this paper, a linear approximation function known as tile coding [2] is utilised to avoid state space from growing exponentially [12].

The details of modified Sarsa(λ) with replacing traces is given in Fig. 3.

In Fig. 3: α is a learning rate, γ is a discount rate. The action selection policy is the exploration strategy, which can be ϵ-greedy policy, i.e. the agent takes a random action with probability ϵ and takes the best action with probability (1 - ϵ). Q(s,a) is the state-action value, δ and θ are matrices to represent action, feature and related feature weights, and $\mathcal{F}$ is the feature to trace the eligible state process. The parameter T is a temperature, and ψ is the temperature-decay factor. Lower temperatures cause a greater difference in selection probability for actions. In the limit as $T = 0$, the action selection becomes the same as greedy action selection.

4 Empirical Study

SA based exploration works similarly to neighborhood search based exploration by searching the set of all possible actions, but reducing the chance of getting stuck in a poor local optimum by allowing moves to inferior actions to be controlled by a randomised scheme [20]. The parameter T is initially high, allowing many inferior actions to be accepted, and is slowly reduced to a value where inferior actions are nearly always rejected. For the annealing-like approach, the temperature T is set to 120 and temperature-decay factor ψ is set to 0.99. The annealing process is illustrated in Fig. 4.

1. Initialise Q(s,a) arbitrarily and e(s, a) = 0, for all s, a.
2. Repeat (for each episode):
3. Initialise s, a;
4. Initialise T $\leftarrow \psi$ T;
5. Repeat (for each step of episode):
6.1. Choose a_p in s based on action selection policy (e.g., ϵ-greedy);
6.2. Choose a_r in s at random;
6.3. Generate an random possibility P_{random};
6.4. Calculate $Q(s, a_r)$ and $Q(s, a_p)$;
6.5. Calculate the possibility of accept P_{accept} using $exp((Q(s, a_r) - Q(s, a_p))/T)$;
6.6. If $Q(s, a_r) \geq Q(s, a_p)$;
6.7. a $\leftarrow a_r$;
6.8. Else If $P_{random} \leq P_{accept}$;
6.9. a $\leftarrow a_r$;
6.10. Else
6.11. a $\leftarrow a_p$;
7. Take action a, observe r, s';
8. Choose a' from s' using action selection policy derived from Q ;
9. $\delta \leftarrow$ r + γQ(s', a') - Q(s, a);
10. e(s, a) $\leftarrow$ 1;
11. For all s, a:
12. Q(s, a) $\leftarrow$ Q(s, a) + $\alpha\delta$e(s, a);
13. e(s, a) $\leftarrow \gamma\lambda$e(s, a);
14. s $\leftarrow$ s';
15. until s is terminal.

Fig. 3. Sarsa(λ) Control Algorithm with Replacing Traces

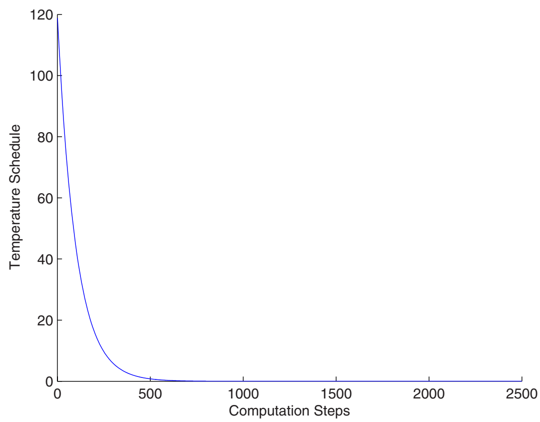


Fig. 4. Temperature Schedule and Computation Steps

The initial start temperature T, the temperature-decay factor ψ, and final temperature T' have to be scaled properly, otherwise, the poor results may occur. In [21], the choice process of their values includes to estimate the mean of the distribution of state values to define a maximum energy scale of the system, its standard deviation to define the maximum temperature scale, and the minimum change in energy to define the minimum-temperature scale.

To compare the realistic performance, the simulation is conducted with different set of parameter values with a number of experiments [13,14]. By setting ϵ to 0.1, the set of optimal parameter values can be found: $\alpha = 0.005$, $\gamma = 0.93$, and $\lambda = 0.9$. The performance may be heavily influenced by the action selection strategy. By running the episodes 2500 times, the average rewards can be obtained using algorithm of Fig. 3 with ϵ-greedy policy. The average rewards can be generated using algorithm in Fig. 3 with annealing-like policy. By comparing the average reward at different episodes, it is clear that the convergence with annealing-like policy is much quicker than that with ϵ-greedy policy.

Table 1. Performance Comparison: Accumulated Average Rewards

Episodes	400	600	800	1000	1200	1400	1600	1800	2000	2200	2400
ϵ-greedy	-28.79	-25.96	-24.25	-23.38	-23.2	-22.6	-22.36	-22.22	-22.07	-21.8	-21.72
SA	-25.98	-23.92	-22.79	-21.96	-21.47	-21.36	-21.07	-20.84	-20.67	-20.59	-20.57

The curves in Fig. 5 illustrate that annealing-like policy can significantly improve the performance, both in speed of convergence and eventually convergence.

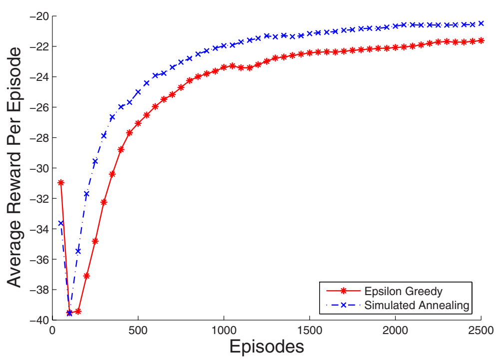


Fig. 5. Convergence Comparison of *epsilon*-Greedy and Simulated Annealing(SA)

5 Related Work

SA is a technique to find the good solution to an optimisation problem by trying random variation of the current solution. SA has been widely applied to optimisation problems [6] by searching for possible solutions and converge at optimal solution. The applications include but are not limited to Travelling

Salesman Problem (TSP), Capacited Vehicle Routing Problems (CVRP), Job Scheduling Problems, Timetabling Problems, and Selection of Communication Protocols/Standards and various other applications [5].

A version of simulated annealing called Adaptive Simulated Annealing is employed with the reinforcement learning algorithm, which shows further improvements in algorithmic convergence properties [3]. A Q-learning algorithm with simulated annealing is introduced to balance exploration and exploitation [7]. In [20], three action selection methods, i.e., neighborhood search based exploration, simulated annealing based exploration, and tabu search based exploration, are evaluated and compared on a discrete reinforcement learning task (robot navigation).

Softmax or Boltzmann distribution Action Selection. A combined use of reinforcement learning and simulated annealing is proposed in [17]. A theoretically established approach tailored to reinforcement learning following Softmax action selection policy are discussed. It has been proven that Boltzmanns formula converges to uniform distribution as T goes to infinity and to the greedy distribution as T goes to 0. In addition, an application example of agent-based routing will also be illustrated.

Another work [3] on adaptive simulated annealing(ASA) based reinforcement learning method is proposed. Here ASA [9] allows far-reaching access of the state space, and permits much faster annealing and hence faster convergence. The action can be generated according to a Boltzmann probability, which provides some kind of "annealing" that is the spirit of the other annealing performed in value function maximisation.

The ϵ-greedy Action Selection with SA. SA based exploration is discussed in [20]. The ϵ-greedy algorithm is a method using near-greedy action selection rule. It behaves greedily (exploitation) most of the time, but every once in a while, say with small probability ϵ (exploration), instead select an action at random. This paper evaluates the role of heuristic based exploration in reinforcement learning. Three methods are compared: neighborhood search based exploration, simulated annealing based exploration, and tabu search based exploration. SA based exploration works by searching the set of all possible actions, but reducing the chance of getting stuck in a poor local optimum by allowing moves to inferior actions. When a non-greedy action is selected, this action is evaluated by SA based exploration approach.

A similar work done in [7] is to explore the possibility of improving the simple ϵ-greedy approach by appropriately reducing ϵ during the learning process. The SA approach is combined with Q(λ). The task of finding the optimal policy in Q-learning is transformed into search for an optimal solution in a combinatorial optimisation problem. Then the Metropolis criterion from SA algorithm is applied to the search procedure in order to control the balance between exploration and exploitation. The improved algorithm is tested in the puzzle simulation domain.

6 Conclusion and Future Work

The parametric optimisation techniques can be combined with TD algorithms to improve the overall performance. This paper provides a comparative study by using the simulated annealing technique to balance between exploration and exploitation. The experimental results demonstrate that the algorithm in Fig. 3 with the annealing-like approach converges much quicker. For a large, stochastic, and dynamic system, utilising the annealing-like technique can reduce computational cost and learn quickly.

Future work includes the comparison of softmax action selection and simulated annealing technique. The ultimate goal is to develop a methodology for adaptively selecting the parameter values in the learning algorithm. In addition, the algorithm will be extended to soccer agents teaming to solve the cooperative learning problems.

References

1. Teambots (2000), http://www.cs.cmu.edu/~trb/Teambots/Domains/SoccerBots
2. Albus, J.S.: A Theory of Cerebellar Function. Mathematical Biosciences 10, 25–61 (1971)
3. Atiya, A.F., Parlos, A.G., Ingber, L.: A Reinforcement Learning Method Based on Adaptive Simulated Annealing. In: Proceedings of the 46th IEEE International Midwest Symposium on, pp. 121–124 (2003)
4. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
5. Chaharsooghi, S.K., Jafari, N.: A Simulated Annealing Approach for Product Mix Decisions. Scientia Iranica 14(3), 230–235 (2007)
6. Dowsland, K.A.: Simulated Annealing. In: Modern Heuristic Techniques for Combinatorial Problems (1995)
7. Guo, M., Liu, Y., Malec, J.: A New Q-learning Algorithm Based on the Metropolis Criterion. Systems, Man and Cybernetics, Part B, IEEE Transactions on 34(5), 2140–2143 (2004)
8. Howard, R.A.: Dynamic Programming and Markov Processes. MIT Press, Cambridge (1960)
9. Ingber, L.: Very Fast Simulated Re-annealing. Mathematical Computer Modelling 12(8), 967–973 (1989)
10. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science 220(4598), 671–680 (1983)
11. Klopf, A.H.: Brain Function and Adaptive Systems–A Heterostatic Theory. Technical report, AFCRL–72–0164, Air Force Cambridge Research Laboratories, Bedford, MA (1972)
12. Leng, J., Fyfe, C., Jain, L.: Reinforcement Learning of Competitive Skills with Soccer Agents. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, Springer, Heidelberg (2007)
13. Leng, J., Jain, L., Fyfe, C.: Simulation and Reinforcement Learning with Soccer Agents. Journal of Multiagent and Grid systems, IOS Press, The Netherlands 4(4) (to be published, 2008)
14. Leng, J., Jain, L., Fyfe, C.: Convergence Analysis on Approximate Reinforcement Learning. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 85–91. Springer, Heidelberg (2007)

15. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equation of State Calculations by Fast Computing Machines. J. Chem. Phys. 21, 1087–1092 (1953)
16. Russel, S., Norwig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs (2003)
17. Stefán, P., Monostori, L.: On the relationship between learning capability and the boltzmann-formula. In: Monostori, L., Váncza, J., Ali, M. (eds.) IEA/AIE 2001. LNCS (LNAI), vol. 2070, pp. 227–236. Springer, Heidelberg (2001)
18. Sutton, R.S.: Learning to Predict by the Method of Temporal Differences. Machine Learning 3, 9–44 (1988)
19. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
20. Vien, N.A., Viet, N.H., Lee, S., Chung, T.: Heuristic Search Based Exploration in Reinforcement Learning. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 110–118. Springer, Heidelberg (2007)
21. White, S.R.: Concepts of scale in simulated annealing. In: AIP Conference Proceedings, vol. 122, pp. 261–270 (1984)

Separating Learning as an Aspect in Malaca Agents

M. Amor, L. Fuentes, and J.A. Valenzuela

Dpto. Lenguajes y Ciencias de la Computación
Universidad de Málaga, Málaga
{pinilla, lff, valenzuela}@lcc.uma.es

Abstract. Current OO frameworks provided with MAS development toolkits provide core abstractions to implement the agent behavior, by using the typical OO specialisation mechanisms. However, these OO designs do not provide proper abstractions to modularize other extra-functional concerns (e.g. learning property), which are normally intermingled with the agent functionality (*tangled code* problem), and spread over different classes or components (*crosscutting concerns* problem). This means that the reusability of the agent architectural components is drastically reduced, so agents are difficult to maintain, extend or adapt. Aspect-oriented technologies overcome these problems by modeling such concerns as *aspects*. This work proposes to separate and modularize the *learning* of software agents following the aspect-oriented solution of the Malaca model. By decoupling the agent functional behavior from the protocol that carries out the learning activities; the development, adaptation and evolution of intelligent agents is substantially improved.

1 Introduction

Some classical learning methods of the Artificial Intelligence are often used in Multi-Agent Systems (MAS) development, to improve their ability to make independent decisions in the execution of some complex tasks (e.g. electronic commerce). A software agent may have the capability to learn and adapt its behaviour from certain events, like feedback from its own actions or interaction with other agents. Agents normally observe and learn from the environment, user actions or agent interactions, to make some deductions that enhance the agent knowledge about the user and the environment in general. Thus agents have to be endowed with appropriate mechanisms to learn from, and adapt to, their environment [1]. This is normally achieved by introducing Machine Learning (ML) algorithms to agent design [2].

However, the incorporation of ML techniques to the agent behaviour is not a trivial task. The problem is that OO frameworks for MAS development are designed centred on the agent domain specific functionality and fail to provide a suitable modularization of other agent properties such as coordination, distribution or learning. Such extra-functional properties of the agent are normally scattered across different classes (*crosscutting concerns problem*), and are intermingled with other basic functionalities (*tangled code problem*), drastically reducing the reusability of the agent architectural components.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 505–515, 2008.

The Malaca agent model [3,4] overcomes these problems by combining CBSD (Component-Based Software Development) and AOSD (Aspect-Oriented Software Development) [5] technologies. Following the CBSD principles, Malaca breaks the domain specific functionality of the agent into completely independent *components*, facilitating their addition or substitution and making the resulting agent more adaptable. Furthermore, to achieve better functional decomposition of the software agent, those features that appear (*cut-across*) in several components are modeled as *aspects* according to current AOSD practice. The main distinctive features of AOSD are that software *base modules* (e.g. objects or components) do not contain any reference or code related to crosscutting concerns (e.g. coordination between agents, learning, etc.), and that each software module permits the injection of crosscutting concerns at specific points, called *join points*, of their execution flows (e.g. after the reception of an input message, or the execution of a method).

Thus the Malaca agent functionality (both basic and domain-specific) is provided by reusable software components and decoupled from the agent extra-functional properties, encapsulated by the aspects. This work proposes to separate and modularize the learning property of software agents following the aspect-oriented solution of the Malaca model. By decoupling the agent functional behavior from the protocol that carries out the learning activities, the development, adaptation and evolution of intelligent agents is substantially improved. By encapsulating a learning algorithm as an aspect is possible to add or delete the learning property to an agent simply by modifying the composition rules of components and aspects, even at runtime.

Such composition rules are defined as part of the MaDL [6] language, a component-aspect oriented architecture description language based on XML, for the configuration of software agents. The developer only has to provide an explicit description of the components and aspects to be plugged into the agent architecture; aspect-component composition information (weaving in AOSD terminology), and other information related with MAS deployment. Note that in Malaca the weaving of components and aspects is performed at runtime, so is possible to adapt an agent to react to new events even at runtime. In order to facilitate the agent development we provide the Malaca Agent Description tool (MAD) to introduce agent descriptions without having to manipulate XML files directly.

The rest of the paper is organized as follows: The next section introduces the case study Reviewing System and how the learning property is typically modeled in an agent (using Jade [7]). Section 3 shows the Malaca agent model and how the learning property is modeled in Malaca as an aspect. In section 4, we demonstrate that the development, adaptation and evolution of Malaca agents is substantially improved modeling learning as an aspect by providing some cohesion, coupling and separation of concerns metrics. We compare the Jade and the Malaca implementations of the proposed case study, and discuss the results.

2 Modelling the Learning Property in a Software Agent

Agent learning can be considered as the way an agent orders its experience in time [8], normally represented by either internal or external events such as feedback from its own actions (either successful or failed), interaction with external world, and cooperation with other agents. Mainly there are two approaches for applying learning to MAS: single-agent learning algorithms, or distributed learning of MAS. We will focus on the former one that consists on improving the agent capacities by applying existing learning algorithms more or less directly to (single) agents in a MAS setting. Single-agent learning neither affects nor is affected by other system agents, at least not in a coordinated or conscious way [1]. We will illustrate these concepts through a case study described next. An implementation in Jade is first showed, for later comparison with our proposal

2.1 A Case Study: The Reviewing System

Normally MAS developers use OO agent toolkits as Jade [7]. We will illustrate the benefits of separating learning as an aspect with respect to current agent toolkits through a Reviewing System case study described in detail in [9,10]. This work presents an implementation of the Reviewing System in AspectT, an agent framework implemented in AspectJ[1] the most well-known aspect-oriented language for Eclipse. This approach separates different agent properties, the learning among others, in different Java classes. The resulting agent is better modularized, since extra-functional agent properties, like interaction, autonomy, learning or mobility, normally intermingled in different classes, can evolve independently. In spite of the use of aspects, one of the main drawbacks is that the user, normally an expert on Artificial Intelligence and not necessarily on aspect-oriented programming (AOP) [5], has to be familiar with the AspectJ language to be able to develop (intelligent) agents. An open challenge is to provide a development environment that facilitates the agent development for non-experts neither in aspect orientation nor in any other advance software development technology.

The Reviewing System in [9,10] models the users involved in the review process as two types of agents: a *Chair* agent, that represents and assists the Program Chair, and several *Reviewer* agents, that represent and assist the review committee members. Firstly, the Chair agent proposes Reviewer agents a list of papers to review, considering their research interests. Then, Reviewer agents accept or reject to review each of the proposed paper. Then, the Chair agent will assign or not a paper to a reviewer depending on their answers.

2.2 The Reviewing System in Jade

In this section we will show a possible implementation of this system using the Jade toolkit. Figure 1 shows the class diagram and part of the review process code of the Chair agent. The *ReviewerPreferences* structure stores the areas of expertise for each Reviewer agent. The *EvaluateReviewerResponse* class encapsulates the reception and processing of Reviewer agents messages notifying the Chair agent their answer to a paper review proposal. This class includes the code for verifying the reviewer answer (*verifyReviewerResponse* method) and also some code for either assigning the paper, or proposing it to a different reviewer; depending on the reviewer answer (*updateReviewAssignment* method of the *ChairAgent* class).

[1] AspectJ language is an Eclipse project, http://www.eclipse.org/aspectj/

Likewise [9], in order to improve the paper assignment process, the agent uses the Least Mean Squares (LMS) [2] to learn about the reviewer preferences. This is an indirect learning since the Chair agent updates its knowledge about the reviewer preferences according to the information received during its interaction with these agents. But, these agents learned about these preferences from the users they represent using also ML techniques (e.g. Temporal Difference Learning (TD-Learning)). Therefore, the *EvaluateReviewerResponse* class, which is mainly part of the basic agent interaction protocol, also undertakes some learning tasks.

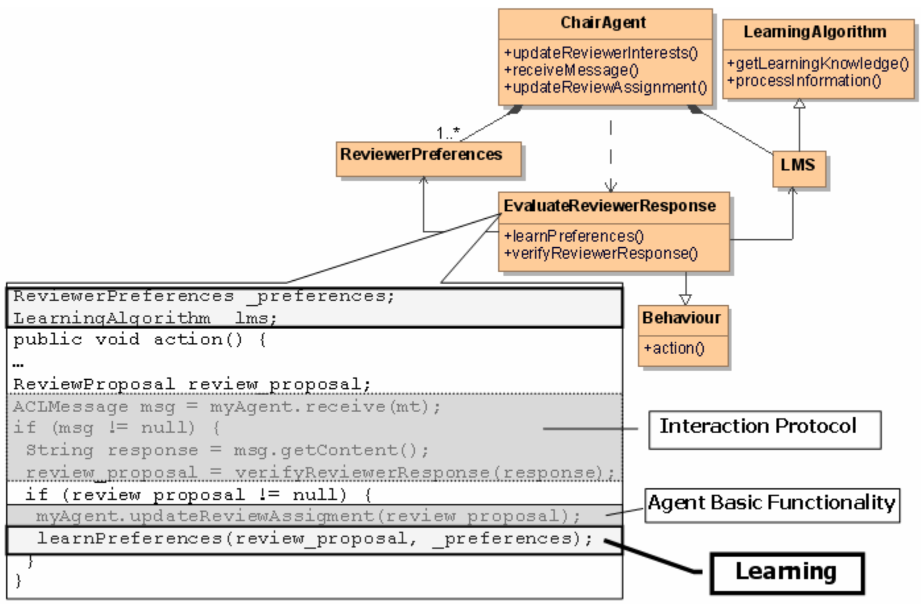


Fig. 1. Partial class diagram of the Chair Agent in Jade

Figure 1 shows that the code of different agent properties is tangled inside the implementation of the *action()* method that has to extend any Jade agent class. According to Jade OO framework we have to include under this method, implementation code for the interaction protocol, an invocation of a review process method that is part of the agent domain specific functionality; and data structures (*ReviewerPreferences* and *LearningAlgorithm*) and a local method related to the learning property (*learnPreferences* method). This method initiates the extraction of information to carry out the learning task and performs the invocation of the learning algorithm encapsulated in the *LMS* class. We can easily appreciate that this Jade design is difficult to reuse and maintain. For example, delete the learning process means inspect this code looking for references to components and data structures related to it, make the accurate changes and recompile the source code of the agent. Likewise, reuse the learning property or perform any change on it, also requires manual code inspection, and recompilation. This is an error-prone task, since entails a very detailed understanding of the implementation code.

3 Learning as an Aspect in Malaca

3.1 The Malaca Agent Model

In this section we will describe in brief the main characteristics of the FIPA-compliant Malaca agent model (a detailed description can be found in [3,4,6]). The architecture of a Malaca agent combines CBSD and AOSD technologies, being the first order entities components and aspects. In Malaca the domain specific functionality of the agent will be provided by components and any other extra-functional property (e.g. interaction protocol) will be modeled as aspects, which promotes the independent evolution of agent functionality and properties.

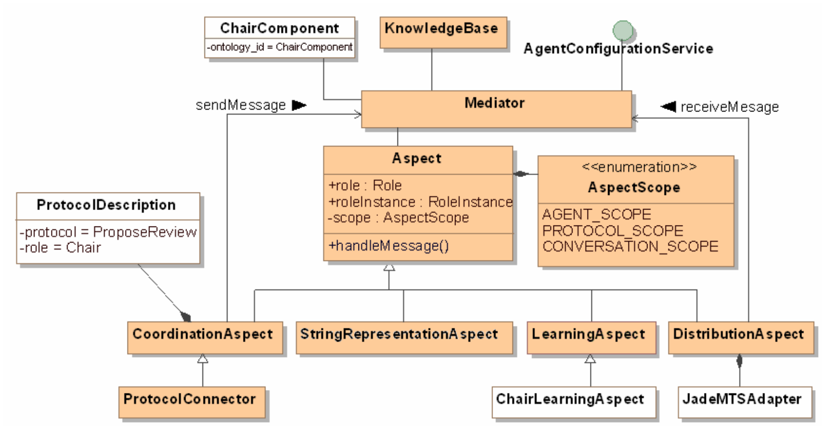


Fig. 2. The Component and Aspect-oriented Model of the Malaca Agent Architecture

Figure 2 shows the internal architecture of the Chair agent that complies with the Malaca agent model. The classes in grey are provided by the Malaca agent framework. The others are application specific, but some of them were (re)used as we will show below. The agent domain specific functionality is provided by independent software components, which can be in-house components, COTS (*Commercial Off-The-Shelf*) components or even Web services. So, for the Chair agent, all the functionality that characterise a program chair is provided by a single component (*Chair-Component*). Additionally, the Malaca framework provides other base components, like a *KnowledgeBase* component that encapsulates the agent knowledge, in our example, the expertise areas of each reviewer.

In the Malaca model the aspect is the entity used to modularize any agent property (i.e. crosscutting concern). A generic aspect entity (*Aspect*) is extended to model any aspect not yet considered by the current implementation. According to AOSD practices, aspects intercept component behavior (a *join point*) and execute an additional behavior not considered in the component (the *advice*). In our model aspects intercept mainly messages, so any property related with the message exchange will be modeled as an aspect. The aspect behavior is codified in the *handleMessage* (either input or output), and others not shown for sake of brevity. But how do aspects know when, what and how they have to intercept the behavior of a component? By means of component and aspect composition rules described below. These rules refer to components using the ontology identifier, and aspects using the *role* and *roleInstance* name. The role of an aspect is then an identifier that describes the property it represents –or the *role* it fulfils in the communication process. The *scope* specifies the application scope of an aspect within an agent. For example: any communication message (AGENT_SCOPE); all messages of an interaction protocol regardless of the conversation identifier (PROTOCOL_SCOPE), the set of messages interchanged in a single conversation (CONVERSATION_SCOPE).

According to the Agent Communication Model defined by FIPA the agent interaction entails (subclasses of *Aspect*, Figure 2): an interaction protocol between agents (our *CoordinationAspect* aspect), the use of a common representation format for ACL messages (the *ACLRepresentationAspect* aspect) and the distribution of aspects through an agent platform (the *DistributionAspect* aspect). The separate modeling of such agent properties promotes the independent reuse in different applications. In our case study the representation format of ACL messages in String and the implementation of the distribution adaptor for Jade were (re)used.

The responsibility of the coordination aspect is twofold. On one hand it controls the different conversations of the agent according to an interaction protocol and on the other, this aspect schedules agent actions implemented as part of software components. A distinctive feature of the Malaca agent architecture is that supported interaction protocols are not hard-coded; instead, they are explicitly represented in ProtDL, which is an XML-based protocol description language that is part of the MaDL language. After launching a Malaca agent the *ProtocolDescription* class is instantiated with the information stored in a ProtDL file (*ProposeReview* in Figure 2). So is possible that an agent reacts to different list of messages only by modifying the ProtDL description of the interaction protocols, which increases its adaptability.

Component and aspect entities are composed using aspect composition rules, or *pointcuts* in AOSD terminology. These rules describe for each component behaviour being intercepted, the aspects that will be applied and in which order. For example, the arrival of a message is intercepted by the aspect weaver (in Malaca called *Mediator*) which initiates the execution of aspects according to the composition rules specified in MaDL. This means that the Mediator is able to dynamically compose components and aspects at runtime, which enhance the agent adaptability. For each input/output message it consults the composition rule(s) and initiates the execution of the corresponding aspects (if there any). The order in which each aspect appears in the rule also determines the order in which the aspects are composed.

The Malaca model defines two join points corresponding to the reception (RECV_MSG) and sending (SEND_MSG) of ACL messages from a software agent perspective. Since the agent model defines two different interception points, we need only two rules to describe aspect composition: one rule describes how to compose aspects when an incoming message is intercepted, while the other is used to describe aspect composition for outgoing message interception. For each aspect is specified if its execution is critical or not. For critical aspects a failed execution means that the delivery/reception of the corresponding message also fails.

A distinctive feature of the Malaca Aspect Model is that the information for composing aspects is defined outside them, facilitating their reuse in different agent architectural configurations. Note also that the composition between components and aspects is not hard coded; it is defined in XML which has many benefits.

3.2 The Learning Aspect in Malaca

In this section we will show how to design the Learning as an aspect for a Malaca agent. In general, when designing the learning aspect we have to decide what information triggers the agent learning, when this information is collected, when we have to apply a learning algorithm that processes this information, and the adaptation of the current agent knowledge.

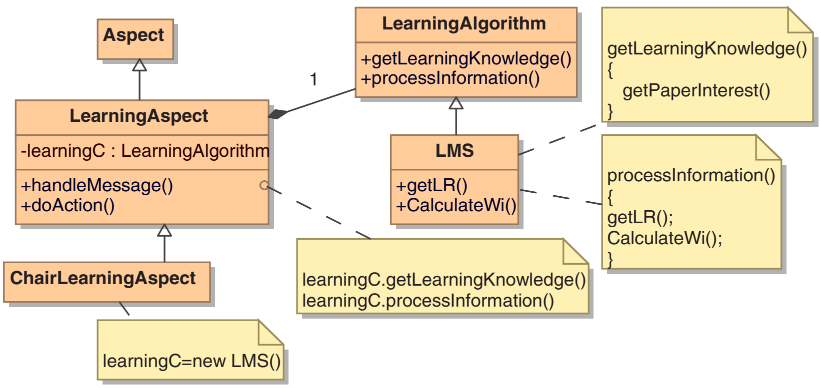


Fig. 3. Learning Aspect Design

As the learning aspect is applied during message interchange, the information that triggers the agent learning will be part of input/output messages, which will be processed before or after the sending or receiving of a message. The aspect behaviour will consist of applying the learning algorithm to the information included in the intercepted messages, and then modifying the agent's knowledge base according to the algorithm output. The learning aspect will be implemented in the *LearningAspect* class, which extends the *Aspect* class and includes a component implementing the learning algorithm. We follow the *Strategy* design pattern [11] that allows keeping a family of different algorithms that implement the learning techniques, so as a client class can choose one. The *LearningAspect* and *LMS* subclasses implement the specific learning algorithms (see Figure 3). Concretely the *Chair* agent uses the *LMS* class that implements the extraction of the information from received messages and the invocation of the LMS learning algorithm. Figure 4 shows how the Mediator weaves the learning aspect.

In Malaca, the incorporation of a new aspect involves to modify the XML document containing the agent configuration in MaDL [6], by only adding the aspect description and updating the aspect composition rules. Figure 5 shows a snapshot of

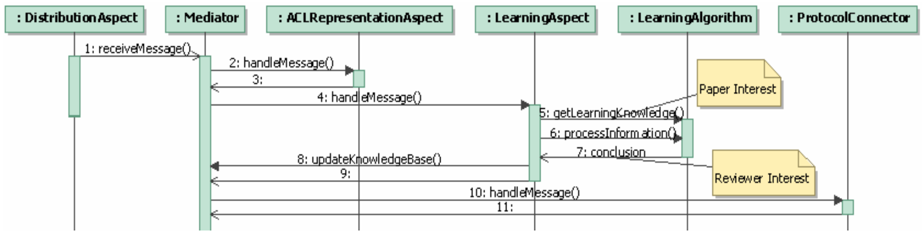


Fig. 4. Learning as an aspect in Malaca

the MAD[2] Editor used to introduce a description of the learning aspect (*ChairLearningAspect*) in MaDL, as part of the *Chair* agent architecture.

Role and role instance names are respectively a global and an instance identifier of the learning aspect. Since learning is applied to any kind of message, is configured

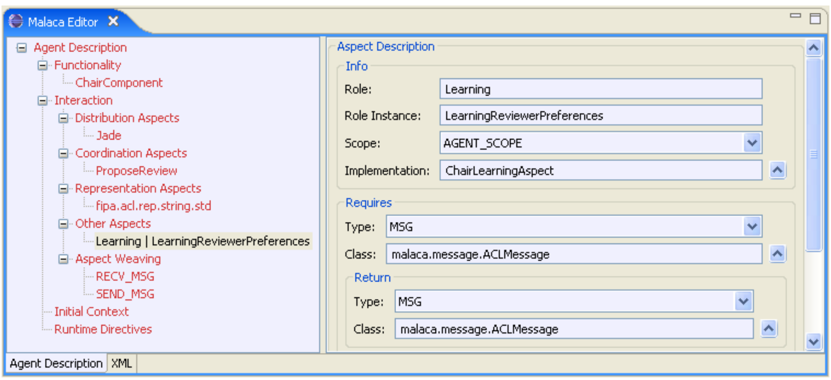


Fig. 5. Learning aspect description in MaDL editor.

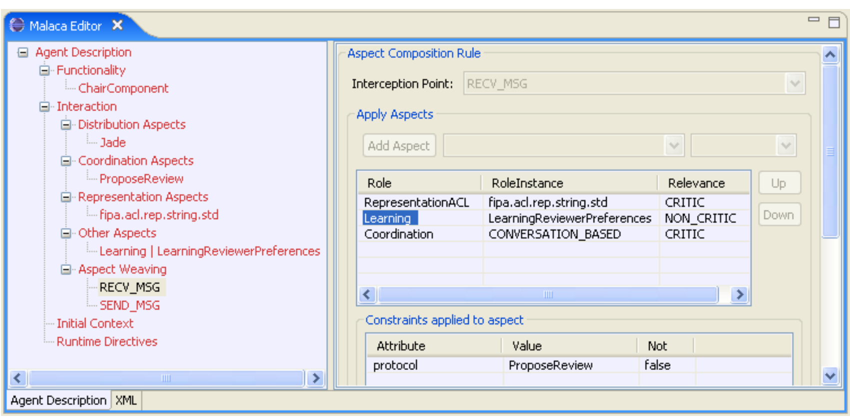


Fig. 6. Composition rule in the reception of a message with the learning aspect

[2] MAD editor is an Eclipse plug-in (http://caosd.lcc.uma.es/malacaTools.htm)

with an *AGENT_SCOPE* scope (i.e. only one instance per agent). The implementation field is bound to the *ChairLearningAspect* class. The signature of the aspect interface is described by means of an input ACL message (*Requires*) and the return is also an ACL message (*Return*). Note that the agent learning is specified independently of the rest of agent properties, making easier agent programming for non-experts, because normally there is no need for writing new code, since (re)use is notably promoted.

To complete the aspect addition is necessary to modify the aspect composition rules to include the application of the learning aspect when the agent receives a reply message from a reviewer. In our example, we have to modify only the composition rule for incoming messages, as shows Figure 6. The new rule says that the learning aspect will be applied to input messages after the application of the representation aspect and before the application of the coordination aspect. This description is completed by specifying that this rule application is constrained to input messages belonging to the *ProposeReview* protocol (this restriction is checked by consulting the *protocol* field of the received message). The incorporation of a new aspect can also be carried out through the agent configuration service of the Mediator (interface *AgentConfigurationService* in Figure 2), that allows to modify the agent architecture at runtime giving the same information as this.

4 Evaluation and Conclusions

The separation of the learning property as an aspect provides valuable benefits at the development of software agents. These benefits can be measured by the augmentation of the cohesion and the reduction of the coupling of the elements of the agent internal architecture, which implies a better architectural design. Both coupling and cohesion can be measured in terms of a metric suite [12,13], which helps us to compare the Jade and the Malaca agent and illustrates the benefits of the latter.

The metric suite focuses strictly on the evaluation of software architecture artefacts, which have a direct impact on architecture modularity. The suite includes metrics for architectural separation of concerns (SoC), architectural coupling, component cohesion, and interface complexity. For the shake of brevity we will focus on the evaluation of cohesion and coupling of the elements that represent learning in both agents (*EvaluateReviewerResponse* in Jade and learning aspect in Malaca), although the complete evaluation and comparison should be extended to every element of both architectures. The cohesion metric computes each component's semantic cohesion by means of a metric named Lack of Concern-based Cohesion (LCC). This metric counts the number of concerns (or agent properties and functions) addressed by the assessed component. The higher the number of different agent properties in the component the lower the cohesion is. he coupling metrics measure the number of components connected to each other, and is quantified by the architectural fan-in and fan-out of every component. The former counts the number of conventional components which require services from the assessed component, while the latter counts the number of components from which the assessed component requires services. A high fan-in and/or fan-out expose a non-desirable tight coupling. Table 1 shows the results of these measurements.

Table 1. Cohesion and coupling measures

	Cohesion	Coupling	
Component	**LCC**	**Fan-in**	**Fan-out**
JADE EvaluateReviewerResponse	4 I O(n)	1 (static composition)	2 I O(m)
MALACA Learning Aspect	1	0 (dynamic weaving)	0

The code included in Figure 1, which corresponds to the Jade agent, shows that the class *EvaluateReviewerResponse* includes functions (methods and invocations to other classes related to learning) required to learn reviewer preferences, but it also includes functions associated to other agent properties (interaction protocol that is related with message distribution and formatting concerns and domain specific functionality). This means a LCC result of four, or in general O(n), an increasing linear function of the number of concerns addressed by this class (or classes with a direct reference to this). This matter is an evidence of a low level of cohesion of this class, which hampers its reuse and upgrading. However, the LCC for the Malaca learning aspect is always 1. This is because all functions related to learning are always encapsulated in one aspect, which does not include code concerning other agent properties. This means that the cohesion of the aspects inside the Malaca agent architecture is high, which facilitates (re)use and evolution of internals components.

The level of coupling among the architectural elements also affects to the reuse and maintenance of the architecture. The fan-in of *EvaluateReviewerResponse* class is one referring to the invocation of the *action()* method by the *ChairAgent*. In Malaca the learning aspect (as any aspect) only interacts with the *Mediator* during the weaving process. The dynamic composition performed by the Mediator is a powerful feature that decides at runtime the aspects and components that are composed facilitating changes. This is a very important difference with Jade implementation, since changing any issue related to learning in Jade requires modifying and recompiling the code of different classes. Inside the Jade agent the composition is static: Class composition is hard coded by means of method invocation using direct references. Regarding fan-out, the *EvaluateReviewerResponse* class has direct references to different classes related to learning included in the *ChairAgent* class. In current implementation this value is two that can be generalised to O(m), where m increases depending on the complexity of the learning strategy. In Malaca, the learning aspect does not invoke services in any component, so the fan-out is always zero, independently of the complexity of the learning strategy.

References

1. Alonso, E., et al.: Learning in Multi-Agent Systems. The Knowledge Engineering Review 16(3), 277–284 (2001)
2. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
3. Amor, M., Fuentes, L., y Troya, J.M.: Training Compositional Agents in Negotiation Protocols Using Ontologies. J. Integrated Computer-Aided Engineering 11(2) (2004)

4. Amor, M., Pinto, M., Fuentes, M.L., y Troya, J.M.: Building Software Agents from Software Components. In: Conejo, R., Urretavizcaya, M., Pérez-de-la-Cruz, J.-L. (eds.) CA-EPIA/TTIA 2003. LNCS (LNAI), vol. 3040, pp. 222–231. Springer, Heidelberg (2004)
5. Aspect-Oriented Software Development, `http://www.aosd.net`
6. Amor, M., Fuentes, L.: A Description Language for Component and Aspect-Oriented Agent Architectures, en Multiagent Systems and Software Architecture (MASSA 2006)
7. Java Agent DEvelopment Framework, `http://jade.tilab.com/`
8. Clancey, W.: Conceptual Coordination: How the Mind Orders Experience in Time. Lawrence Erlbaum Associates, New Jersey (1999)
9. Garcia, A., et al.: The Learning Aspect Pattern. In: Proceedings of Pattern Languages of Programs (PLoP 2004), Illinios, USA (2004)
10. Garcia, A., Lucena, C.: Taming Heterogeneous Agent Architectures with Aspects. Communications of the ACM, July 2006 (accepted for publication)
11. Gamma, E., et al.: Design Patterns Elements of Reusable Object-Oriented Software. Addison-Wesley editors, Reading (1995)
12. Sant'Anna, et al.: On the Quantitative Assessment of Modular Multi-Agent System Architectures. NetObjectDays (MASSA) (2006)
13. Sant'Anna, C., et al.: General Architecture Evaluation Process/ Metrics. AOSD-Europe Deliverable D85, AOSD-Europe-ULANC-25 (June 2007)

Learning Semantic Web from E-Tourism

Waralak V. Siricharoen

Computer Science Department, School of Science,
University of the Thai Chamber of Commerce, 126/1 Dindeang,
Bangkok, Thailand
waralak_von@utcc.ac.th, lak_waralak@yahoo.com

Abstract. Inquiring an information system in natural language is engaging in the special tourism domain because users usually have very different backgrounds regarding their computer literacy. The constant fast growth in travel related information measure makes it increasingly difficult to find, organize, access and maintain the information required by consumers (users). The Semantic Web provides enhanced information access based on the exploitation of machine-processable meta-data.E-tourism is a perfect candidate for Semantic Web because it is information-based and depends on the World Wide Web, both as a means of marketing and transaction channel. E-tourism/e-travel software adapted from original e-commerce, ready for creating instantly online reservation/booking. The Semantic Web relies heavily on the formal ontologies that structure underlying data for the purpose of comprehensive and transportable machine understanding. Therefore, the success of the Semantic Web depends strongly on the proliferation of ontologies. Ontologies can assist organization, browsing, parametric search, and in general, more intelligent access to online information and services. The ideas proposed in this paper, are particularly interested in the new possibilities afforded by Semantic Web technology in the area of knowledge management practical to the travel industry. This paper also discusses some ontological trends that support the growing domain of online tourism. The preview of e-tourism is introduced in general. The paper also gives the example concepts of existing e-tourism using ontologies display in graphical model presented in ontologies editor tool called Protégé and show the example of e-tourism ontologies description in OWL and RDFS syntax. The last part of the paper is a summary on the e-tourism ontologies projects.

Keywords: Ontologies, E-commerce, E-tourism, Semantic Web, Travel, Tourism.

1 Introduction

Tourism is an information based business. Tourists have to leave their daily environment for consuming the product. At the beginning of the 21st century, the structure of demand and supply in the tourism industry is undergoing significant changes. Social and economic changes, for instance age profile, life styles and organization of work, together with the fast distribution of the Internet, increasing e-business and the availability of online public services have had a strong impact on the demand for tourism

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 516–525, 2008.

products and their modes of provision [1]. The competitiveness and success of the tourism industry depend on the impact of social, economic and technological changes within society as a whole. Key among these global trends is the increase in the use of electronic services [2]. Today, tourism service suppliers are a group of many different users, most of them offering electronic services.

Tourism has become the world's largest trade and its expansion shows a constant year-to-year raise. Kim addresses in the paper [3] that "Competitive benefit is no longer ordinary, but increasingly driven by science, information technology and innovation. The Internet is already the major source of tourist destination information for travelers". This is support that the tourism business mission is to revolutionaries the traditional tourism industry to next generation e-tourism powered by Semantic Web technology. It will be realized by an advanced e-tourism Semantic Web portal which will connect the customers and virtual travel agents from anywhere at anytime with any needs and requests [4]. Tourism Information Systems are a new form of business systems that provide and support e-tourism and e-travel organizations, such as airlines, hoteliers, car rental companies, leisure suppliers, and travel agencies. One class of these systems relies on travel related information sources, such as Web sites, to create tourism products and services. ICTs enable tourism businesses to make tourism products and services directly available to a large number of consumers, and to interact with them as well as with other tourism producers and distributors [5].

Semantic Web will bring the revolution to this area by not only exponentially extending the dissemination and exchange channels with unlimited access, unlimited time and unlimited locations, but also assisting users with smart information searching, integrating, recommending and various intelligent services[6]. The Semantic Web will also be crucial to the development of Web applications such as e-commerce, providing users with much more sophisticated searching and browsing capabilities as well as support from intelligent agents such as shopbots (shopping "robots" that access vendor Web sites, compare prices etc.). Examples of the use of ontologies/taxonomies to support searching and browsing can already be seen at e.g., Yahoo Shopping and Amazon.com. Therefore, e-tourism is the one of the decent application areas for Semantic Web technologies and it is also a good test-bed to prove the efficiency and utility of Semantic Web technologies [7]. The e-tourism ontology provides a way of viewing the world of tourism. It organizes tourism related information and concepts. The ontology will allow achieving interoperability through the use of a shared vocabulary and meanings for terms with respect to other terms [8].

2 Semantic Web and E-Tourism

Overall, e-tourism comprises electronic services, which include three basic service categories: Information services (e.g. destination and hotel information); Communication services (e.g. email, discussion); Transaction services (e.g. reservation and booking, payment). These services are typically offered via the Internet and are accessible from a variety of locations, from private PCs at home or at work to electronic kiosks and other devices in public places. The World Wide Web (WWW) as known it today is a huge collection of information or information superhighway. The number of websites on the WWW is growing daily. However, this expands of information is not as good as to an increase of functionality: information extraction has become a difficult task. Current technologies do not provide means to reuse existing information efficiently. Common search engines can perform keyword-based searches. Finding a certain part of information is very difficult. However, the number of results usually is enormous and not manageable by the human reader. For the human user it is simply impossible to go through all the websites that are delivered as results to a query. Objective of the Semantic Web is to make the information on the Web understandable and useful to computer applications in addition to humans. The Semantic Web promises a solution to this problem: Semantically annotated websites can not only be understood by the human reader but also by machines. Enriching websites with machine-readable semantics will enable more intelligent and efficient searching and further processing of data without requiring the human user to interfere [9]. Ontologies define data models in terms of classes, subclasses, and properties. For instance, we might define a carnivore to be a subclass of animals. In Fig. 1. shows very simple example ontology for animals [10].

```
class-def animal                          % animals are a class

class-def plant                           % plants are a class
    subclass-of NOT animal                % of things that are not animals

class-def carnivore                       % carnivores are a class
    subclass-of animal                    % which is a subclass of animals
    slot-constraint eats
        value-type animal                 % that eat animals
```

Fig. 1. Example of animal ontology [10]

The challenge to develop software package [9] for online commerce is to find a solution to cope and integrate the non-standard way of defining e-tourism products and services. There are no standards or common criteria to express transportation vehicles, leisure activities, and weather conditions when planning for a vacation package, several ways can be found among all the existing Web sites. The information from several travel, leisure, and transportation online sites, it point out the lack of standards in the tourism domain. Some of the differences founded among several sites are the following example. The price of tourism related activities and services are expressed in many different currencies (Thai baht, euros, dollars, British pounds, etc.). To finding a solution to improve on this lack of standards in the tourism field by automatically understanding the different ways of expressing tourism products and services, extracting its relevant information and structuring. The sophisticated technologies, such as semantics and ontologies, are good candidates to enable the development of dynamic information systems [11]. Ontology can be constructed for e-tourism. Tourism is a data rich domain. Data is stored in many hundreds of data sources and many of these sources need to be used in concert during the development of tourism information systems. The e-tourism ontology provides a way of viewing the world of tourism. It organizes tourism related information and concepts. The e-tourism ontology provides a way to achieve integration and interoperability through the use of a shared vocabulary and meanings for terms with respect to other terms. The e-tourism ontology was developed using OWL (Web Ontology Language). OWL was proposed by the W3C for publishing and sharing data, and automating data understanding by computers using ontologies on the Web [6]. Ontology is mentioned by Tom Gruber which used to refer to “an explicit specification of a conceptualization [of a domain]. In other words, ontology refers to a formalization of the knowledge in the domain. Ontology is the concept which is separately identified by domain users, and used in a self-contained way to communicate information. Combination of concept is the knowledge base or knowledge network. Some of the reasons why someone want to develop an ontology are to share common understanding of the structure of information among people or software agents, to enable reuse of domain knowledge, to make domain assumptions explicit, to separate domain knowledge from the operational knowledge, to analyze domain knowledge [12]. An ontology structure holds definitions of concepts, binary relationship between concepts and attributes. Relationships may be symmetric, transitive and have an inverse. A minimum and maximum cardinality constraint for relations and attributes may be specifies. Concepts and relationships can be arranged in two distinct generalization hierarchies [13]. Concepts, relationship types and attribute abstract from concrete objects or value and thus describe the schema (the ontology) on the other hand concrete objects populate the concepts, concrete values instantiate the attributes of these objects and concrete relationship instantiate relationships. Three types of relationship that may be used between classes: generalization, association, and aggregation [5].

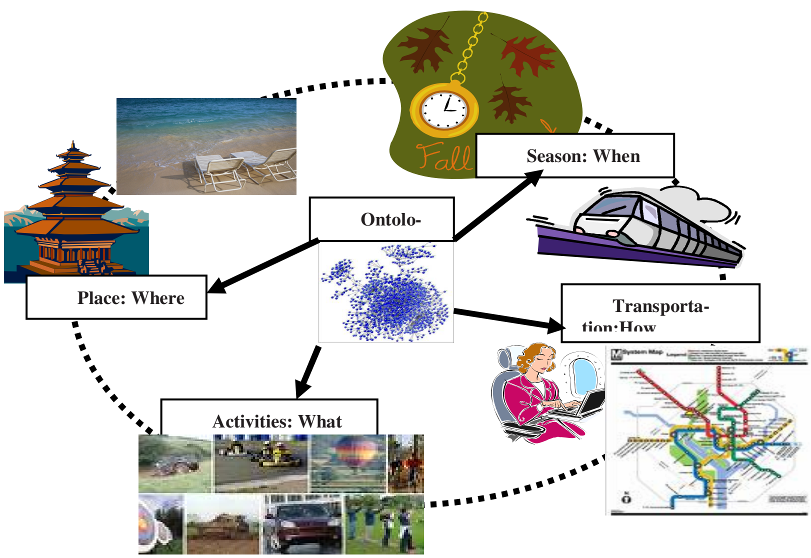


Fig. 2. The e-tourism ontologies explanation

Ontology is able to answers four types of questions that can be asked when developing the e-tourism package. These questions involve the predicates What, Where, When, and How. Examples of typical questions are in Fig. 2. [4]: What can a tourist do while staying?, Where are located the interesting places to see and visit, When can the tourist visit a particular place? , How can the tourist get to its destination to see or do an activity?.

Classes are the focus of most ontologies [12]. A class can have subclasses that represent concepts that are more specific than the superclass. Slots describe properties of classes and instances. Ontologies for the Semantic Web are characterized as Resource Description Framework (RDF) ontologies, and are being built using OWL and other languages based on RDF. This ontology growth in direct support for RDF and various species of OWL has created some controversy [14]. Shared ontologies allow for different systems to come to a common understanding of the semantics of learning concepts. The present the required ontology model including the formal expression of ontology, mapping to Extensible Markup Language (XML) representation and the corresponding system architecture for binding web services. Two important technologies for developing the Semantic Web are XML and RDF. XML is accepted as a standard for data interchange on the Web. It is a language for semi-structured data.

RDF uses XML and it is at the base of the semantic Web. It is a metadata language for representing information in the Web and provides a model for describing and creating relationships between resources. Resource Description Framework Schema (RDFS) is technologically advanced compared to RDF. RDFS allows users to define resources with classes, properties, and values. A class is a structure of similar things and inheritance is allowed. RDFS property can be viewed as an attribute of a class. RDFS properties may inherit from other properties, and domain and range constraints can be applied to focus their use. In OWL an ontology is a set of: Classes, Properties, Constraints on the classes and properties. OWL adds a layer of expressive power to RDFS. Uses formal logic to describe the semantics of classes and properties. Build on top of RDF and RDFS. Over the years a vast amount of research has been carried on how to represent and reason about knowledge. In Europe funding has been heavily concentrated on the development of OIL (Ontology Inference Layer), a language for defining ontologies. In the US, DARPA funded a somewhat similar project called DAML (Distributed Agent Markup Language). More recently these activities have been combined into a project to work on a merged ontology language, DAML+OIL. In late 2001 the W3C set up a working group called WebOnt to define an ontology language for the Web, based on DAML+OIL. All of these ontology languages aim to provide developers with a way to formally define a shared conceptualization of a domain. They encompass both a means of representing the domain and a means of reasoning about that representation, typically by means of a formal logic. In the case of DAML+OIL this is Description Logic [10]. Jakkilinki [15] provide an overview of the development methodology and applications for tourism ontologies. Ontologies are created using ontology development tools, such as Protégé [16]. A Java-based ontology editor with OWL Plugin: that means that it allows ontology implementation as an applet on the Web. This permits multiple users to share the ontology. The W3C has recently finalized the OWL as the standard format in which ontologies are represented online. With OWL it is possible to implement a semantic description of the travel domain by specifying its concepts and the relationships between the concepts.

Usually, there is several tourism ontologies were considered for reuse, before considering built the new ontology. In e-tourism different ontologies have been developed for different areas. However, sometimes in different countries or regions around the world, the existing ontologies might not meets the needs to describe regional distinctions for any specific areas. An international standard is the thesaurus on Tourism & Leisure Activities of the World Tourism Organization (WTO). It is a very extensive collection of terms related to the area of tourism. The example of e-tourism ontology developed by DERI [17], the project started with a list of terms that should be included in the ontology. On the one hand it was helpful to have a voluminous collection of terms, and on the other hand it was misleading because the broad range of terms sometimes led to too detailed concepts, which had to be taken out in a later stage of the development [17]. Ontologies enhance the semantics by providing richer relationships between the terms of a vocabulary.

Ontologies can serve as point of reference to map against. The intelligent agents and more to come each day could then make suggestions on consumers; make arrangements in consideration of consumer preferences. For these agents, the Semantic Web infrastructure would be based on core travel ontologies that would be published on fixed URI's (Universal Resource Indicators) as OWL files. Ontologies would allow these providers to publish metadata about their travel services and contact information. The following code is showing the example of two different e-tourism ontologies presented in the same Ontology Editor Program called Protégé [16]. And also below shows example of OWL ontologies description in e-tourism ontology. It show that there are classes *Accommodations, Guestroom*, etc. which the *Accommodations* classes have object property called *hasRoom*. These classes and properties are quite normal for tourism/travel domain.

```
<owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="Accomodation">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="hasRoom"/>
        </owl:onProperty>
        <owl:someValuesFrom>
          <owl:Class rdf:ID="Guestroom"/>
        </owl:someValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf
```

Above shows the other example of RDFS ontologies description concerning the tourism system. This ontology's description also represented in the next part shows the example of RDF ontologies description. It show that there are classes *Address, AirbusPlane, Airport, BedandBreakfast* etc. which the *AirbusPlane* classes is the subclass of class *Plane*. And class *Bed andBreakfast* the subclass of class *LodgingFacility*. It show that class *LodgingFacility* in this ontology is similar to class *Accommodations* in the previous ontology shown.

```
<rdf:RDF xmlns:rdf="&rdf;" xmlns:a="&a;">
<a:Class rdf:about="http://www.ina.fr#AbstractEntity"/>
<a:Class rdf:about="http://www.ina.fr#Address">
<a:subClassOf
   rdf:resource="http://www.ina.fr#ComplexDatatype"/></
   a:Class>
<a:Class rdf:about="http://www.ina.fr#AirbusPlane">
<a:subClassOf rdf:resource="http://www.ina.fr#Plane"/>
</a:Class>
<a:Class rdf:about="http://www.ina.fr#Airport">
<a:subClassOf
   rdf:resource="http://www.ina.fr#TransportFacility"/>
</a:Class><a:Class
   rdf:about="http://www.ina.fr#Artefact">
</a:Class>
```

3 Summary of E-Tourism Ontologies.

3.1 Harmonise Project

Harmonise Project is funded by European Commission as part of the 5th Framework Programme. This project aims at building a technological infrastructure base on a shared ontology, to enhance the cooperation of European SMEs in tourism sector. The interoperability problem in tourism can be resolved by putting emphasis on the combination of a social consensus process with the application of new technologies. Harmonise project is a base for another European project HARMO-TEN (Tourism Harmonisation Trans-European Network eTEN), which aims to create an electronic space for tourism stakeholders where all businesses in the marketplaces should be enabled to exchange their information in a seamless, semiautomatic manner, independent from geographical, linguistic and technological boundaries.

3.2 HI-TOUCH

The aim is to develop software tools to be used by travel agency sales assistants for providing a tourist prospect with the best-adapted offer. Assistance tools are based on ITC new tools (XML, Java, Flash, ontological data bases and semantic descriptors, multi-lingual thesaurus). One of the work-packages includes development of semantic database: multi-lingual thesaurus, inventory of tourism resources and implementation of authoring tools on the semantic database [18].

3.3 SATINE

Semantic-based Interoperability Infrastructure for Integrating Web Service Platforms to Peer-to-Peer Networks. It aim is to realize a semantic based infrastructure which will allow the Web services on well-established service registries like UDDI or ebXML to seamlessly interoperate with Web services on P2P Networks. The travel ontologies to be developed and the semantics of the Web services will be based on standard specifications like the one produced by Open Travel Alliance[19].

3.4 IM@GINE IT

Intelligent Mobility AGents, Advanced Positioning and Mapping Technologies INtEgration Interoperable MulTimodal, location based services. IM@GINE IT project aims to develop one and single access point, through which the end user can obtain location-based, intermodal transport information, mapping and routing, navigation and other related services everywhere in Europe, anytime, taking into account personal preferences of the user. One of the milestones is the development of common transport and tourism ontologies for semantic web applications [20]. These ontologies shown in table 1 and the project described in the previous part is the only small proposed ontologies of e-tourism in the market. There are many of them available for anyone to download and reuse it.

4 Conclusions

Ontologies and ontology-based information retrieval have the potential to significantly improve the process of searching information on the World Wide Web. Concept search and browsing can ease the trouble of searching the web using keyword-based techniques. This is especially important in information-based business, such as e-tourism. The travel industry is facing rapid changes with the advent of the Semantic Web technologies. There is now the need for developing an infrastructure to manage the online travel information and deliver to consumers what they want. New superior consumer services can be deployed such as travel market overview and price comparison. Ontologies will play an important role as they promise a shared and common understanding of traveling concepts that reaches across people and application systems [6]. Semantic Web technology pave the way to enhanced Knowledge Management (KM) solutions that are based on semantically related knowledge pieces of varying granularity. Although early semantic web-based KM approaches and solutions have shown the benefits of ontologies and related methods, there still exist a large number of open research issues that have to be addressed in order to make semantic web technologies fully effective when applied to KM solutions. Delivering the Semantic Web to the travel industry depends upon: (1) the syntactical and semantic mark-up of travel content; (2) the development of better knowledge analysis and modeling tools; (3) widespread adoption of interoperable knowledge representation languages, and (4) the construction of suitable ontologies.

Ontologies are becoming increasingly important as a component of online commerce offerings. They are useful (and arguably necessary) in supporting at least navigation, browsing, user expectation setting, and parametric search. Sources of class taxonomies exist, tools for piecing ontologies together are growing, and some sources of parameter information are becoming available. Challenges remain for users in reusing available ontological information, because as standards are still forming, most vocabulary information needs to be augmented, and although some tools exist, most are still on a development path to becoming complete tool suites suitable for mass deployment. These challenges are surmountable and they should reduce over a short time [21]. The main strategies to gain an accomplishment are accelerating and increasing electronic information, building tourists' and entrepreneurs' trust and confidence in e-commerce, developing e-commerce capacity, marketing via electronic devices, and promoting the development of e-marketplace for tourism.

The area of tourism especially e-tourism seems to a perfect application area for Semantic Web technologies. As information dissemination and exchange are the key backbones of travel industry, the Semantic Web can considerably improve e-tourism. Thai entrepreneurs have to adapt themselves well to suit the e-tourism environment such as serving niche markets, becoming subsidiaries of large enterprises to handle domestic business or niche products, and developing web sites to connect directly with their customers. Support from the government will accelerate the e-commerce development. The travel industry is facing rapid changes with the advent of the Semantic Web technologies. There is now the need for developing an infrastructure to manage the online travel information and deliver to consumers what they want. New superior consumer services can be deployed such as travel market overview and price comparison. Ontologies will play an important role as they promise a shared and common understanding of traveling concepts that reaches across people and application systems. The Semantic Web is certainly to become a global database, and if its development is evolutionary and distributed, then there are issues of accessibility, trust and credibility. Not all data sources will have universal access, so there needs to be a robust and extensible security model. Not all data sources will be equally reliable. If instead of just returning an answer to a query a Semantic Web application could also attach a proof of how that answer was derived, then the querying application could potentially do some reasoning about how 'believable' that fact is.

References

1. Jentzsch, A.: Tourism Standards. XML Clearinghouse, http://www.xmlclearinghouse.de/reports/Tourism%20Standards.pdf
2. Millard, J.: The nature of e-tourism services, Prisma Strategic (2003), http://www.prisma-eu.net
3. Kim, C.: E-tourism: An Innovative Approach for the Small and Medium-sized Tourism Enterprises (SMITEs) in Korea (2000), http://www.oecd.org/dataoecd/56/13/34268048.pdf
4. Cardoso, J.: E-tourism: Creating Dynamic Packages using Semantic Web Processes (2005), http://www.w3.org/2005/04/FSWS/Submissions/16/paper.html
5. Vongdoiwang, W., Batanov, D.N.: Similarities and Differences between Ontologies and Object Model. In: CCCT 2005 Conference Proceeding (2005)
6. Siricharoen, W.V.: Using Ontologies for E-tourism. In: The 4th WSEAS / IASME International Conference on Engineering Education (EE 2007) Proceeding (2007)
7. Cardoso, J., Sheth, A.: Semantic Web Services, Processes and Applications. Springer, Heidelberg (2006), http://www.dme.uma.pt/jcardoso/Books/Semantic-Web-Services-Processes-and-Applications/Slides/Ch01.ppt
8. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. In: Knowledge Engineering Review (1996)
9. Siorpaes, K.: OnTour System Design (2004), http://e-tourism.deri.at/ont/docu2004/OnTour%20-%20System%20Design.pdf

10. Hewlett-Packard Development Company: Introduction to Semantic Web Technologies (2007), http://www.hpl.hp.com/semweb/sw-technology.htm
11. Awad, E., Ghaziri, H.: Knowledge Management. Pearson Education, London (2004)
12. Klein, M., Fensel, D., Harmelen, F.V., Horrocks, I.: The Relation between Ontologies and Schema-languages: Translating OIL-specifications in XML-Schema (2000), http://www.cs.vu.nl/~mcaklein/papers/oil-xmls.pdf
13. Volz, R., Oberle, D., Studer, R.: Views for light-weight web ontologies. In: Proceeding of SAC 2003, Melbourne, Florida, USA (2003)
14. McGuinness, L., Harmelen, F.V.: OWL Web Ontology Language Overview (2004), http://www.w3.org/TR/owl-features/
15. Jakkilinki, R., Sharda, N., Ahmad, I.: Ontology-based Intelligent Tourism Information Systems: An overview of development methodology and applications. In: Proceeding of TES 2005: Tourism Enterprise Strategies: Thriving – and Surviving – in an Online Era, Melbourne, Australia (2005)
16. Protégé: what is protégé-owl? Stanford Medical Informatics (2007), http://protege.stanford.edu/overview/protege-owl.html
17. DERI: E-Tourism Working Group. DERI International (2005)
18. Hi-Touch Working Group. Semantic Web methodologies and tools for intra-European sustainable tourism (2003), http://www.mondeca.com/articleJITT-hitouch-legrand.pdf
19. Flügge, M.: SATINE: Semantic-based Interoperability Infrastructure for the Tourism Industry (2006), http://www.fokus.fraunhofer.de/projekte/Satine/Satine-2-06.pdf
20. Bekiaris, E.: Factsheet ICT for Transport and The Environment Center for research and technology Hellas/ Hellenic Institute of Transport (Certh/HIT) (2004), http://ec.europa.eu/information_society/activities/esafety/doc/esafety_library/projects_fact_sheets/imagine_it.pdf
21. McGuinness, L.: Ontologies and Online Commerce. IEEE Intelligent Systems 16(1), 8–14 (2001)

Incremental Biomedical Ontology Change Management through Learning Agents

Arash Shaban-Nejad and Volker Haarslev

Department of Computer Science and Software Engineering, Concordia University,
H3G1M8 Montreal, Quebec, Canada
{arash_sh, haarslev}@cs.concordia.ca

Abstract. Biomedical knowledge bases and ontologies constantly evolve to update the knowledge in the domain of interest. One problem in current change management methodologies is the over-reliance on human factors. Despite the advantages of human intervention in the process of ontology maintenance, including a relative increase of the overall rationality of the system, it does not guarantee reproducible results of a change. To overcome this issue, we propose using intelligent agents to discover and learn patterns for different changes and their consequences. In this paper, we present a novel multi-agent-based approach, to manage the evolving structure of biomedical ontologies. This framework aims to assist and guide ontology engineers through the change management process in general, and aids in tracking and representing the changes, particularly through the use of category theory. It provides an efficient way to automatically capture, validate, and implement a change.

Keywords: Bio-Ontologies, Multi-Agent, Learning, Change Management, Category Theory.

1 Introduction

Bioinformatics is a challenging domain in knowledge management. Biological data are highly dynamic, and the large biomedical knowledge sources contain complexly interrelated elements, with various levels of interpretation. With advances in life science, many features and functionalities must be added to or removed from existing knowledge bases in the biomedical domain. Bio- ontologies and controlled vocabularies evolve continuously to provide meaningful and valid information. Due to complexity and heterogeneity of their structure, changing one component can have wide-ranging, unpredictable effects. Different versions of an ontology may respond differently to queries, so a system based on frequently changing ontologies may yield responses of questionable quality. The automatic representation, validation and reproduction of changes with minimal human intervention raise challenging issues in ontology evolution. Many bio-ontologies are known to be seriously defective, from both terminological and ontological perspectives. Our approach is based on the RLR framework for employing agents in an integrated, ontology-driven infrastructure to capture the pattern of changes and validate the results. To represent changes in ontology-driven biomedical applications and formalize interactions between agents, we

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 526–535, 2008.

also propose the use of category theory as a mathematical notation, independent of any specific ontology language or implementation. We have chosen category theory as the main formalism in our framework, because it proved itself as an efficient vehicle to examine the process of structural change in living and evolving systems [30]. As an experiment, we have employed the RLR framework to manage the evolving structure of the Skin disease Ontology (SKDON). Due to ambiguities in medical mycology from the lack of a standard and consistent vocabularies with comprehensible semantic, there is a need [1] for a standard set of vocabularies to support dermatological practice and enhance the accuracy of clinical knowledge management. The SKin-Disease ONtology (SKDON) is an integrated formal ontology based on OWL-DL, and aims to provide a shared and common understanding of the concepts in the domain of medical mycology for dermatologists and other researchers working on life science. In our work, we have concentrated on disorders of the skin and related tissues, such as hair and nail due to fungi. We have designed the SKDON as an essential part of the FungalWeb semantic web infrastructure [2] to gather, retrieve and manage consistent laboratorial and clinical information for patient care. SKDON is created from several distributed resources, including structured/unstructured texts, online databases, and existing controlled vocabularies, such as MeSH [3], ICD-9 (http://www.cdc.gov/nchs/icd9.htm), SNOMED (http://www.snomed.org/) and Disease database (http://www.diseasesdatabase.com/). Cross referencing between the FungalWeb ontology, SKDON and MeSH "Chemicals & Drugs" category provides valuable information about the disease, the involved fungus and the drugs prescribed. Change in any of the resources can alter the definitions in the target ontology.

2 The RLR Framework

The RLR framework aims to Represent, Legitimate, and Reproduce the changes and their effects (Figure 1). It helps to capture, track, represent, and manage the changes in a formal and consistent way, enabling the system to generate reproducible results.

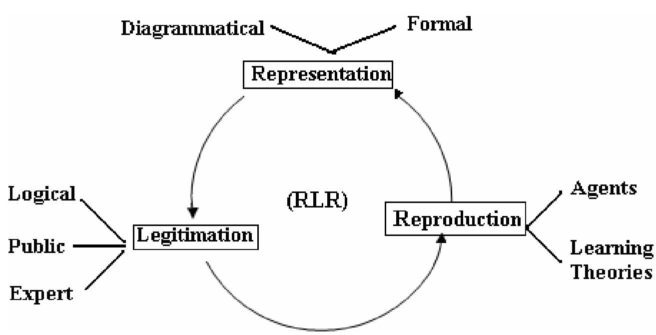


Fig. 1. The RLR framework

In this framework, various ontological changes can be represented in either formal or diagrammatical ways. Each change will be legitimated and validated logically, then approved publicly and by experts. To reproduce the results of changes and automate the change management process, agents are recruited to learn change patterns and their consequences.

2.1 Representation

This phase is responsible for consistently updating the representations of new knowledge. Many of the problems in ontology evolution are basically problems about the nature and representation of change. For the formal representation of changes, we use Description Logics, and for diagrammatical representation, we employ a method based on discrete state model and category theory [4].

2.2 Legitimation

Legitimation, in our context, is defined as the verification of the legitimacy and consistency of a change in the domain of interest. This phase assesses the impact of a potential change before the change is actually made. Experts and logical reasoners should study a change based on its consistency with the whole design, in various degrees of granularity. Then, the final approval is needed from end-users. Logical legitimation can be obtained by a reasoning agent.

2.3 Reproduction

Over-reliance on human factors is a problem in current change management methodologies. Despite the advantages of in maintenance, including higher rationality, human intervention does not guarantee the reproducibility of results of a change [29]. To overcome this issue, we propose using intelligent agents that discover patterns for different changes and their consequences.

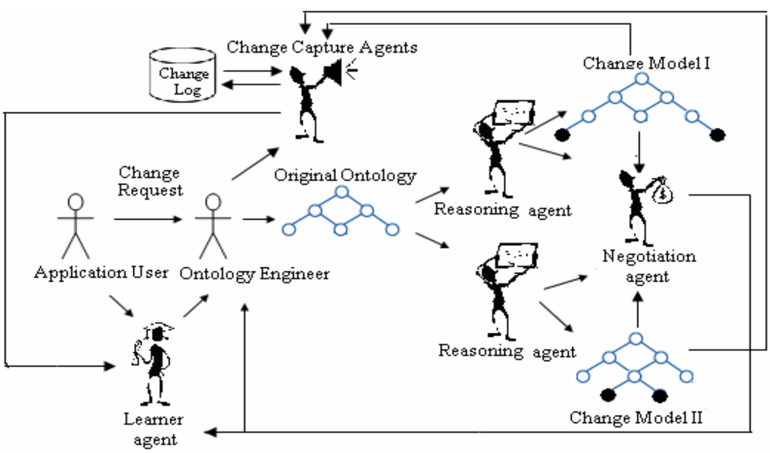


Fig. 2. The change management process using agents

3 Agents and Pattern of Change

Intelligent agents have the ability to find, identify, and collect desired information from multiple resources about various actions under changing conditions [5]. Agents are able to work rationally in order to capture changes in dynamic and heterogeneous environments, and to respond properly to these changes [6]. In the RLR framework, we have used four types of agents: Change capture agents, Learner agents, Reasoning agents, and Negotiation agents. Figure 2 demonstrates the interactions between them.

3.1 Change Capture Agents

This agent family is responsible for discovering, capturing, and tracking the changes in ontology, by processing one or more change logs. They detect real-world alterations and report them as new facts with which to update the knowledge base of an agent. Changes can occur on a random or scheduled basis. The change capture agents act like triggers in a database. We have defined the following three different types of change-capture agents:

- **Action Control Agents (ACA):** The action control agents consist of user activities and legal operations, which together capture changes such as deletion, insertion, and updates to ontology elements, and can store all the data related to different types of change in change logs.
- **Explorer Agents (EA):** The explorer agents capture changes by processing and reading change logs in parallel, in a specified time range. By logically determining transactions, the explorer agents generate the appropriate messages for the corresponding services.
- **Log-Reading Agents (LRA):** The log-reading agents read the log files in a specified time period. This information will be passed on to a learning agent in order to create patterns for different changes. Later, the information can be used to Undo or Redo a change.

Together, these agents determine which ontological elements have changed. To capture ontological changes, we also use annotation properties such as: Timestamps, Version and Status on ontological elements.

3.2 Learner Agent

As an application is used and evolves over time, the change logs can accumulate invaluable data and information about various types of changes. A learner agent can use these historical records of changes that occur over and over in a change process to derive a pattern. After several changes, possibly from various releases, it would be feasible to estimate the rate and direction of future changes for a system by generating rules or models. In RLR, the reasoner and negotiation agents can change the generated rules, and send modifications to the learning agent. The learning agent starts with limited, uncertain knowledge of the domain, and tries to improve itself, relying on adaptive learning based on semantics provided by the ontological backbone. The adaptive learner agent plays an important role in the reproduction phase, where we look for patterns to bootstrap the process of change management. The discovery of temporal patterns for event-based data is addressed by P.S. Kam, et al. [7], while Höppner tackled the problem with the discovery of informative temporal rules for defining temporal patterns in [8]. Learning rules for discovering temporal patterns is described by L. Sacchi, et al. [9, 10]. The RLR-Learning agent uses the Sacchi's algorithm [9] for extracting temporal rules to learn patterns of evolving ontological data [9].

3.3 Reasoning Agent

A reasoning agent is a software agent that controls and verifies the logical validity of a system, revealing inconsistencies, hidden dependencies, redundancies, and misclassifications. It automatically notifies users or other agents when new information about the system becomes available. We use RACER [11] as a description logic reasoner agent, along with other semi-formal reasoners in the RLR framework.

When the agent is faced with a change, it ought to revise its conceptualization [12] based on the new input by reasoning about the consistency of the change using both prior and new knowledge. Several attempts [13, 31] have been made, to provide reasoning services for category-based systems. We also use a semi-automated reasoning system for basic category-theoretic reasoning based on a first-order sequent calculus [14]. It captures the basic categorical constructors, functors, and natural transformations, and provides services to check consistency, semantic coherency, and inferencing [14].

3.4 Negotiation Agent

Negotiation happens when agents with conflicting interests desire to cooperate [15]. In the RLR framework, the negotiation agent acts as a mediator allowing the ontology engineer and other autonomous agents to negotiate the proper implementation of a specific change while maximizing the benefits and minimizing the loss caused by such change. A human expert may then browse the results, propose actions and decide whether to confirm, delete, or modify the proposals, in accordance with the intention of the application.

In our framework, negotiation is defined based on the conceptual model of argumentation [16]. In this context, an argument is described as a piece of information that allows an agent to support and justify its negotiation stance or influence that of another agent [15, 17]. Employing argumentation to analyze belief revision with the intention of updating an agent's knowledge also has been studied by M. Capobianco [24], based on dialectical databases.

4 Category Theory for Representing Agent Interactions

Category theory is a new domain of mathematics, introduced in 1945 [18]. Using categories, one can recognize certain regularities in distinguishing a variety of objects, capture and compose their interactions, differentiate equivalent interactions, identify patterns of interacting objects and extract some invariants in their action, and decompose a complex object into its basic components [19]. Categorical notations consist of diagrams with arrows. Each arrow $f: X \rightarrow Y$ represents a function. Representation of a category can be formalized using the notion of diagram.

As presented in [4], category theory has great potential as a mathematical medium to represent, track, and analyze changes in biomedical ontologies. In addition we demonstrated [4] its capability to analyze some of the common operations during ontology evolution. After describing the ontological concepts within categories representing a modular hierarchy of domain knowledge, we have employed category theory as a formalism to analyze ontological changes and agent interaction in different stages of the RLR framework.

Agents perform actions in a context by using rules. The change of the rules is a main adaptation principle [20] for learning in our framework. For describing our adaptive agents we follow G. Resconi's formalization method [20]. Each rule includes a finite or infinite semantic unity symbolized as S_1, IN, P_1 and OUT, respectively representing the input statement, the domain of the rule, the rule, and the range of the rule (denoting the value of an agent's action). Generally, when working in a static environment, we deal with only one family of rules for each context, but when the environment is dynamic, it is very likely that these rules change into other rules. Therefore, a single change in an ontological element triggers other changes in rules and contexts. As an agent gradually learns the different rules for various contexts, a communication channel between these rules and between different agents is necessary. Such changes are demonstrated in [20] as follows (Figure 3).

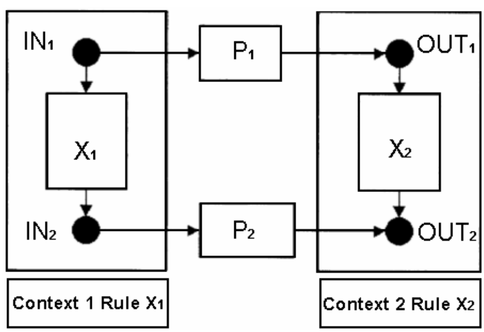


Fig. 3. Demonstration of the semantic unity of the changes of the rule X_1 in the context 1 into the rule X_2 in the context 2 (adapted from [20])

Agent interactions can be simulated by categories. We have used category theory formalism, along with General Systems Logical Theory (GSLT) explained in [21], to describe agent communication. For example, the communication between different semantic unities as shown in [20] can be represented as follows:

Category theory can also be used for modeling agent interaction protocols [22] yielding, a practical image of adaptive learning agents, their semantic unities and adaptation channels [20].

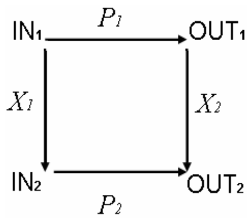


Fig. 4. Categorical representation that demonstrates how rules *P1* and *P2* enable the transformation of the rule *X1* into the rule *X2* [20]

5 Application Scenario

Since skin disorders have been historically categorized by appearance rather than scientific and systematic facts [23], the existing taxonomy must be modified to update the ontological truth. Many terms in current medical mycology vocabularies describing skin disorders originate as verbal descriptions of appearance, foods, people, mythological and religious texts, geographical places, and acronyms [24]. Many names and terms are highly dependent on individual or regional preferences [24], causing redundancy, vagueness, and misclassification in current vocabularies. We study various alterations in fungal taxonomy [4]. An example of the proposed changes are *Trichophyton Soudanense* to *Trichophyton Violaceum*, *Trichophyton megninii* to *Trichophyton rubrum*, and *Trichophyton equinum* to *Trichophyton Tonsurans.*

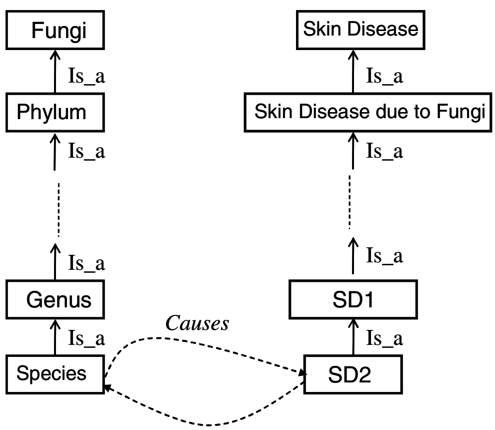


Fig. 5. Changing the fungi name can change the related disease name

As another example, the pathogenic fungus *Candida glabrata* is now called *Torulopsis glabrata* [26]. Usually changes in fungi taxonomy can alter the related disease name and description (Figure 5). For instance, the name of the fungus, *Allescheria boydii* which can cause various infections in humans, was changed to *Petriellidium boydii* and then to *Pseudallescheria boydii* within a short time [25]. Consequently, the infections caused by this organism were referred to as *allescheriasis*, *allescheriosis*, *petriellidosis*, and *pseudallescheriosis* in the medical literature [28]. In our domain, we need to update and improve the ontological structure of the FungalWeb and SKDON Ontologies regularly for the annotation of fungal genes and analyzing the role of the fungi species in various diseases. For example, the older version of the FungalWeb Ontology did not have sufficient terminology to annotate genes involved in *Malassezia* infections. To meet this new requirement, the updated version of the ontology has gained 26 additional terms addressing these infections.

As it is shown in Section 4, category theory can be used in RLR to represent agent interactions. It is also used as formalism for analyzing changes in FungalWeb Ontology and SKDON. Changes to any part of the ontologies may cause the conceptual design changes the state and also may cause alterations to other dependent artifacts. To represent different states of our conceptualization, we use a categorical discrete state-model, which describes states and events in diagrammatical notion.

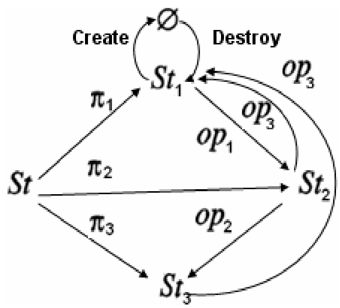


Fig. 6. A Class diagram that represents the transition between states [4]

Based on our application we designed our class diagrams following the method described in [27, 4] (Figure 6), which can be used to create patterns for learning agents. The *Opi* arrows in this figure represent the operations for the class, wherein the operation or event Op_1 causes an object in state St_1 to undergo a transition to state St_2. The operation Op_1 has no effect upon the object if it is in any other state, since no arrow labeled Op_1 originates from any other state. The object Ø in the diagram is the null state. The create arrow represents the creation of the object by assigning an identifier and setting its state to the initial defined state, and the destroy arrow represents its destruction.

6 Discussion and Challenges

Ontologies in general must change to update their ontological 'truth'. The heterogeneity of biomedical ontologies and the volatility of their knowledge sources increase the odds of different structural alterations. One issue in the domain of ontology evolution is lack of formal change models with clear, comprehensible semantics. In this manuscript we have introduced a novel multi-agent framework to handle changes in bio-ontologies based on category theory. This framework assists an ontology engineer to capture, track, represent and manage the changes in a formal and consistent way which enables the system to create reproducible results. Using category theory with its dynamic nature in our model allows capturing the full semantics of evolving bio-ontologies as well as providing a formal basis to represent agent interactions.

In the process of employing category theory as the core formalism for the RLR framework, we had to deal with a variety of challenges, including the reasoning issues and management of conceptualization changes. However, we are able to provide basic reasoning and inferencing for categories, though we still must improve the reasoning capability to cover more advanced services. The representation of conceptualization changes is another challenge, especially for abstract concepts and notions. To overcome this, we are working on grammatical change algorithms in linguistics and language evolution. Minimizing human intervention is an issue in the "Reproduction" phase, although improvement of the learning and negotiation algorithms for the agents may reduce the problem.

References

1. Ackerman, B.: Need for a Complete Dictionary of Dermatology Early in the 21st Century. Arch Dermatol 136(23) (2000)
2. Baker, C.J.O., Shaban-Nejad, A., Su, X., Haarslev, V., Butler, G.: Semantic web infrastructure for fungal enzyme biotechnologists. Journal of Web Semantics 4(3), 168–180 (2006)
3. National Library of Medicine, Medical Subject Headings, Bethesda, MD (1994)
4. Shaban-Nejad, A., Haarslev, V.: Categorical Representation of Evolving Structure of an Ontology for Clinical Fungus. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 277–286. Springer, Heidelberg (2007)
5. Devedžic, V.: Knowledge Modeling - State of the Art. Integrated Computer-Aided Engineering 8(3), 257–281 (2001)
6. Li, L., Wu, B., Yang, Y.: Agent-Based Approach for Dynamic Ontology Management. KES (3), 1–7 (2005)
7. Kam, P.S., Fu, A.W.: Discovering Temporal Patterns for Interval-Based Events. In: Kambayashi, Y., Mohania, M., Tjoa, A.M. (eds.) DaWaK 2000. LNCS, vol. 1874, pp. 317–326. Springer, Heidelberg (2000)
8. Höppner, F.: Knowledge discovery from sequential data. PhD thesis, Technical University Braunschweig, Germany (2003)
9. Sacchi, L., Bellazzi, R., Larizza, C., Porreca, R., Magni, P.: Learning Rules with Complex Temporal Patterns in Biomedical Domains. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS (LNAI), vol. 3581, pp. 23–32. Springer, Heidelberg (2005)
10. Sacchi, L., Larizza, C., Combi, C., Bellazzi, R.: Data mining with Temporal Abstractions: learning rules from time series. Data Mining Knowledge Discovery 15(2), 217–247 (2007)
11. Haarslev, V., Möller, R.: Description Logics for the Semantic Web: Racer as a Basis for Building Agent Systems. KI 17(3), 10–15 (2003)
12. Capobianco, M., Chesñevar, C.I., Simari, G.R.: Argumentation and the Dynamics of Warranted Beliefs in Changing Environments. Journal of Autonomous Agents and Multi-Agent Systems 11(2), 127–151 (2005)

13. Poigné, A.: Elements of Categorical Reasoning: Products and Coproducts and some other (Co-) Limits. In: Poigné, A., Pitt, D.H., Rydeheard, D.E., Abramsky, S. (eds.) Category Theory and Computer Programming. LNCS, vol. 240, pp. 16–42. Springer, Heidelberg (1986)
14. Kozen, D., Kreitz, C., Richter, E.: Automating Proofs in Category Theory. In: Furbach, U., Shankar, N. (eds.) IJCAR 2006. LNCS (LNAI), vol. 4130, pp. 392–407. Springer, Heidelberg (2006)
15. Rahwan, I., Ramchurn, S.D., Jennings, N.R., McBurney, P., Parsons, S., Sonenberg, L.: Argumentation Based Negotiation. Knowledge Engineering Review 18(4), 343–375 (2003)
16. Van Eemeren, F.H., Grootendorst, R.F., Henkemans, F.S.: Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Applications. L. Erlbaum Associates, NJ (1996)
17. Jennings, N.R., Parsons, S., Noriega, P., Sierra, C.: On argumentation-based negotiation. In: International Workshop on Multi-Agent Systems (IWMAS 1998), Boston, USA (1998)
18. Eilenberg, S., Mac Lane, S.: General Theory of Natural Equivalences. Transactions of the American Mathematical Society 58, 231–294 (1945)
19. Ehresmann, A.E., Vanbremeersch, J.P.: The Memory Evolutive Systems as a Model of Rosen's Organism-(Metabolic, Replication) Systems, vol. 16, pp. 137–154. Springer, Heidelberg (2006)
20. Resconi, G., Jain, L.C.: Intelligent Agents: theory and applications. Studies in Fuzziness and Soft Computing, vol. 155. Springer, Berlin (2004)
21. Resconi, G., Hill, G.: The Language of General Systems Logical Theory: A Categorical View. In: 3rd European Congress on Systems Science, Rome, pp. 1091–1096 (1996)
22. Johnson, M.W., Miller, T., McBurney, P.: Modelling Multi-agent Interaction Protocols using Categories Enriched over Pointed Sets. Technical Report, U of Liverpool (2007)
23. Papier, A., Chalmers, R.J.G., Byrnes, J.A., Goldsmith, L.A.: Framework for improved communication: the Dermatology Lexicon Project. Journal of the American Academy of Dermatology 50(4), 630–634 (2004)
24. Al-Aboud, K., Al-Hawsawi, K., Ramesh, V., Al-Aboud, D., AL-Githami, A.: An Appraisal of Terms Used in Dermatology. SKINmed 2(3), 151–153 (2003)
25. Odds, F.C., Arai, T., Di Salvo, A.F., Evans, E.G.V., Hay, R.J., Randhawa, H.S., Rinaldi, M.G., Walsh, T.J.: Nomenclature of fungal diseases. In: A report from a Sub-Committee of the Intl' Society for Human and Animal Mycology (ISHAM) (1992)
26. Cushion, M.T., Stringer, J.R.: Has the Name Really Been Changed? It Has for Most Researchers. Clinical Infectious Diseases 41, 1756–1758 (2005)
27. Whitmire, S.A.: Object Oriented Design Measurement. John Wiley & Sons, New York (1997)
28. Odds, F.C., Rinaldi, M.G.: Nomenclature of fungal diseases. Curr. Top. Med. Mycol. 6, 33–46 (1995)
29. Flouris, G.: On Belief Change and Ontology Evolution. Ph.D. thesis in the Department of Computer Science, University of Crete (2006)
30. MacFarlane, A.I.: Dynamic structure theory: A structural approach to social and biological systems. Bulletin of Mathematical Biology 43(5), 579–591 (1981)
31. Girard, J.Y., Lafont, Y., Taylor, P.: Proofs and Types. Cambridge University Press, Cambridge (1989)

An A-Team Approach to Learning Classifiers from Distributed Data Sources

Ireneusz Czarnowski, Piotr Jędrzejowicz, and Izabela Wierzbowska

Department of Information Systems, Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
{irek, pj, iza}@am.gdynia,pl

Abstract. Distributed data mining is an important research area. The task of the distributed data mining is to analyze data from different sources. Solving such tasks requires a special approach and tools, different from those dedicated to learning from data located in a single database. This paper presents an approach to learning classifiers from distributed data based on data reduction (the prototype selection) at a local level. The problem is solved through applying the A-Team concept implemented using the JABAT environment, which supports implementation of multiple-agent teams. The paper includes a general overview of the JABAT, the problem formulation and some technical details of the proposed implementation. Finally, the computational experiment results validating the approach are shown.

Keywords: Multi-agent system, learning classifiers, distributed data mining.

1 Introduction

Usually machine learning and data mining algorithms base on the assumption that all the training data can be pooled together in a centralized data repository. In the real life there are, however, numerous cases where the data have to be physically distributed due to some constraints. Applying traditional data mining tools to discover knowledge from distributed data sources may not be possible due to a variety of reasons (huge data volume, data privacy or others) [7]. Therefore, there is a need for data mining tools that can perform the analysis of the distributed data. The need to extract potentially useful patterns out of separated, distributed data sources created the new and important research area known as the distributed data mining or knowledge discovery from multi-databases.

Knowledge discovery from the multi-database is considered to be a more complex and difficult task than knowledge discovery from the single database [8]. In [8] a methodology for the two-stage data mining was discussed. The approach was based on extraction of prototypes from distributed data sources. The first stage performs local data analysis and the second combines the local results into a global one. In [9], a meta-learning process was proposed as an additional learning tool for combining a set of locally learned classifiers (decision trees in particular) into a global classifier. Meta-learning methodologies view data distribution as a technical issue and treat distributed data sources as parts of a single database. It was pointed out in [11] that

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 536–546, 2008.

such an approach offers rather a narrow view of the distributed data mining, since the data distributions in different locations often are not identical. Another approach to confront the discussed problem is to select out of the distributed databases only the relevant data. In [10] a relevance measure was proposed to identify the relevant databases for mining. In [17] the authors presented yet another approach to learning classifiers from distributed data sets. In this approach the extraction of useful patterns was based on applying the population learning algorithm proposed in [21].

Agent-oriented software engineering seems to be an attractive tool for implementation of methods and systems dedicated for distributed learning. In such an approach each site can have one or more associated agents that process the local data and communicate the results to other agents that control and manage the discovery knowledge process. For example, in [7], [9], [12] different distributed agent-based data-mining systems, denoted respectively as BODHI, PADMA and JAM, are shown.

In this paper the implementation of an agent-based environment, called JABAT, for solving the problem of learning from distributed data is proposed. JABAT (JADE-based A-Team) environment supports the construction of the dedicated A-Team architectures used for solving a variety of computationally hard optimization problems [6].

The paper deals with a distributed machine classification. The ultimate goal of learning is to produce a classifier that minimizes the performance criterion of classification error. The proposed approach to learning from distributed data involves two stages: at the local level the selection of reference vectors (called prototypes) from distributed data is carried out. At the global level integration and pooling of the selected vectors and generation of the global hypothesis take place.

At the local level the instance reduction algorithm is used. The instance reduction algorithm, originally proposed in [13], aims at obtaining a compact representation of the data set that includes reference vectors and consists of non-redundant information. In [13] it has been also demonstrated that in many cases reducing the data set size can increase efficiency of the supervised learning.

In this paper it is proposed to solve problems of instance selection and of learning from data, both of which are NP-hard ([20], [19]), through constructing a JABAT-based multi-agent tool. The JABAT environment has been chosen since it can produce satisfying results for problems from NP [6]. Moreover this environment provides an additional possibility of carrying the selection of instances in a distributed environment using multiple agents.

The paper is organized as follows: Section 2 presents a short overview of the JABAT. Section 3 gives some technical details of the proposed JABAT implementation for solving the problem of learning from distributed data sources. The approach involves two stages of mining – the local and the global, both based on the prototype selection. In Section 4 of the paper the computational experiment results are showed. Finally, the last section contains conclusions and suggestions of further research.

2 Overview of JABAT

During the last decade, a number of significant advances have been made in both the design and implementation of autonomous agents. A number of agent-based approaches have been proposed to solve different types of optimization problems

[1], [3], [4]. One of successful approaches to agent-based optimization is the concept of an asynchronous teams (A-Teams), originally introduced by Talukdar [5]. An A-Team is a collection of software agents that cooperate to solve a problem by dynamically evolving a population of solutions. An A-Team usually uses a combination of approaches, which often have been inspired by natural phenomena. An A-Team is also a cyclic network of autonomous, asynchronously working agents and shared, common memories. Each agent possesses some problems solving skills and each memory contains a population of temporary solutions of the problem to be solved. The agents cooperate by selecting and modifying these solutions.

JABAT has been designed as a middleware allowing implementation of the A-Team architecture with a view to solving computational optimization problems. The main feature of JABAT is its ability to solve instances of several different optimization problems in parallel using multiple hardware platforms. The user can easily add or delete a computer from the system. In both cases JABAT will adopt to the changes, commanding the agents working within the system to migrate.

The JABAT environment, is expected to be able to produce solutions to difficult optimization problems using a set of agents, each representing an improvement algorithm. To escape getting trapped into a local optimum an initial population of solutions called individuals is generated. Individuals, during computations, are eventually improved by agents, thus increasing chances for reaching a satisfactory solution.

Main functionality of the proposed tool includes generation of the initial population of solutions; application of solution improvement algorithms which draw individuals from the common memory and store them back after attempted improvement using some user defined replacement strategy and continuation of the reading-improving-replacing cycle until a stopping criterion is met.

The above functionality is realized by two main types of classes. The first one includes *OptiAgents*, which are implementations of improvement algorithms. The second are *SolutionManagers*, which are agents responsible for maintenance and updating of individuals in the common memory. Each *SolutionManager* is responsible for finding the best solution for a single instance of the problem. All agents act in parallel and communicate with each other exchanging solutions that are either to be improved (when solutions are sent to *OptiAgent*) or stored back (when solutions are sent to *SolutionManager*). An *OptiAgent* has two basic behaviors defined. The first is sending around messages on readiness for action including the required number of individuals (solutions). The second is activated upon receiving a message from some *SolutionManager* containing the problem instance description and required number of individuals. This behavior involves improving fitness of individuals and resending the improved ones back to the sender.

To create a *SolutionManager* two pieces of information are needed: the name of the class that represents task and the name of the chosen replacement strategy. The replacement strategy defines how the initial population is created, how solutions are chosen to be sent to optimising agents, how solutions that have been received from optimising agents are merged with the population, and when the process of searching stops. A simple strategy can, for example, draw a random solution and, when an improved solution is received, replace the worst solution in the population with it. A more sophisticated strategy can introduce recursion: it allows for division of the task into smaller tasks, solving them and then merging the results to obtain solutions to the initial task. The strategy may be used for these problems for which specific methods of *Task* and *Solution* classes are defined that are responsible for dividing and merging.

Apart from *OptiAgents* and *SolutionManagers* there are also other agents working within the system: responsible for initializing the process of solving of an instance of a problem, organizing the process of migrations and moving optimization agents among available computers that have joined the JABAT platform, writing down the results. To summarize, from the functional point of view, JABAT based on JADE engine [2] provides the basic services necessary to create distributed peer-to-peer applications dedicated for solving computationally hard optimization problems.

3 The JABAT Tool for Learning Classifiers

3.1 Problem Formulation

The problem of learning from data can be formulated as follows: Given a data set D, a set of hypothesis H, a set of performances P, the learning algorithm L outputs a hypothesis $h \in H$ that optimize P. In pattern classification application, h is a classifier (i.e. decision tree, artificial neural network, naive Bayes, k-nearest neighbor, etc.). The data D consists of training examples. Each example is represented by a vector of n+1 attributes, where n+1 attribute corresponds to the class label and the remaining attributes represent inputs to the classifiers. The goal of learning is to produce a hypothesis that optimizes the performance criterion.

In the distributed learning, data set D is distributed among data sources $D_1,\ldots,D_K$. which are stored in separate sites. In the distributed learning a set of constraints Z can be imposed on the learner. A typical constraint may prohibit the transfer of data from the separated sites to the central location. Another constraint can be imposed by the need to preserve privacy of the distributed data.

Thus, in the distributed learning there are subsets $D_1,\ldots,D_K$, of the data set D located across the sites 1,..,K, the set of constraints Z, the hypothesis class H and the performance criterion P. The task of the distributed learner L_d is to output a hypothesis $h \in H$ that optimize P using operations allowed by Z.

Based on the above one can observe that the problem of learning from the single (centralized) data set D is a special case of learning from distributed data where K=1 and Z is empty. Further details in respect to formulation of the general problem of learning from distributed data can be found, for example, in [14].

This paper proposes solving the problem of learning from distributed data in two stages (levels) of data mining. The first stage (local level) is dedicated to the selection of prototypes, independently on each separated site $D_1,\ldots,D_K$, and to produce the reduced data sets $S_1,\ldots,S_K$ of local patterns. In the second stage (global level) a set of global patterns is synthesized from sets $S_1,\ldots,S_K$ and finally the reduced set $S \in D$ is produced. It is assumed that reduced data sets preserves basic features of the analyzed distributed data. It is also assumed that the distributed databases have the same set of attributes and the datasets are syntactically homogenous.

To sum up, in our case the task of the distributed learner L_d is to output the hypothesis $h \in H$ that optimize performance criterion P using operations allowed by the set of constrains Z and using data located in K sites subsets $S_1 \in D_1, \ldots, S_K \in D_K$ of the data set D.

3.2 Prototype Selection

Prototype selection aims at creating a compact representation of the data that consists only of the relevant vectors. Prototype selection can be viewed as a data reduction approach, which removes a number of instances from the original training set in order to produce the reduced training set.

The proposed algorithm is based on calculating, for each instance from the original set, the value of its similarity coefficient, and then grouping instances into clusters consisting of instances with identical values of this coefficient, selecting the representation of instances for each cluster and removing the remaining instances, thus producing the reduced training set. The algorithm involves the following steps:

Step 1. Transform each data instance $X_k = \{x_{ij}\}$ (i=1…N_k, j=1…n+1), normalizing the value of each attribute x_{ij} into interval [0, 1] and then round it to the nearest integer, that is 0 or 1, where N_k is the number of instances in D_k, and k=1,…,K.

Step 2. For each instance from the original set X_k calculate the value of its similarity coefficient I_{ik}:

$$I_{ik} = \sum_{j=1}^{n+1} x_{ij} s_{jk}, \ i = 1...N_k,$$

where:

$$s_{jk} = \sum_{i=1}^{N_k} x_{ij}, \ j = 1...n+1.$$

Step 3. Map data instances (i.e. vectors X_k) into clusters denoted as Y_{vk}, v=1…t, where t is the number of different values of I_k. Each cluster contains input vectors with identical value of the similarity coefficient I_k.

Step 4. Select instances to be retained in each cluster. Let $|Y_{vk}|$ denote the number of instances in the cluster v, v = 1,…,t . Then the following rule for selecting instances are applied:

- If $|Y_{vk}| = 1$ then $S_k = S_k \cup Y_{vk}$,where z = 1,..,k.
- If $|Y_{vk}| > 1$ then $S_k = S_k \cup \{ x^v_{jk} \}$, where x^v_{jk} are reference instances from the cluster Y_{vk} selected by applying any optimization technique. In this paper the reference vectors are selected by the agent-based algorithm with the classification error as the performance criterion.

The above method of instance reduction was proposed in the earlier paper of the authors [13]. In [13] it was shown that the proposed technique can result in reducing the number of instances and still preserving the quality of the data mining results. It was also demonstrated that in many cases reducing the training set size can increase efficiency of the supervised learning.

3.3 Learning Classifiers from Distributed Data

To implement both stages of the proposed approach to learning classifiers from the distributed data the JABAT environment has been used. All the required classes have been defined in the package called LCDD *(Learning Classifiers for Distributed Data)*. Two main classes include *LCDD_Task* inheriting form the *Task* class and *LCDD_Solution* inheriting from the *Solution* class. These classes have been adopted from the earlier JABAT-based implementation of the two-dimension data reduction technique proposed in [15]. In this approach the data reduction process in case of the centralized, non-distributed database was also based on the idea of selecting the reference instances and removing the irrelevant attributes.

The *LCDD_Task* identifies the location of data set and clusters of the potential reference instances produced using the algorithm described in Section 3.2.

The *LCDD_Solution* contains representation of the solution. It consists of: the list of number of selected reference instances. The list, on its first t positions, defines how many elements from Y_v, $v = 1,\dots,t$ are placed in the cluster v, next positions represent the numbers of these elements from Y_v; the classification accuracy of the solution.

Each optimization agent operates on a single individual (solution) randomly drawn form the population of solutions stored in the common memory by the *SolutionManager.* An optimization agent tries to improve quality of the solution provided by the *SolutionManager*. This is done iteratively, using some local search procedure. After the stopping criterion is met the resulting individual is sent back to the *SolutionManager*, which, in turn, updates common memory by replacing a randomly selected individual with the improved one. Generally, the *SolutionManager* manages the population of solutions, which during the initial phase is generated randomly. To assure the required diversity, the initial population of solutions is drawn in such a way that random individuals represent different numbers of reference instances in each cluster.

The discussed JABAT-based tool is based on a strategy, denoted as *StrategyDiv,* that divides the learning classifier task into subtasks. It allows searching for solutions to the problem of learning classifiers from distributed data on two levels – local and global. Thus, the task of learning from data corresponds to the global level. Each subtask corresponds to the independent individual prototype selection problem and is performed for each distributed data site separately (the location of this data is given as a parameter of the subtask).

The local learning classifier aims at finding the optimum solution at the local level. A solution of the distributed learning classifiers problem at the local level is represented by the set of reference vectors (instances) i.e. by the compact representation of the data set from given local level. Each set of reference vectors is selected in the process that optimizes the classification accuracy. To obtain value of the classification accuracy any classification tool can be used.

The *StrategyDiv* is carried until all the subtasks have been solved. Afterwards solutions from the local levels are used to obtain the global solution: the compact representation of the all distributed data sets. This is done through merging local level solutions and transferring the reference instances from distributed sources to the centralized learning stage. The *StrategyDiv* merges the results to obtain solutions to the initial task where the global classifier is constructed. At the global level, the global hypothesis is determined by integrating the solutions $S_1,\ldots,S_K$ from individual data sources $D_1,\ldots,D_K$ respectively.

To solve the problem of the data reduction two types of agents that represent different improvement procedures have been implemented. In each case the agent's classes are inherited from the *OptiAgent* class. The first procedure – local search with tabu list modifies a solution by replacing a randomly selected reference vector from a randomly chosen cluster with some other randomly chosen reference vector thus far not included within the improved solution. The modification takes place providing the vector to be replaced is not on the tabu list. After the modification, the newly added reference vector is placed on the tabu list and remains there for a given number of iterations. This number depends on the cluster size and decreases for smaller clusters. The second procedure – local search, modifies the current solution either by removing a randomly selected reference vector from a randomly chosen cluster or by adding a randomly chosen reference vector thus far not included within the improved solution.

In both cases the modified solution replaces the current one if it is evaluated as a better one using the classification accuracy as the criterion. If, during the search, an agent successfully improves the received solution then it stops and the improved solution is transmitted to the *SolutionManager*. Otherwise, agents stop searching for an improvement after having completed the predefined number of iterations.

4 Computational Experiment

To validate the proposed approach it has been decided to carry out a computational experiment. Classification accuracy in case of the distributed data and the distributed data reduction as proposed earlier has been compared with the results obtained by pooling together distributed databases and applying the classification tool to the full dataset. Generalization accuracy was used as the performance criterion. The classifier was the C 4.5 algorithm.

Table 1. Average accuracy (%) of C 4.5 results obtained for the distributed datasets

Problem	Number of datasets				Full dataset
	2	3	4	5	
Customer	68.45	70.4	74.7	75.21	73.32
Adult	86.2	87.2	86.81	87.1	82.43

Table 2. Percentage and number of prototypes sent to the central site

	Number of datasets								Full dataset
Problem	2		3		4		5		
Customer	112	0.52%	174	0.81%	215	1.00%	245	1.13%	21600
Adult	341	1.26%	399	1.47%	447	1.65%	515	1.90%	27140

The experiment involved two data sets – *customer* (24000 instances, 36 attributes, 2 classes), *adult* (30162, 14, 2). The respective datasets have been obtained from [18] and [16]. The reported computational experiment was based on the "10 cross validation" approach. At first the available datasets were randomly split into training and test sets in approximately 9/10 and 1/10 proportions. The second step involved the random partition of the previously generated training sets into the training subsets each representing a different dataset placed in the separate location. Next, each of the thus obtained datasets was reduced using the proposed instance reduction approach. The reduced subsets were then integrated on the global level into a training set and the C 4.5 classifier (with pruned leaves) was trained. Ten such trials were run ten times each, using a different dataset partitions as the test set for each trial.

The above described experiment was repeated four times for the four different partitions of the training set into a multi-database. The original data set was randomly partitioned into, respectively, the 2, 3, 4 and 5 multi-database.

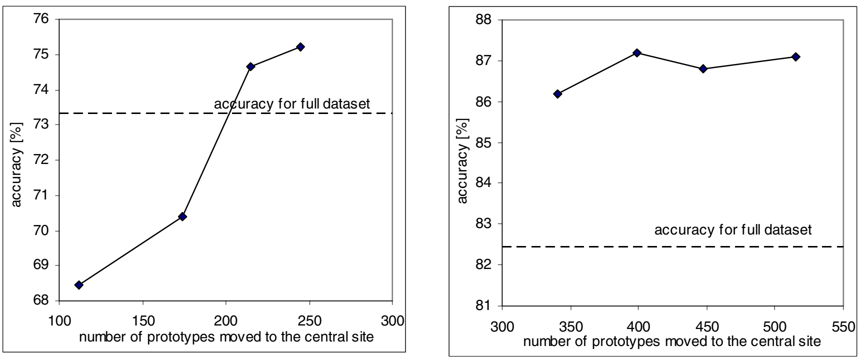


Fig. 1. Accuracy of classification versus the number of prototypes moved to the central site, respectively for problems: *customer* (on the left) and *adult* (on the right)

Experiment results are shown in Table 1-2 and in Fig. 1. The results have been averaged over all experiment runs carried. Table 1 shows accuracy of classification performed on the test sets by the C 4.5 classifier. The results have been obtained by applying the reduction of instances approach to multi-database. Additionally, Table 1 in the column “Full dataset”, presents accuracy of classification by C 4.5 algorithm applied after integration of all instances into the centralized database. Table 2 shows percentages and numbers of instances moved to the central site from the distributed datasets.

Fig. 1 shows accuracy of classification versus the number of prototypes moved to the central site from four different datasets for *customer* and *adult* problem respectively. It also shows the distribution of the classification accuracy on test data with respect to the size of data sent to the central site. These results show that even a small fraction of data representing the multi-database can give quite good classification accuracy.

Generally, it should be noted that the approach to learning classifiers from distributed data, based on data reduction on local levels, produces reasonable to very good results. For example, pooling the selected instances, from two distributed datasets, together into the training set assures classification accuracy of 68.45% and 86.2% for *customer* and *adult* respectively. For the comparison, accuracy of classification for the full dataset equals 73.32% and 82.43% respectively. However, in the first case the size of the data sent to the central site equals 0.5% and 1.3% of "full dataset".

It should be noted that in the case of *customer* data the proposed data reduction technique at the local level assures reasonable results with a high data compression rate. Also in the case of *adult* data, the data reduction on the local level resulted in both a very good accuracy of classification, better then for "full dataset", and a very high data compression rate.

5 Conclusion

This paper presents an approach to learning classifiers from distributed data based on data reduction (the prototype selection) at the local level. The problem is solved through applying the A-Team concept implemented using the JABAT environment. The paper presents some technical details of the implementation and results of the computational experiments.

The proposed approach involves two stages. At the first one the team of agents carries-out data reduction procedure aiming at elimination of redundant training examples and producing compressed training set. Computations are carried in parallel and independently for each distributed dataset located in different sites. At the second stage instances selected from the distributed databases are pooled together to provide input to data mining tools used in the knowledge discovery processes.

Computational experiment results confirm that the proposed technique and its JABAT implementation can produce very good results for the problem of learning classifiers from distributed dataset. The suggested algorithm allows selecting important information from many distributed databases, producing significantly reduced training set and still assuring good or very good generalization accuracy at the global level.

The experiments were carried using the C4.5 learning algorithm as the classification tool. The future work is needed to extend computational experiments to other learning algorithms.

Future work will also concentrate on extending agents mobility in JABAT. Generally, JABAT allows for agent migration to locations where the distributed data is physically stored. However, in the proposed implementation the agents did not have possibility to move to specified data location, they could migrate only between available containers i.e. between computers that have joined the main computer where the JABAT engine was running. Thus, the implementation of the migration schemes to local sites, where data is stored, is planned. Such a migration may help in more effective use of the available computational resources and eliminate necessity of transferring data between sites. The direct access to data may also prove advantageous from the data safety point of view.

References

1. Aydin M.E., Fogarty T.C.: Teams of Autonomous Agents for Job-shop Scheduling Problems: An Experimental Study, Journal of Intelligent Manufacturing, 15(4), 455–462 (2004)
2. Bellifemine F., Caire G., Poggi, A., Rimassa G.: JADE. A White Paper, Exp, 3(3), 6–20 (2003)
3. Marinescu D.C., Boloni L.: A Component-based Architecture for Problem Solving Environments, Mathematics and Computers in Simulation, 54, 279–293 (2000)
4. Parunak H.V.D.: Agents in Overalls: Experiences and Issues in the Development and Deployment of Industrial Agent-Based Systems. International Journal of Cooperative Information Systems, 9(3), 209-228 (2000)
5. Talukdar S., L. Baerentzen, A.Gove, P. de Souza: Asynchronous Teams: Co-operation Schemes for Autonomous, Computer-Based Agents, Technical Report EDRC 18-59-96, Carnegie Mellon University, Pittsburgh (1996)
6. Jędrzejowicz P., I. Wierzbowska: JADE-Based A-Team Environment, Lecture Notes in Computer Science, Springer Berlin/ Heidelberg, 3993, 719-726 (2006)
7. Kargupta, H., Park, B.-H., Hershberger, D., Johnson, E.: Collective Data Mining: A New Perspective Toward Distributed Data Analysis. In: Kargupta, H., Chan, P. (eds.) Accepted in the Advances in Distributed Data Mining, AAAI/MIT Press (1999)
8. Zhang, X.-F., Lam, C.-M., Cheung, W.K.: Mining Local Data Sources For Learning Global Cluster Model Via Local Model Exchange. IEEE Intelligence Informatics Bulletine 4(2) (2004)
9. Prodromidis, A., Chan, P.K., Stolfo, S.J.: Meta-learning in Distributed Data Mining Systems: Issues and Approaches. In: Kargupta, H., Chan, P. (eds.) Advances in Distributed and Parallel Knowledge Discovery, ch. 3, AAAI/MIT Press (2000)
10. Liu, H., Lu, H., Yao, J.: Identifying Relevant Databases for Multidatabase Mining. In: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 210–221 (1998)
11. Tsoumakas, G., Angelis, L., Vlahavas, I.: Clustering Classifiers for Knowledge Discovery from Physical Distributed Database. Data & Knowledge Engineering 49(3), 223–242 (2004)
12. Stolfo, S., Prodromidis, A.L., Tselepis, S., Lee, W., Fan, D.W.: JAM: Java Agents for Meta-Learning over Distributed Databases. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pp. 74–81. AAAI Press, Menlo Park (1997)
13. Czarnowski, I., Jędrzejowicz, P.: An Approach to Instance Reduction in Supervised Learning. In: Coenen, F., Preece, A., Macintosh, A. (eds.) Research and Development in Intelligent Systems XX, pp. 267–282. Springer, London (2004)
14. Caragea, D., Silvescu, A., Honavar, V.: A Framework for Learning from Distributed Data Using Sufficient Statistics and its Application to Learning Decision Trees. International Journal of Hybrid Intelligent Systems (2003)
15. Czarnowski, I., Jędrzejowicz, P.: An Agent-based Algorithm for Data Reduction. In: Proc. of the Twenty-seventh SGAI International Conference on Artificial Intelligence Cambridge, England (to appear, 2007)
16. Merz, C.J., Murphy, M.: UCI Repository of Machine Learning Databases, Irvine, CA: University of California, Department of Information and Computer Science (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`

17. Czarnowski, I., Jędrzejowicz, P.: Instance Reduction Approach to Machine Learning and Multi-Database Mining. In: Proceedings of the Scientific Session organized during XXI Fall Meeting of the Polish Information Processing Society, Informatica, ANNALES Universitatis Mariae Curie-Skłodowska, Lublin, pp. 60–71 (2006)
18. The European Network of Excellence on Intelligence Technologies for Smart Adaptive Systems (EUNITE) – EUNITE World Competition in domain of Intelligent Technologies, http://neuron.tuke.sk/competition2
19. Wróblewski, J.: Adaptacyjne metody klasyfikacji obiektów. PhD thesis, University of Warsaw, Warsaw (in Polish) (2001)
20. Krawiec, K.: Evolutionary Feature Programming. Cooperative learning for knowledge discovery and computer vision. Ph.D thesis, Poznan University of Technology, Poznan (2004)
21. Jędrzejowicz, P.: Social Learning Algorithm as a Tool for Solving Some Difficult Scheduling Problems. Foundation of Computing and Decision Sciences 24, 51–66 (1999)

The User Preference Learning for Multi-agent Based on Neural Network in Ubiquitous Computing Environment

Eungyeong Kim[1], Hyogun Yoon[1], Malrey Lee[1,*], and Thomas M. Gatton[2]

[1] School of Electronics & Information Engineering, ChonBuk National University 664-14, 1Ga, DeokJin-Dong, Jeonju, Chon Buk, 561-756, South Korea
[2] School of Engineering and Technology, National University, 11244 North Torrey Pines Road, La Jolla, CA 92037, USA
{rotnrwk, kosher}@kongju.ac.kr, mrlee@chonbuk.ac.kr, tgatton@nu.edu

Abstract. In order to provide users with intelligent services, ubiquitous computing needs to make user situation analysis in consideration of users' movement. Thus, the present study proposed the multi-agent model for users to share context information and for user situation analysis and context learning using Bayesian Neural Network. The proposed user context structure of multi-agent distinguishes between dynamic and static contexts according to the volume of context change, and defines correlations among objects. Therefore, multi-agent can be aware of users' situation.

Keywords: Ubiquitous computing, Multi-agent, Context awareness, Mobile device.

1 Introduction

Since Mark Weiser proposed ubiquitous computing in 1993, service environment has been changed in a way of recognizing users' context using mobile devices and providing necessary real-time services. These are mainly designed to monitor the user's situation and to support the service based on multi-agent and events. And then, multi-agent learns users' ability to cope with services caused by user surrounding situations and events. When the users are in a similar situation or do a similar action, the multi-agent based on the learning result can find service resources and provide them to the users more quickly [1][2][3].

However, these systems can recognize user's context for the service environment of the specification area. Moreover, if a user gets out of the area, the user's multi-agent cannot understand the user's situation because of the different type of the context resources. Even if the type of context resources is uniform, it takes a lot of time and effort to process a large quantity of context resources because user’s mobile devices have a restricted processing ability. Therefore, we need a multi-agent system that can

---

[*] Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 547–556, 2008.

find context resources exactly for a specific user and understand the user's situation. And then service resources, adaptable for the user's situation, should be provided in real-time [3][4][5][6][7][8].

This paper proposes the multi-agent model for sharing the context information, analyzing the situation of users and context learning method using Bayesian Neural Network. The structure of the proposed multi-agent model was designed to process services through context information sharing, context analysis and learning process. Moreover, the proposed system was designed adaptive for mobile devices in consideration of users' movement. User contexts are recognized by collecting information on the user's location and service history through the users' mobile device, and context information shares. For context information sharing, we used hybrid P2P, which is a P2P structure with hybrid functions. In addition, for situation analysis and learning process, we applied Bayesian Neural Network modified for portable devices.

In this paper, Section 2 describes the contents of this study, and Section 3 explains the structure and the characteristics of the proposed situation analysis multi-agent model. Section 4 explains the method and contents of multi-agent model evaluation, and Section 5 summarizes the study and draws conclusions.

2 Ubiquitous Computing System

In ubiquitous computing, context awareness is a process for recognizing the user's place, time, state and surrounding situation and providing services intelligently according to the context. The concept of context awareness was defined by Schilit[9], Dey[2][3][6][10][15], Pascoe[11], Crowley[14], etc. According to their definitions, context awareness understands the user's state and surroundings specifically and monitoring changes in such information in real-time. In addition, a context-aware model is for analyzing the user's needs accurately.

The ubiquitous computing system can be to provide services in 5 any (Anytime, Anywhere, Any Network, Any Device, Any Service) quietly while the user is not aware of the existence of computers and networks (5C: Computing, Communication, Connectivity, Contents, Calm). For this, ubiquitous computing needs USN (ubiquitous Senor Network) for connection between electronic spaces and physical spaces, context mining for analyzing users' situation, and agent technique for providing necessary information automatically [3].

Representative researches ubiquitous computing include Oxygen Project [17][21]at MIT, CoolTown Project at HP, Gaia [18] at the University of Illinois, Aura at Carnegie Mellon University [19], and RCSM (Reconfigurable Context-Sensitive Middleware) at the Arizona State University.

The Oxygen Project at MIT is making various technological challenges in order to support very dynamic and diverse human activities. In particular, they are developing human-centered computing technologies and applications freely usable anywhere. Until now, the Oxygen Project has developed device technologies (E21s, H21s), network technologies (N21s), software technologies, recognition technologies and user technologies.

The CoolTown Project was launched by Tim Kindberg at the HP Internet and Mobile Systems Laboratory, aiming at the implementation of real World Wide Web for connection between the real world and virtual world and the development of software and service information equipment for the implementation [16]. Based on the technology, HP is developing CoolTown Art, conference rooms, schools, museums, etc. suitable for ubiquitous computing environment.

Gaia developed by the University of Illinoisaims at the realization of an active space for requesting information necessary in physical spaces where the user lives and executing tasks using various available resources. An active space is a natural combination of electronic spaces and physical spaces. To form such a space, Gaia agents infer various conditions on the current situation, respond to logical queries, and adapt behavioral patterns to different situations through various inference mechanisms including first order logic, temporal logic and fuzzy logic and learning mechanisms such as Bayesian learning and reinforcement learning based on a context prediction model [18].

The Aura Project at Carnegie Mellon University aims at the provision of computing environment that minimizes users' attention, taking note of the fact that users' intention or interest, which is an extremely scarce resource, hardly increases. The research areas of the Aura Project cover from hardware to operating system, middleware, application programs and users [19][22].

RCSM (Reconfigurable Context-Sensitive Middleware) under development at the Arizona State University aims at providing middleware for the easy design of applications demanding active connection between contexts and services and ad-hoc communication. RCSM provides services in which context-adaptive applications are combined with ad-hoc communication. RCSM pursues object-oriented development environment and is closely related with CORBA. In addition, RCSM applications are composed of context-adaptive objects, and components sensitive to performance such as ad-hoc communication are implemented as hardware using FPGA (Field Programmable Gate Arrays) in order to solve the performance problem of embedded equipment [5].

Ubi-uCAM (an Unified Context-Aware Application Model for ubiHome) developed by the Gwangju Institute of Science and Technology in Korea is a context-based application service models. Ubi-uCAM is composed of ubi-Sensor and ubi-Service that use intelligent sensors and application services. Ubi-Sensor and ubi-Service create information and exchange the information with each other using processing and networking models. In particular, ubi-Service creates integrated context by combining first-drafted contexts in the form of 5W1H provided by multiple ubi-sensors and recognizes users' state more precisely through the integrated context. In addition, it provides the integrated context to currently connected ubi-sensors so that the sensors can draft contexts more precisely [8].

3 A Model of Learning Multi-agent

3.1 Structure of Multi-agent

Recently, cellular phones are widespread throughout the world, though they are various among countries. Thus, cellular phones are suitable for getting services in ubiquitous computing. In addition, users can access necessary services regardless of where they are. Thus, this paper proposed multi-agent model that can recognize and process a user's context via the user's mobile device. For analyzing user's situation, the multi-agent model used context information such as the location of the cellular phone, phone calls, the local environment of the local server, users' services, etc.

For this process, we designed a context-aware and processing model with structure as in Figure 1.

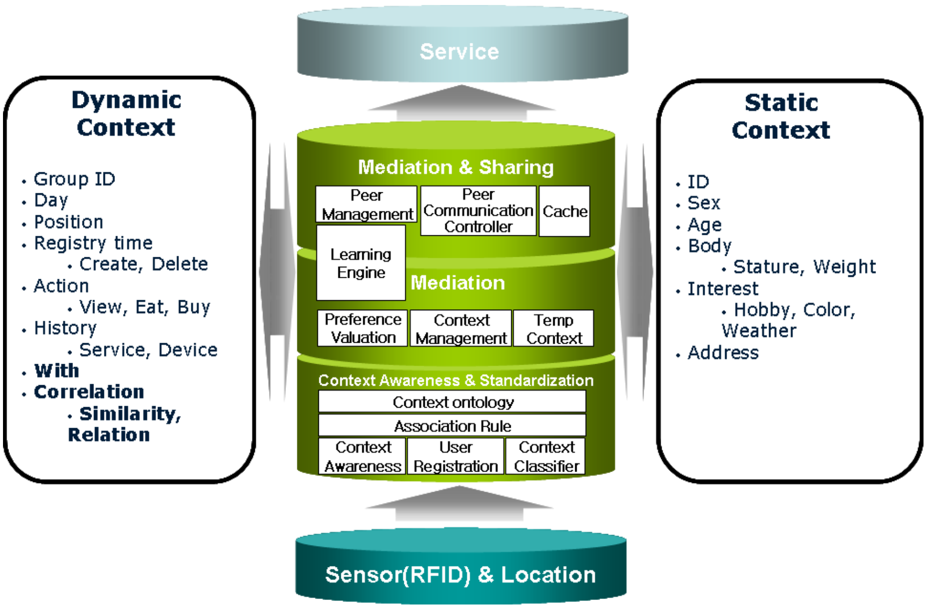


Fig. 1. This shows a figure consisting of agent of fourth level. Each agent achieves sequential and independent work

The model is composed of 'context awareness & Standardization agent' for recognizing and normalizing context, 'Mediation agent' for analyzing the user's situation based on the context, context database for recording context, 'Sharing agent' for sharing context, and 'Service agent' for providing services.

Context awareness & Standardization agent

This agent is a sensor that registers the recognized user at the local server and is provided by the local server with local information and other user's shared information. Then, it performs normalization according to the user's context type. It is composed of 'context awareness and classifier', 'association rule' and 'ontology'. 'Context awareness and classifier' can use mobile antennas around the user. In this study, however, it locates the user using APs (access points) and receives local information and shared information from the local server. Delivered information and the user's phone calling history are given association according to the user's context type. This process applies 'association rule'. In addition, ontology sets relationship (connection strength) on the user's context and integrates the context.

Mediation agent
A Mediation agent evaluates the user's preference with Bayesian Neural Network, and adjusts connection strength. As a result, the proposed model can learn the user's context and cope with additional contexts. It is composed of ‘preference evaluation’, ‘context management’ and ‘temp context’. ‘Preference evaluation’ compares the integrated data as input information with the user's history, and evaluates the user's preference. ‘Context management’ classifies contexts resulting from the evaluation into dynamic contexts and static contexts. ‘Temp context’ stores the classified contexts temporarily and sends them to ‘Sharing agent’.

Sharing agent
A Sharing agent transmits the results of analysis to the local server and is provided with additional service information necessary. For this, it is composed of ‘Peer communication controller’, ‘Peer management’ and ‘Cache’. ‘Peer communication controller’ maintains communication with the local server, and shares and manages the analyzed user information. ‘Peer management’ manages the signal of the controller and informs the local server of peers' state. ‘Cache’ is used for fast transmission of service information from the local server to the cellular phone.

Context database
Context database divides user contexts into dynamic contexts and static contexts according to variation, and stores them. Dynamic contexts are information related to the user's movements and service history. Static contexts are information related to the user's personal details.

3.2 User Preference Learning

User preference learning is the process of ‘Mediation’ explained in Section 3.1. In addition, users' actions and goals are determined and learned using Bayesian Neural Network. The results of learning supports the collection of service information and appeals for the sharing of context information. Figure 2 is a user preference learning model that uses normalized context information as training data of MLP based on Bayesian Neural Network.

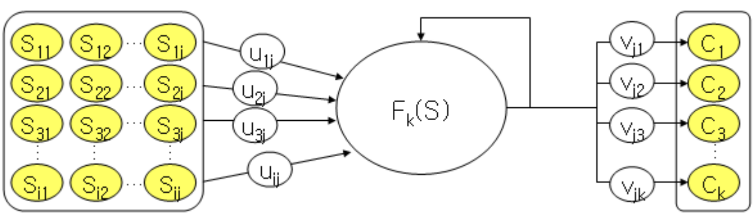


Fig. 2. Structure of MLP for user’s context awareness based on Bayesian Neural Network

The input layer of MLP is the user's context information and shared information from the local server. In addition, its output layer is set according to the type of the user's context database. For evaluation, we modified Bayesian Neural Network in accordance with the processing capacity of cell phones. The process below is for evaluating users' context information.

When S_i is shared context information and C_k is the user's context information, the evaluation function is expressed as (1).

$$f_k(S) = b_k + \sum_j v_{jk}(tanh(a_j + \sum_i u_{ij} S_i)) \tag{1}$$

In Equation 1, u_{ij} is connection weight between i input node and j hidden node, and v_{jk} is connection weight between j hidden node and k output node. a_j is bias of j hidden node, and b_k is bias of k output node. Accordingly, it has a regression model in which the dependent variable has a real number. In addition, the conditional distribution of C_k, which is the dependent variable for input S, has normal distribution (2) in which the mean is $f_k(S)$ and the standard deviation is δ_k. Because each output value is always independent from input, it satisfies the user's state value.

$$p(C|S) = \prod_k \frac{1}{\sqrt{2\pi}\delta_k} e^{\frac{-(f_k(S) - C_k)^2}{2\delta_k^2}} \tag{2}$$

In addition, in a classification model that has the dependent variable among k classes, dependent value C_k can be redefined as the conditional probability of multiple classes using k output nodes of Softmax model [19].

$$p(C|S) = \frac{e(f_k(S))}{\sum_{k'} e(f_{k'}(S))} \tag{3}$$

Weight and bias minimize the error between result values and dependent values for learning in training data composed of input value S_i and dependent value C_i. In this case, $C_{(k+1)}$ for new input can be predicted easily. (4) expresses the distribution of output value $C_{(n+1)}$ predicted for given new input value $S_{(n+1)}$.

$$\begin{aligned} &p(C_{(n+1)}|S_{(n+1)}, (S_1, C_1), (S_2, C_2), \ldots, (S_n, C_n)) \\ &= \int p(C_{(n+1)}|S_{(n+1)}, \theta) p(\theta|(S_{(1)}, C_{(1)}), \ldots, (S_{(n)}, C_{(n)}) d\theta \end{aligned} \tag{4}$$

In (4), θ indicates the parameter denoting all weights and biases. If the distribution of input value has not been modeled, the probability function of θ is expressed as (5).

$$L(\theta|(S_1, C_1), \ldots, (S_n, C_n)) = \frac{v_k}{N_{Tot}} \prod_{i=1}^{n} p(C_i|S_i, \theta) \tag{5}$$

Given input S_i and parameters, the distribution of dependent value C_{ii} can be defined as a network model of a regression model (2) and a Softmax classification model (3). The best method of predicting $C_{(n+1)}$ is estimating the mean of predicted distribution using a squared error loss function. Then the mean is the expectation value of post-distribution of θ and if this is summed up into a regression model it is expressed as (6).

$$\hat{C}_{(n+1)} = \int f_k(S_{(n+1)}, \theta) p(\theta|(S_{(1)}, C_{(1)}), \ldots, (S_{(n)}, C_{(n)})) d\theta \tag{6}$$

The result value of (6) extracts through ‘service’ service information within the range of expectation value from service information in the server.

Through this process, the proposed model learns the user's preference. In addition, extracted service information is delivered to the user's cell phone and the contents of service are recorded in the context database.

4 Experiment and Evaluation

In order to evaluate the proposed model, we conducted a questionnaire survey of some 400 person’s movie goers using experimental data. With the collected data, we evaluated the weights of contexts that can influence services (7).

$$\begin{aligned} Cov(CX, CY) &= E[(CX-\mu_{CX})(CY-\mu_{CY})] \\ &= E(CXCY) - \mu_{CX}\mu_{CY} \end{aligned} \tag{7}$$

(7) is an equation that evaluates the correlation between user contexts and services for deriving the weights of the contexts. CX is context variables based on which users choose a movie and CY is the variables of service information.

Table 1. Corrections between contexts and service

Gender	Hobby	Special Ability	Interest	Internet Recommendation	Company	Acquaintances Recommendation
Man	8.2%	4.7%	6.1%	12.1%	52.4%	16.5%
Women	7.5%	4.8%	4.6%	11.8%	51.5%	19.8%

Table 1 shows the results of evaluating the experimental data. The most influential context variable in users' choice of movies was company and acquaintances' recommendation. This suggests that if services are recommended to users based on locally shared information, the users' satisfaction with service can be enhanced. Information recommended using these results as the weights of Bayesian Neural Network was evaluated with MAE (Mean Absolute Error) (8).

$$|E| = \frac{\sum |C_{i+1} - r_i|}{N} \tag{8}$$

MAE is the mean absolute error in the experiment, and means the mean prediction error for the whole of predictions made. In MAE $C_{(i+1)}$, is predicted preference obtained from Bayesian Neural Network, and r_i is actual preference. N is the total number of predictions. Prediction of preference for services was evaluated as in Figure 9.

According to the results of evaluation, error in satisfaction with service was different according to genre and company. Single movie goers chose motives based on their hobby and interest regardless of genre. However, those in company with family or a lover chose movies based on the company's taste and opinion. In addition, error was similar when the company was friends and when the company was family or a lover.

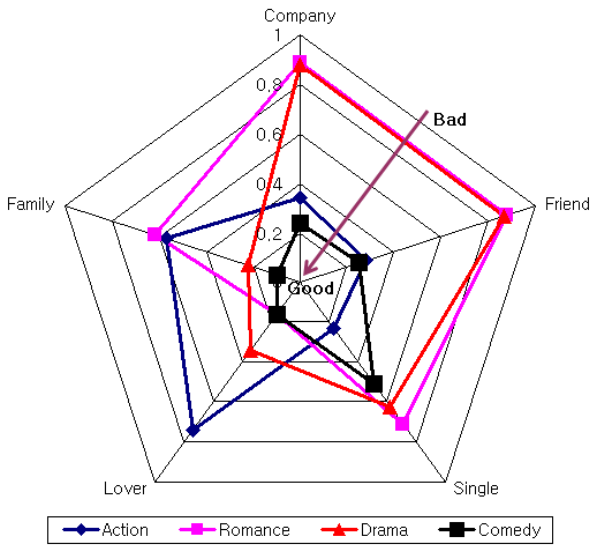


Fig. 3. This shows a figure of MAE result by Gender and companion

Accordingly, it is highly probable that users choose services considering not only themselves but also surrounding situation and company.

5 Conclusions

This paper proposed a model of multi-agent that recognizes and learns contexts through the user's cell phone in mobile environment in consideration of the user's movement and the real-time services. The proposed model was designed adaptive for mobile devices in consideration of users' movement. In addition, hybrid P2P was applied so that users' context information is shared and used as data for learning users'actions.

In the experiment for evaluating the proposed model, we measured the weights between context information collected through a questionnaire survey and services, and applied the weights to Bayesian Neural Network. The results showed that users' satisfaction with service is highly dependent on the users themselves, their company and surrounding contexts. In addition, the proposed model of multi-agent can learn users' actions from shared information and this enables the provision of services adequate for the users' contexts in dynamic or new environment.

Acknowledgment

"This work was supported by grant Rol-2006-000-10147-0 from the Basic Research Program of the Korea Science & Engineering Foundation and the second stage Brain Korea 21 Project"

References

1. Weiser, M.: Some Computer Science Problems in Ubiquitous Computing. Communications of the ACM, 75–84 (1993)
2. Dey, A.K., Abowd, G.D.: Towards a Better Understanding of Context and Context Awareness. GVU Technical Report GIT-GVU -99-22. Submitted to the 1st International Symposium on Handheld and Ubiquitous Computing (HUC 1999) (1999)
3. Dey, A.K., Abowd, G.D.: Towards a Better Understanding of Context and Context Awareness. In: Proceedings of the CHI 2000 Workshop on The What, Who, Where, When and How of Context Awareness (2000)
4. Yoon, H.J.: Researches and Trends of Ubiquitous Technology. Ubiu.net Report. 1 (2006)
5. Moran, T.P., Dourish, P.: Introduction to This Special Issue on Context-Aware Computing. Human-Computer Interaction (HCI) 16 (2001)
6. Salber, D., Dey, A.K., Abowd, G.D.: The Context Toolkit: Aiding the Development of Context-Aware Applications. In: The Workshop on Software Engineering for Wearable and Pervasive Computing, Limerick Ireland (2000)
7. Couder, P., Kermarrec, A.M.: Improving Level of Service of Mobile User Using Context Awareness. In: 18th IEEE Symposium on Reliable Distributed System, pp. 24–33 (1999)
8. Jang, S., Woo, W.: ubi-UCAM: A Unified Context- Aware Application Model. LNAI(Contex03), pp.178–189 (2003)
9. Schilit, B.N., Adams, N.I., Want, R.: Context-Aware Computing Applications. In: IEEE Workshop on Mobile Computing Systems and Applications, pp. 85–90 (1994)
10. Dey, A.K.: Understanding and Using Context. Personal and Ubiquitous Computing Journal 5(1), 4–7 (2001)
11. Pascoe, J.: Adding generic contextual capabilities to wearable computers. In: Proceedings of 2nd International Symposium on Wearable Computers, pp. 92–99 (1998)
12. Kim, C.S., Kim, J.M., Bae, S.J., Woo, S.M., Kim, H.Y.: Infrastructure Construction and Management Technologies for Ubiquitous Computing Environment. Analysis of Information Communication Trends 19(5), 9–20 (2004)
13. Lim, C.S.: Ubiquitous: Technology or Service? In: Electronic Commerce Integration Forum (ECIF 2005) (2005)
14. Crowley, J.L., Coutaz, J., Rey, G., Reignier, P.: Perceptual Components for Context Aware Computing. In: Borriello, G., Holmquist, L.E. (eds.) UbiComp 2002. LNCS, vol. 2498, pp. 117–134. Springer, Heidelberg (2002)
15. Dey, A.K.: Providing Architectural Support for Building Context-Aware Applications. Ph.D. Thesis, Department of Computer Science, Georgia Institute of Technology, Atlanta (2000)
16. TRON Project, `http://www.assoc.tron.org/jpn/tp.html`
17. MIT the media Lab, `http://www.media.mit.edu`
18. Romn, M., Hess, C.K., Cerqueira, R., Ranganathan, A., Campbell, R.H., Nahrstedt, K.: Gaia: A Middleware Infrastructure to Enable Active Spaces. IEEE Pervasive Computing 1, 74–83 (2003)
19. Garlan, D., Siewiorek, D., Smailagic, A., Steenkiste, P.: Project Aura: Towards Distraction-Free Pervasive Computing. IEEE Pervasive Computing 1(2), 22–31 (2002)
20. Preuveneers, D., Bergh, J.V., Wagelaar, D., Georges, A., Rigole, P., Clerckx, T., Berbers, Y., Coninx, K., Jonckers, V., Boscchere, K.D.: Towards an extensible context ontology for ambient intelligence. In: Markopoulos, P., Eggen, B., Aarts, E., Crowley, J.L. (eds.) EUSAI 2004. LNCS, vol. 3295, pp. 148–159. Springer, Heidelberg (2004)

21. Wennlund, A.: Context-aware Wearable Device for Reconfigurable Application Networks. Department of Microelectronics and Information Technology (IMIT2003) (2003)
22. Project Aura, `http://www-2.cs.cmu.edu/~aura/`
23. Rossi, F., Brieuc, C.-G.: Functional multi-layer perceptron: a non-linear tool for functional data analysis. Neural Networks 18, 40–60 (2005)
24. Datta, A.: MobiGrid: Peer-to-Peer Overlay and Mobile Ad-Hoc Network Rendezvous - a Data Management Perspective. In: Eder, J., Missikoff, M. (eds.) CAiSE 2003. LNCS, vol. 2681, pp. 16–20. Springer, Heidelberg (2003)

Performance Evaluation of Mobile Grid Services

Sze-Wing Wong and Kam-Wing Ng

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong SAR
{swwong, kwng}@cse.cuhk.edu.hk

Abstract. This paper presents the performance evaluation of Mobile Grid Services developed by using the MGS application programming interface. The MGS API, constructed by combining an existing mobile agent system (JADE) and a generic grid system toolkit (Globus), is proposed to support the development of Mobile Grid Services (extended Grid services with mobility during execution). To examine the performance of the Mobile Grid Services, a series of experiments are conducted. The results, analysis and the overheads estimation are presented in this paper.

1 Introduction

Grid computing has been drawing a lot of attentions from both academia and industry in recent years. A Grid [1] is a set of resources distributed over wide-area networks that can support large-scale distributed applications. Grid computing enables users and systems to dynamically share resources, balance the loading on resources, and perform possible parallel processing.

Regarding to current grid technologies, the services provided by grid services providers are usually stationary. They are fixed on the Grid node and cannot be moved to other nodes even if these other nodes have lots of idle resources. This static type of services results in many drawbacks such as continuous connection and overloading with service requests. If the grid services can become mobile such that they can be moved to more appropriate nodes, grids can maximize their resource usages by coordinating both services and resources.

Mobile Grid Services can be realized by enhancing the original static Grid Service with the ability of migration. Since this is still a young research area, there is not much work in the literature. Our research aims is to develop a middleware framework to support Mobile Grid Services in a secure manner. A preliminary middleware framework [2] has been already proposed. Moreover, the MGS API [3] has been implemented to ease the development of the Mobile Grid Services. Furthermore, a set of security mechanisms for the framework [4] has been also proposed in order to supplement the early description and enhance the MGS API. In this paper, the performance evaluation of the Mobile Grid Services will be presented.

This paper is structured as follows: The next section gives an explanation of the Mobile Grid Services. In Section 3, an overview of our proposed Mobile Grid Services framework and the MGS API is illustrated. Section 4 presents and analyzes the

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 557–566, 2008.

results of the experiments for performance evaluation. Some works related to Mobile Grid Services are reviewed in Section 5 and finally we give a conclusion in Section 6.

2 Mobile Grid Services

Grid Services are "stateful" Web Services which conform to the OGSA [5] standard. To improve the Grid Services, Mobile Grid Services are proposed as an extension. Unlike the original static Grid Services, Mobile Grid Services are able to move from nodes to nodes in the Grid because of their mobility capability. They can leave their hosts and migrate to other Grid nodes containing more idle resources. Mobile Grid Services can be seen as a special kind of Web Services with state and mobility.

With the help of the mobility, the services can travel throughout the Grid to get information from Grid nodes, execute on those nodes and bring the results back to their original hosts. This can improve the practicability and flexibility of Grid Services. Due to the service mobility, the overload problem can be relieved even if a large amount of service requests are present at the same time. For remote data access, the continuous connection between Grid nodes is essential for static services. However, this is less critical for Mobile Grid Services where their execution can be moved to the node storing the data directly.

3 Mobile Grid Services Framework

3.1 Framework Overview

We are working on a project aiming to develop a middleware framework for Mobile Grid Services [2]. The framework is constructed by joining the Globus Toolkit [6] and the Java Agent Development Framework (JADE) [7].

In our framework, a JADE mobile agent (supporting weak mobility) is used to encapsulate the actual working unit of the application task. JADE containers are setup throughout the grid to provide platforms for mobile agents to move to. For each application, a Globus Grid Service will create the corresponding mobile agents and act as a relay between users and the mobile agents. In this way, Mobile Grid Services are realized as a type of grid services which distribute the actual working tasks to mobile agents. The advantage of this design is that the existence of the relay part makes the Mobile Grid Services conform to the Globus grid service architecture. At the same time, the mobile agent part is able to exploit useful resources in other grid nodes.

Three main components in each Mobile Grid Service are Agent Manager, Task Agent and Monitor Agent: Agent Manager is responsible for managing the agents used in its own service and for redirecting information (requests and results) between the client and agents; Task Agent is a JADE mobile agent doing the actual service task of the Mobile Grid Service; Monitor Agent is a JADE mobile agent which is responsible for monitoring the grid resource information (from a Mobile Grid Service called Resource Information Service) and making migration decisions for the Task Agent.

3.2 MGS API Overview

The MGS [3] application programming interface is written in the Java programming language. It is built on libraries of the Java Agent Development Framework (JADE) for the implementation of mobile agent management and communication. At the same time, it uses some of the services provided in the Globus Toolkit 4.

This API consists of a collection of libraries for supporting the development of Mobile Grid Services. As an enhancement, security supports [4] have been added to the MGS API to protect the execution of Mobile Grid Services in potentially hostile environments. The supported security mechanisms include authentication, authorization, message integrity and confidentiality, agent permission and agent protection.

4 Performance Evaluation of MGS

Experiments were conducted to evaluate the performance of Mobile Grid Services developed by using the MGS API. In the experiments, the execution times of different Mobile Grid Services were measured under different settings. From their execution times, we can examine the performance and the overheads of the Mobile Grid Services.

The experiments were carried out in a grid with four nodes running Fedora Core 3. Each node is equipped with 3.2GHz dual-Xeon CPU and 2 GB memory. The machines were connected through a dedicated 10Mbit/sec switched Ethernet. They were all running with Globus toolkit 4.0 [6] and JADE 3.4. Their Java Virtual Machine versions were JDK5.0.

In the experiments, the four nodes were called Hosts A, B, C and D. Host A acted as the home host where the experimental services were running on it. On Host B, the JADE main container and the Resource Information Service were running. The roles of Host C and Host D were providing available platforms for agents to migrate to.

The details, results and analysis of the experiments will be presented in the following sections.

4.1 MGS Performance

In this experiment, the load balancing performance of MGS with different numbers of available hosts was measured.

To test the performance of the Mobile Grid Services, a sorting service was implemented by using the MGS API. When a service was requested, a new Task Agent would be created. Monitor Agent was created with the Task Agent in order to handle the resource information and make migration decision. After that, 600000 integers were generated randomly and insertion sorting was carried out in the Task Agent. After it finished the sorting, it would notify the Agent Manager of the service. As the service was a Mobile Grid Service, the Task Agent could migrate to other hosts with better resources during the sorting process.

The sorting service was setup on Host A in normal mode (no security facilities running) first. To test the load balancing, the sorting service was requested 50 times simultaneously with various numbers of hosts (1, 2, 3 and 4) available in the grid. Their execution times were measured as the time from the Task Agent creation to the receiving of Task Agent's final notification.

The experiment was repeated by requesting the service for 10 times and 1 time. To obtain more reliable results, 2, 10 and 5 trials were performed respectively in the experiments simultaneously requesting service for 50, 10 and 1 times. The average execution times of the sorting service under different settings are shown in Fig. 1 and Table 1.

Table 1. Average execution times (in seconds) of the sorting service under different settings

	1 host	2 hosts	3 hosts	4 hosts
50 requests	11645.0	4571.3	2415.9	1772.0
10 requests	1939.6	962.9	717.6	691.2
1 request	266.6	266.5	266.6	267.2

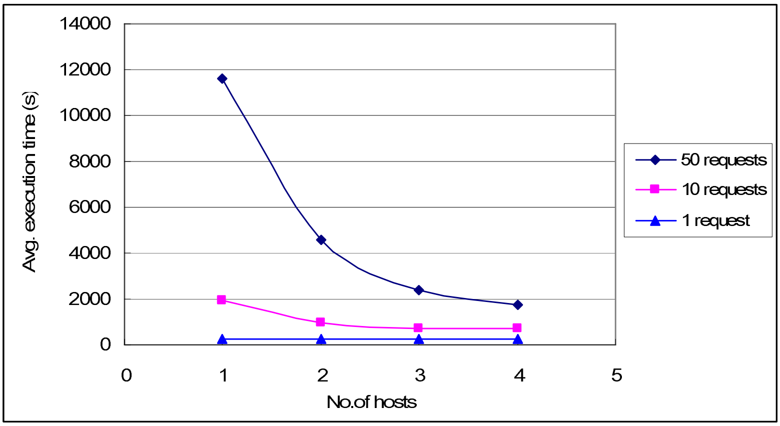


Fig. 1. The changing of the execution time against the number of available hosts

For invoking 50 sorting services at the same time, the host was overloaded heavily due to the work load of services when one host existed only. The created Task Agents in the services needed to share the CPU cycles and to perform the execution in turns. The result was that the average execution time of each service was over 11000 seconds (which was 43 times more than the single service's execution time). When one more host was available, the average execution time reduced to about 4600 seconds which was 39.26% of the single host's execution time. The work load of the tasks was shared between the two hosts in the grid and the service performance improved significantly. When there were three hosts, the execution time further improved to 20.75% of the single host's execution time. For using four hosts, the average execution time reduced to 15.22%. The curve was similar to the curve "y=1/x" and we can see that the execution time tended to be further decreased when more hosts are available. This shows that the overloading problem can be solved and the performance of the service can be improved by providing more hosts and resources for service migration.

When 10 sorting services were executed together in the home host only, overloading occurred and average execution time was nearly 2000 seconds (over 7 times of the single service's execution time). When the grid size increased to 2, 3, and 4, the average execution time improved to 49.65%, 37.00%, and 35.63% of execution time in a single host respectively. We can observe that the improvement was minor when available hosts changed from 3 to 4. Their execution times appeared to approach a boundary.

The reason is that all Task Agents were created on the home host (Host A) and they would initially execute at there. At this part of execution, the Task Agents needed to compete for resources (mainly CPU cycles) in the overloaded home host before they had opportunities to migrate. Owing to the design and the configuration of the Resource Information Service and the Monitor Agents, the formation of the load balancing required some time to finish. Thus, each agent spent certain time on the former part of the service execution but only little work could be done due to the overloading. When several hosts were available (i.e. the execution time became short), the overall execution times of the Task Agents were mainly contributed by this overloading period. Further adding available hosts in the grid could not speed up the load balancing and so could not shorten the length of this period (Amdahl's Law). Therefore, the execution time of the services could not improve significantly when approaching the boundary even if more hosts are added.

For requesting one service only, the average execution time was maintained at around 267 seconds when the size of the grid changed from 1 to 4. This phenomenon can be explained by the fact that the home host's resources being consumed by a single service was little such that agent migration was not triggered. Therefore, all the measured execution times were similar which represented the time required for a Task Agent to execute at the home host until it ended.

From Fig. 1, we can find that the curve for 50 services is the steepest among the three curves and it tends to be flattening slower. This shows that load balancing performs better when the home host is more overloaded. It is reasonable because a higher level of overloading will lead to a longer execution time which can lessen the effect of boundary execution time caused by the gradual load balancing formation.

Discussion. From this experiment, we can see that Mobile Grid Services are able to relieve the overloading problem and to make use of idle resources in the grid by employing their migration ability. In fact, the testing service used in this experiment is implemented in a general way. We can improve the service performance by specifically modifying the service in the implementation stage. By proper Task Agent implementation, the Task Agents in the service can be moved to other available hosts randomly at the beginning of their execution. This can avoid too many Task Agents (created by multiple service requests) working on a particular host initially. This procedure is suitable for the services which are expected to be requested heavily within a short period (i.e. a large number of Task Agents are created simultaneously).

4.2 MGS Overheads

A set of new Mobile Grid Services was used in this experiment. This experiment concentrated on the general overheads, migration overheads and message overheads of Mobile Grid Services in different modes.

A new service called Service X was implemented by using the MGS API. When the Service X was requested, a new Task Agent would be created. Monitor Agent would not be created such that no unexpected migration would occur to influence the result. At home host (Host A), the Task Agent would carry out 100000 times the prepared task. For each prepared task execution, 200 random integers were generated for doing summations with a variable and other 200 random integers were generated for subtractions. Meanwhile, it would send and receive five clear ACL messages (messages used by JADE mobile agents for communication) to and from an agent located in Host B. After that, the Task Agent would migrate to Host C and then Host D to carry out 100000 times the prepared task respectively. Finally, it would move back to Host A.

In the experiment, Service X setting up on Host A in normal mode was invoked. The time started from the Task Agent initialization to the end of execution was measured as the execution time of the service.

Service X was modified to a set of new services for measuring the execution times of the service under different settings. The new services were equal to Service X except for the specific differences below:

1. No migration.
2. No ACL message sending and receiving (including versions with and without migration).
3. Running in secure mode (including versions with and without migration).
4. Using signed message instead of clear message (including versions with and without migration).
5. Using encrypted message (including versions with and without migration).
6. The prepared task was changed from 200 times to 100 times of random integer summations and subtractions.
7. Service developed in standard Grid Service (GS) instead of Mobile Grid Service (MGS). The migration and message transmission were missing due to the absence of related supports.
8. Standard Grid Service version. The prepared task was changed from 200 times to 100 times of random integer summations and subtractions.

All the new services were executed separately and their execution times were recorded. For each service and setting above, 10 trials were performed. The average results of the experiments were used to estimate the general overheads, migration overheads and message overheads of Mobile Grid Services. The results will be shown as follows:

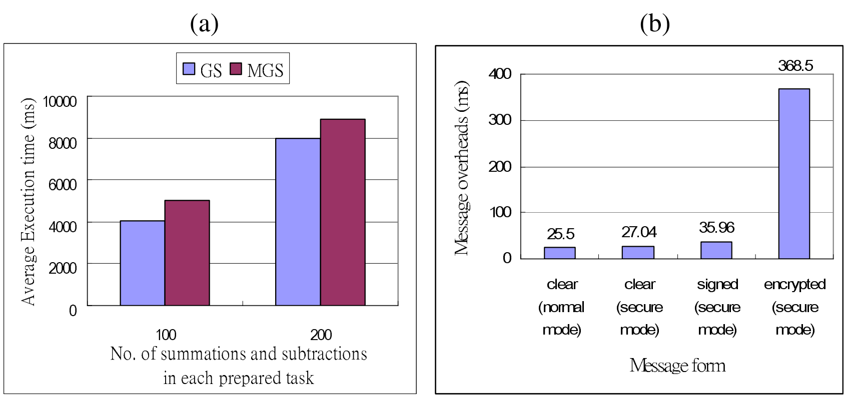


Fig. 2. (a) Execution times of Service X in form of standard Grid Service (standard GS) and Mobile Grid Service (MGS); (b) message overheads for different message forms and running modes

In Fig. 2(a), the execution times of the service without message transfer in the form of standard Grid Services (GS) and Mobile Grid Services (MGS) are compared. For both the prepared task consisting of 100 or 200 summations and subtractions, the MGS version requires a longer time to finish. This is reasonable as MGS has a general overhead caused by the agent creation and the additional works in agent execution (e.g. examining incoming messages and checking if stopping criteria are met). For 100 addition and subtraction, the MGS version executed 984.8 milliseconds (24.30%) longer than the standard Grid Service version. For 200 addition and subtraction, the MGS version executed 901.4 milliseconds (11.28%) longer than the standard Grid Service version.

We can find that the MGS overheads are similar for different prepared task contents. Actually, the MGS overheads are mainly contributed by the essential works in agent execution such as examining incoming messages and they are independent from the prepared task in each iteration. By taking the average value of the execution time's differences, we can estimate that the general overhead of Mobile Grid Services is about 0.003 milliseconds per iteration. For example, the overheads will contribute to 0.3% of the overall execution time if the task spends 1 millisecond in each iteration.

From the experimental results, we can also see that the task execution time in each iteration determines the proportion of the MGS general overheads in the overall execution. Longer execution in each iteration will make the MGS overheads less significant.

Fig. 2(b) shows the estimated message overheads in different running modes and message forms. The values are calculated by one-fifth of the execution time differences between the service without ACL messages transmission in normal mode and services under other settings (since five message sending and receiving were involved). The values stand for the costs of sending and receiving one ACL message under different modes.

From Fig. 2(b), we can see that secure mode does not bring significant overheads to the service handling clear ACL messages. The costs for clear message are very similar in normal and secure mode. Both values are less than 30 milliseconds and this shows that handling clear ACL messages is a light-weight operation. The overhead for signed messages (35.96 milliseconds) is higher but it is still not a heavy operation. Comparatively, the handling of encrypted messages introduces a much larger overhead. The cost for an encrypted message is 10 times higher than that for a signed message. However, it is reasonable as encryption and decryption are known as time-consuming processes. In fact, the message overheads above will be affected by the size of the ACL messages.

Fig. 3 shows the migration overheads of the Mobile Grid Services under different running modes and message forms. The migration costs represent the overheads of each agent migration in the Mobile Grid Services. They are calculated by one-third of the execution time differences between service with migration and without migration under the same setting (since three migrations were involved).

From the graph, we can observe that the migration costs are alike under different settings. By taking the average value, the overhead for each migration is about 330 milliseconds in normal mode while 410 milliseconds in secure mode. Migration

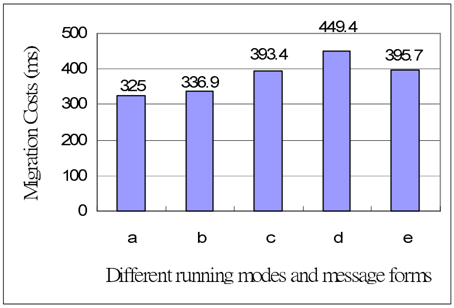


Fig. 3. The migration costs under different running modes and message forms, where a = normal mode without message transmission; b = normal mode with clear message transmission; c = secure mode with clear message transmission; d = secure mode with signed message transmission; e = secure mode with encrypted message transmission

introduces more overheads in secure mode than normal mode because extra secure process and checking is present such as authentication and agent permission checking. In fact, the migration cost will be affected by the size of the agent execution state. For example, an agent with a lot of instance variables should have higher migration overheads.

Discussion. From this experiment, we know the general overheads, migration overheads and the message overheads of the Mobile Grid Services. Although those overheads cannot be completely eliminated, service developers can maintain good service performance by appropriate Task Agent implementation (e.g. minimize the number of iterations, migration and message transmission in the Task Agent).

In the MGS Framework, we have tried to diminish the influence of the migration overheads by minimizing the number of migration in the automatic load balancing process. Unnecessary migrations are avoided by the “Gain” checking in the migration decision policy of the default Monitor Agent. This checking makes sure that the resource values (i.e. CPU, RAM and HD) of the decided host have a significant improvement (more than 5%) when compared with the current host. Otherwise, migration will not take place.

On the other hand, only part of Monitor Agents will receive resource data in each round of data distribution. This can prevent all Task Agents from migrating to the currently best host together and overloading that hosts. Unnecessary migrations due to this kind of collective migration can be avoided.

5 Related Work

There exist a number of grid development toolkits; however they support static services only. Globus [6] is the most popular one among them. It offers good resources for grid application development. It provides software services and libraries for resource monitoring, resource discovery, resource management, security, file management, fault detection, and portability. Nevertheless, it does not support service mobility. Up to now, none of the existing grid development toolkits support Mobile Grid Services nor provide complete security measures for them.

In [8] and [9], the authors present the formal definition of Grid Mobile Service (GMS) and some critical factors in GMS implementation. GMS is a concept similar to the Mobile Grid Services which aim to enhance the static Grid Services with mobility by making use of Mobile Agent technology. The GMS approach tends to extend and modify the OGSA standards, the WSDL and the SOAP protocols to achieve mobility while our design tends to maintain the Globus Grid service architecture (follows the OGSA standards) and make the framework more conformable to the existing and future Globus-based software. Unlike our framework, the implementation details of the GMS project are limited and security measures are absent.

Poggi *et al* [10] present another approach to realize a development framework for agent grid systems. This new approach is to extend agent-based middleware to support grid features instead of adding mobile agent support to grid middleware. In the paper, the authors present an extension of the JADE framework by introducing two new types of agent supporting intelligent movement of tasks as well as intelligent task composition and generation. This provides the basis for the realization of flexible agent-based grid systems. The project is still at an early stage and no framework is built yet.

6 Conclusion and Future Work

In this paper, we briefly describe an extension of the original stationary Grid Service - Mobile Grid Service. Besides, the middleware framework of Mobile Grid Services and the application programming interface for easy service development - MGS API are introduced. After that, the details of the experiments conducted for performance evaluation as well as their results and analysis are shown. From the experimental results, we can see that Mobile Grid Services are able to achieve load balancing in the grid and relieve the overloading problem. On the other hand, the estimation of the general overheads, migration overheads, message overheads and agent protection overheads are also presented.

In the future, the Resource Information Service and the Configurable Monitor Agent in the MGS API will be improved to provide better load balancing performance. Extra resource metrics will be studied and added to the resource information exchange system in the MGS framework such that the Monitor Agent can make appropriate migration decisions as soon as possible.

Acknowledgements

The work described in this paper was partially supported by the following grants: Research Grants Council of the Hong Kong Administrative Region (Project Number: CUHK4220/04E), and RGC Research Grant 2007-2008 Direct Allocation (Project Code: 2050406).

References

1. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid, Enabling Scalable Virtual Organizations. International Journal of Supercomputer Applications (2001)
2. Wong, S.-W., Ng, K.-W.: A Middleware Framework for Secure Mobile Grid Services. In: International Workshop on Agent based Grid Computing at 6th IEEE Interna-tional Symposium on Cluster Computing and the Grid (CCGrid 2006), May 16-19 (2006)
3. Wong, S.-W., Ng, K.-W.: MGS: An API for Developing Mobile Grid Services. In: OTM 2006. LNCS, vol. 4276, pp. 1361–1375. Springer, Heidelberg (2006)
4. Wong, S.-W., Ng, K.-W.: Security Support for Mobile Grid Services. The International Journal of Web Services Practices (IJWSP 2006) (2006)
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, Globus Project (2002)
6. Globus toolkit, `http://www.globus.org/`
7. Java Agent Development framework (JADE), `http://jade.tilab.com/`
8. Guo, S.-F., Zhang, W., Ma, D., Zhang, W.-L.: Grid Mobile Service: Using Mobile Software Agents in Grid Mobile Service. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai (2004)
9. Zhang, W., Zhang, J., Ma, D., Wang, B., Chen, Y.: Key Technique Research on Grid Mobile Service. In: Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004) (2004)
10. Poggi, A., Tomaiuolo, M., Turci, P.: Extending JADE for Agent Grid Application. In: Proceedings of the 13th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE 2004) (2004)

Are You Satisfied with Your Recommendation Service?: Discovering Social Networks for Personalized Mobile Services

Jason J. Jung[1], Kono Kim[2], Hojin Lee[3], and Seongrae Park[3]

[1] Yeungnam University, Korea
[2] SaltLux, Korea
[3] KTF, Korea
j2jung@intelligent.pe.kr, kono@saltlux.com,
{horilla, srpark}@ktf.com

Abstract. Most of recommendation mechanisms have been attempting to identify a set of cohesive user groups (or clusters) of which members in the same group might be interested in a common area (e.g., movies and musics) than others. Practically, this assumption is not working well, because the statistical analysis to extract simple demographic features (e.g., ages, genders, and so on) can not find out personal context in a certain situation, i.e., a more specific recommendation for each person. In order to solve this problem, we want to take into account social relations (particularly, kin relations) to identify each person. As aggregating the social networks, we can build a social network for making various social relations extractable. Most importantly, we are introducing our experiences on discovering social networks for providing personalized mobile services. Real customer information has been provided from KT Freetel (KTF), one of the major telecommunication companies in Korea. This work is an on-going research project for delivering personalized information to mobile devices via the social networks.

1 Introduction

To efficiently support personalized service, various types of information can be applied for modeling a target user's preference. One of well-known approaches, the so-called collaborative filtering [1,2], is to compare profiles of people. Such profiles are composed of ages, genders, occupation, and so on. Main assumption of this approach is that the more closer profile should be the more like-minded people. It means that two persons whose age are same are more probably interested in the same movie, rather than people who are older (or younger).

However, in real world, current personalized services have not shown efficient performance, people are not satisfied with the services at all. We think that most of the personalization mechanisms are simply trying to find out hierarchical clustering structure (this is very similar to the decision tree) identifying cohesive user groups of which members in the same group might be interested in a common area (e.g., movies and musics) than others [3]. This statistical analysis to extract simple demographic features by comparing user profiles [4] (e.g., ages, genders, and so on) can not find out personal

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 567–573, 2008.

context, i.e., what they are looking for in a certain situation. In other words, the personal recommendation for each user is supposed to be more specific.

In order to solve this problem, we mainly take into account two more conditions. Firstly, social affinity is regarded as a reasonable evidence to predict the personal context. For example, social relations (particularly, kin relations and friendships) can be assumed to identify each person's context more specifically. When he is looking for his father's birthday present, it is much more probable that he is looking for what his father wants than what he does. As aggregating the social networks, we can build a social network for making various social relations extractable. This social network can play a role of integrating multiple contexts which are inter-related between each other (e.g., parents and children).

Second condition is location, i.e., geographical position where you are. The personal context basically changes over time. In order to support better personalization, the service should be timely activated. For example, a user who is in a department store is expected to buy a certain product, rather than to go to restaurant.

More importantly, these two types of conditions can be merged for better personalized recommendation. Given two person who are i) linked (highly related) on a social network and ii) located in (or moving to) a close place, we can recommend very reasonable information to them. Especially, in this research project, we have been focusing on the mobile users joining KTF (KT Freetel) services. The problem is that the social networks are hidden. Thereby, we want to discover the hidden social network from usage patterns of mobile devices.

The outline of this paper is as follows. In the following Sect. 2, we will describe the problem on mobile communication. Sect. 3 and Sect. 4 show our heuristic approach to discover social networks among mobile users, and the whole system architecture, respectively. Sect. 5 will mention some of related work on personalization and building social network between mobile users. Also, we want to discuss some important issues. In Sect. 6, we draw a conclusion of this paper, and address the on-going and future work of this project.

2 Problem Description

To establish better personalization, we want to discover a social network from a given dataset. Generally, a social network is a graph-structured information.

Definition 1 (Social network). *A social network $\mathcal{S}$ is defined as*

$$\mathcal{S} = \langle \mathcal{N}, \mathcal{A} \rangle \tag{1}$$

where $\mathcal{N}$ and $\mathcal{A}$ are a set of participants $\{n_1, \ldots, n_{|\mathcal{S}|}\}$ and a set of relations between the participants, respectively.

In particular, $\mathcal{A}$ is simply represented as an adjacency matrix where

$$\mathcal{A}_{ij} = \begin{cases} 1 \text{ if } n_i \text{ links to } n_j; \\ 0 \text{ otherwise,} \end{cases}$$

and it is not necessarily symmetric, because we want to consider the directionality of the relations. For instance, the matrix is shown, as follows.

$$\mathcal{A} = \begin{array}{c} \\ n_1 \\ n_2 \\ \dots \\ n_{|\mathcal{S}|} \end{array} \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & \dots & n_{|\mathcal{S}|} \end{array} \\ \begin{pmatrix} 0 & 1 & \dots & 0 \\ 1 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & 0 \end{pmatrix} \end{array} \tag{2}$$

More importantly, in this work, the social network can contain multiple context together.

Definition 2 (Multiplex social network). *A multiplex social network $\mathcal{S}^+$ is defined as*

$$\mathcal{S}^+ = \langle \mathcal{N}, \mathcal{A}, \mathcal{C} \rangle \tag{3}$$

where $\mathcal{N}$ and $\mathcal{A}$ are the same components as normal social networks $\mathcal{S}$. Additional component $\mathcal{C}$ is a specified social relation attached to $\mathcal{A}$.

A simple example of multiplex social networks is shown in Fig. 1. Links in this social network are attached with six different contexts (e.g., *isSisterOf*, *isMotherOf*, and so on).

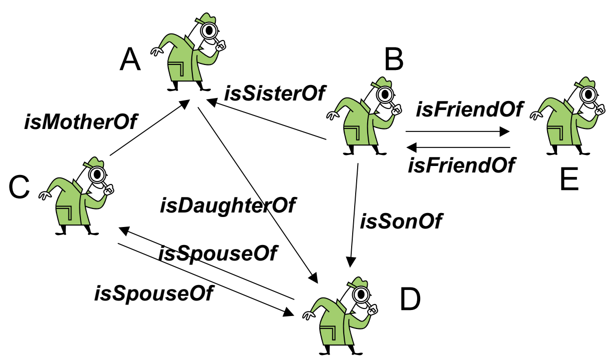


Fig. 1. A multiplex social network

Now, we want to explain about the datasets sampled from KTF legacy databases where raw records are stored. Mainly, it consists of three parts; i) registration profiles, ii) device (and service) specifications and iii) their calling patterns of over 60 thousand subscribers (mobile users). A set of fields of registration profiles are composed of

1. Name (first name and last name)
2. Social security number (partially encrypted)
   - Date of birth
   - Gender
   - Place of birth
3. Address
4. Job, Hobby, etc. (not provided).

As second part, the information about the devices and services are

1. Phone number
2. Device model
   - Communication protocol types (e.g., 2G, 3G)
   - Bell, sound, color, CDMA, GPS, KBANK types
3. Payment
   - Payment method (e.g., bank transfer, credit card, etc.)
   - Delegated person (e.g., name and social security number)
   - Payment creditability (e.g., history, postpone)

Third kind of information is indicating usage pattern of calling and SMS.

1. Calling
   - Calling number
   - Time of calling
   - Duration of calling
2. SMS
   - Receiver number
   - Time of Texting
   - Types of Texting (e.g., SMS and MMS)
3. Data communication types
   - Service types (e.g., CDMA/WCDMA, BREW/WIPI/VOD, DPP, RTSP, etc.)
   - Status of Charge (e.g., Start, Stop, Interim-Update, PPP Session Stop, and Accounting-On/-Off)
   - Amount of sent/received packets
4. Location
   - Scheme (e.g., GPS, CELL, and GPS+CELL)
   - Map viewer and map image formats (e.g., BMP, SIS, GML, and CGML)
   - Location information (e.g., X, Y, Z coordinations, etc.)

3 Heuristic Approach

These datasets are applied to predict social relations between mobile users by our heuristics. We have been tried to formalize the scenarios which are easily understandable to people in a common sense. Thus, each scenario can be matched to a set of social relations. Given a certain social relation, we have investigated as many cases as possible. To do so, we have built a decision tree by two ways of;

- interviewing with experts in KTF, and
- machine learning software packages (e.g., clementine and weka).

For example, in case of family relations such as *isFatherOf* and *isFamilyWith*, which are the most important social relation in this project, we can say that n_A is a father of n_B when;

- $(Payment(n_B) = n_A) \wedge (Age(n_A) - Age(n_B) \in [30, 50]) \wedge (Lastname(n_B) = Lastname(n_A))$

– $(Location(n_A, AtNight) = Location(n_B, AtNight)) \wedge (Age(n_A) - Age(n_B) \in [30, 50]) \wedge (Lastname(n_B) = Lastname(n_A))$

In addition, this logical expressions can be dynamically updated over time. More importantly, the social relations are semantically represented as concept hierarchy. For example, *isFatherOf* and *isBrotherOf* are subclasses of *isFamilyWith*. Thus, when the given information is not clear or enough, we can replace it to one of its superclass relations. (In fact, this issue is planned to work in near future.)

4 System Architecture

Our system is called NICE, which stands for New-generation Intelligent Content dElivery. The whole of system architecture is shown in Fig. 2.

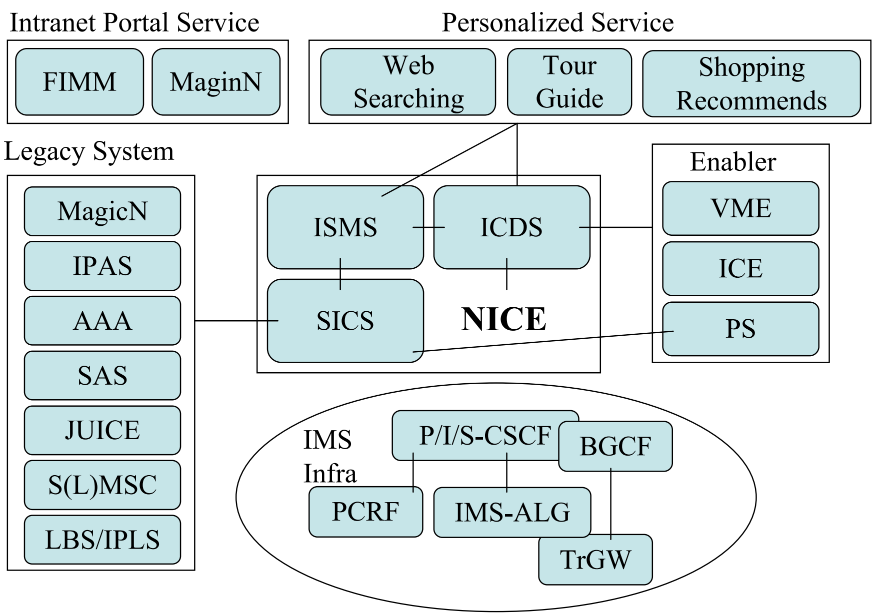


Fig. 2. System architecture of NICE

Our system is mainly divided into four parts. Legacy system is basically playing a role of data storage, as follows.

– IPAS - Internet Protocol Accounting Server
– AAA - Authorization, Authentication, Accounting
– SAS - Subscriber Analysis System
– JUICE - Storage database server for all subscriber's profiles
– S(L)MSC - Short (Long) Messaging Service Center
– LBS/IPLS - Location Based Server/Internet Protocol Location Server

As second component, IMS (IP Multimedia Subsystem) infrastructure can collect calling patterns from WCDMA, and be aware of charging status. The modules are shown, as follows.

- (P/I/S) CSCF - Call Session Control Function
- BGCF - Breakout Gateway Control Function
- PCRF - Policy & Charging Rule Function
- IMS & ALG - IP Multimedia Subsystem & Application Layer Gateway
- TrGW

These modules are related to fulfilling personalized service dynamically.

Third part is enabler for controlling the data streams.

- VME - Video Media control Enabler
- ICE - IMS Community Enabler
- PS - Presence Server

Finally, four part is main components of NICE system.

- ISMS (Intelligent Subscriber Management Server) can manage and support intelligent subscriber information cleansing (generation, storage and update), personalized social network builder, and subscriber status/location information.
- ICDS (Intelligent Content Deliver Server) can manage and support subscriber's log-on/-off, dynamic user interface, user sessions, SIP/HTTP/MSRP/RT(C)P protocol, personalized pop-up, intelligent web searching, and personalized web content.
- SICS (Subscriber Information Collection Server) can conduct subscriber's information collection/summarization/scheduling, subscriber information storage, and subscriber information transferring.

5 Related Work and Discussion

Most similarly, there have been two works. In [5], they have proposed a way of measuring the closeness between two persons to build a social network. Mainly they are focusing on the calling patterns, when (and how frequent) people are calling.

Also, in [6], their project "reality mining" has introduced experimental results for evaluating several hypothesis. These hypothesis has been compared to self-reports provided by the human users.

Personalization based on multi-agent systems has been introduced in MAPIS [7]. With regards to the business-oriented work, in [8], personalization process on e-commerce has been conducted by three modules; i) marketing strategies, ii) promotion patterns model, and iii) personalized promotion products. Especially, location-based personalization services has been implemented, e.g., NAMA [9].

6 Concluding Remarks and Future Work

This work is an on-going research project for building social network between mobile users. In this paper, we have described our research project for delivering personalized content to mobile users. In order to do so, we have introduced our heuristic approach to construct meaningful social networks between mobile users. Each context included in a social network has been combined with spatial context to better recommendation.

Future work, we are planning to put the calling patterns into the social network which has been built by expert's heuristics. It will make the social network more robust, and dynamically evolvable. Evaluation method has been also considered to verify whether the personalized service is reasonable or not. Finally, as proposed in [10], we have to implement a visualization functionality for better understandability.

Acknowledgement. This work was supported by the IT R&D program of MIC/IITA. (2007-S-048-01, Development of Elementary Technologies for Fixed Mobile IP Multimedia Convergence Services on Enhanced 3G Network).

References

1. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: Grouplens: applying collaborative filtering to usenet news. Communications of the ACM 40(3), 77–87 (1997)
2. Pazzani, M.J.: A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review 13(5-6), 393–408 (1999)
3. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: CSCW 2000: Proceedings of the 2000 ACM conference on Computer supported cooperative work, pp. 241–250. ACM, New York (2000)
4. Bonhard, P., Harries, C., McCarthy, J., Sasse, M.A.: Accounting for taste: using profile similarity to improve recommender systems. In: CHI 2006: Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 1057–1066. ACM, New York (2006)
5. Kazienko, P.: Expansion of telecommunication social networks. In: Luo, Y. (ed.) CDVE 2007. LNCS, vol. 4674, pp. 404–412. Springer, Heidelberg (2007)
6. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
7. Petit-Rozé, C., Strugeon, E.G.L.: MAPIS, a multi-agent system for information personalization. Information & Software Technology 48(2), 107–120 (2006)
8. Changchien, S.W., Lee, C.F., Hsu, Y.J.: On-line personalized sales promotion in electronic commerce. Expert Systems with Applications 27(1), 35–52 (2004)
9. Kwon, O., Choi, S., Park, G.: NAMA: a context-aware multi-agent based web service approach to proactive need identification for personalized reminder systems. Expert Systems with Applications 29(1), 17–32 (2005)
10. Jung, J.J.: Visualizing recommendation flow on social network. Journal of Universal Computer Science 11(11), 1780–1791 (2005)

SETNR/A: An Agent-Based Secure Payment Protocol for Mobile Commerce

Chung-Ming Ou[1] and C.R. Ou[2]

[1] Department of Information Management
Kainan University, Luchu 338, Taiwan
cou077@mail.knu.edu.tw
[2] Department of Electrical Engineering
Hsiuping Institute of Technology, Taichung 412, Taiwan
crou@mail.hit.edu.tw

Abstract. Non-repudiation of a mobile payment transaction ensures that when a buyer (B) sends some messages to a seller (S), neither B nor S can deny having participated in this transaction. An evidence of a transaction is generated by wireless PKI mechanism such that B and S cannot repudiate sending and receiving the purchase order respectively. SETNR/A protocol is proposed to improve the weakness of lacking non-repudiation mechanism from SET and SET/A for credit card-based transactions; on the other hand, agent-based protocol is ideal for complicated payment system. Broker generates a mobile agent for B which carries encrypted purchase order to S. A trusted third party (TTP) acts as a lightweight notary for evidence generations. One advantage of this agent-based payment protocol is to reduce inconvenience for mobile clients such as connection time and search for suitable merchant servers, etc.; it also provides necessary security mechanisms for mobile payment transactions.

1 Introduction

The security of credit card-based payment system has been a concerned issue for a long time. For example, credit card frauds which are performed by hackers' eavesdropping over transacting connections; on the other hands, dispute of a transaction could also jeopardize the mobile commerce [1].

In order to protect user's credit card information while transacting with payment systems, VISA and MasterCard, in association with major software and cryptography companies, developed SET (Secure Electronic Transaction) protocol [2]. One major advantage of SET is the separation of information disclosure, namely, ordering information and payment information. Merchant never knows credit card information; and financial institutes, which authorizes payment transaction, never knows ordering information.

Payment by credit cards is an attractive and efficient mobile payment; comparing to other cash-based and check-based payment systems, it is more consumer-based and suitable for mobile transactions [3]. However, for mobile payment systems, SET may be too demanding for limited computational capacity, slower connection speeds such

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 574–583, 2008.

as mobile handsets. This investigation leads to the research of agent-based applications. The digital signature-based authentication proves to be implemented efficiently within 3G communications [4]. One feasible idea is that mobile device simply sends purchase order out; it needs not to connect to base station afterwards. Later on, a message will be sent back to mobile device as a completion of purchasing. For example, Ramao and Silva [5] improved the SET protocol and proposed the agent-based SET/A protocol guided by SET rules for better performance and efficiency in e-commerce.

The purpose of non-repudiation is to collect, maintain, make available and validate irrefutable evidence concerning a claimed event or action in order to resolve disputes [6], [7]. However, neither SET nor SET/A considers non-repudiation mechanism for purchase orders. Therefore, neither cardholder nor merchant can obtain solid evidence for dispute resolutions. In this paper, our major goal is to improve SET/A by "embedding" some non-repudiation mechanism. This new protocol is named SETNR/A. The use of a broker between the wired and the wireless network can ease the access to web information from the mobile devices [8].

Mobile agents are considered to be an alternative to client-server mobile commerce systems. A mobile agent of the host is a set of code and data which can execute the code with the data as parameter in some trusted processing environment (TPE) or on some merchant hosts. However, there are several issues related to security and trust while considering mobile agent-based electronic commerce [9], [10], [11].

The arrangement of this paper is as follows. In section 2, we introduce preliminary knowledge for agent-based mobile payment protocols. In section 3, we propose an agent-based secure mobile payment protocol with non-repudiation mechanism; in section 4, we also analyze of the security mechanisms of this proposed protocol, namely, dispute resolutions.

2 Preliminary Knowledge of Designing Agent-Based Payment Protocol

2.1 Basic Structure for 3G Mobile Payment Services

The architecture for mobile payment system is composed of the following entities: a cardholder represented by mobile equipment (ME), WPKI, a merchant, a bank and a broker. These entities are also issued certificates by some certification authority. ME utilizes the USIM (Universal Subscriber Identity Module) to store mobile clients' information. ME is capable of verifying digital signatures to authenticate other entities, if necessary. We also deploy a middleware called the broker to help ME authenticate the merchant server such that attackers cannot impersonate this seller.

2.2 WPKI

The WPKI is the core cryptographic mechanism for non-repudiation protocol. There are two major WPKI operations in our agent-based payment protocol, one is the PKI encryption and decryption (key exchange); the other is the digital signature generation and verification.

A Certificate Authority (CA) issues certificates to mobile subscribers and server certificates to merchant servers, TTP, Home Revocation Authority (HoRA) and banks. A CA needs to provide certificate management service to ensure the validity of certificate. These entities would continue WPKI operations if and only if related certificates are valid.

We suggest that a mobile client be USIM-based 3G mobile equipment for efficient signature generation and verification. In this paper, we will use ***cardholder*** to represent mobile client; this means mobile client will pay items and services by his/her credit card.

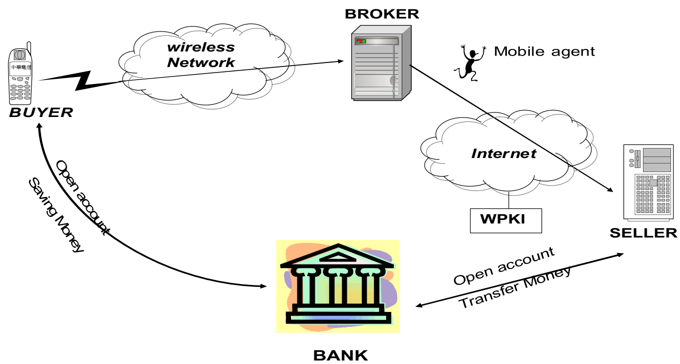


Fig. 1. The architecture of an agent-based mobile payment system

Trusted Third Party (TTP) is a notary server which simply generates necessary evidences for buyers and sellers. TTP needs to access CA's repository to retrieve necessary certificates and verify digital signatures. For those generated evidences, TTP will store these information in its public directory from which buyers and sellers may fetch evidences. TTP acts as a "lightweight" notary in this non-repudiation mechanism that only notarizes purchase order by requests. TTP also provides directory services accessible to the public.

Host Revocation Authority (HoRA) issues host certificates (HC) to merchant servers; these certificates bind mobile agent execution capability to the merchant host identity. When a merchant server acts maliciously, HoRA only needs to revoke this server's HC to prevent the broker from sending agents to it. The functionality of HoRA to detect the status of merchant servers can be referred to [11]. HoRA issues the host revocation List (HRL), which is a digitally signed list of revoked HCs.

2.3 Broker

Cardholder wants to pay the merchant according to the purchase order. The broker acts as a mediator between the cardholder in the wireless network and the merchant servers in the Internet, see Fig. 2. The broker must distinguish malicious servers from the honest ones according to HRL to avoid sending agents to them. HoRA will issue an updated HRL to the broker if a merchant server is detected to be malicious. Beside agent generations, Broker can verify certificates and digital signatures on behalf on agents.

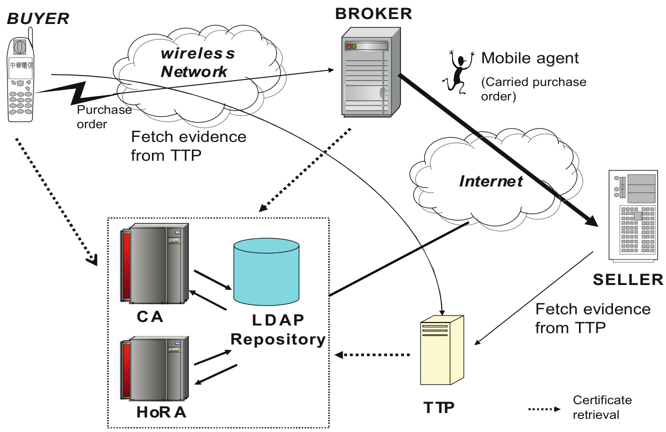


Fig. 2. Architecture of WPKI based Non-repudiation

2.4 SET Protocol

SET introduces an innovation of dual signatures to guarantee OI-PI pairs, where OI and PI represent ordering information and purchase information. It also provides safe and reliable business among merchants and cardholders over Internet. There are five participants in the SET protocol: cardholder (C), issuer, merchant (M), acquirer and payment gateway (PG). Dual signature and payment processing are out of the scope in this paper and [2] is an excellent reference. The SET protocol proceeds in four phases, where C, M and PG are involved in the second one:

1. Certificate Issuance: CA issues signature and key exchange certificates to transacting parties, if they had not been issued certificates before.
2. Purchase request: the cardholder purchase items or services from the merchant and supplies payment information to PG.
3. Payment Authorization: PG authorizes payment from cardholder to the merchant such that the latter could commit the transaction.
4. Payment Capture: the merchant captures funds being authorized by PG.

We concentrate on purchase request in this paper, which is described as follows (see Fig.3).

1. The cardholder issues the purchase request to the merchant.
2. The merchant returns a digitally signed reply with its signature certificate and the key-exchange certificate of the PG.
3. The cardholder verifies the certificates through CA and generates the OI-PI pair. PI is encrypted with a randomly generated symmetric key generated by cardholder; a digital envelope is generated by encrypting both (encrypted) PI and symmetric key together with PG's public key (Env_{PG}(PI, K)). This digital envelope is sent to the merchant with OI and cardholder's certificate.
4. The merchant verifies the cardholder's certificate and forwards the digital envelope to PG.
5. After PG has authorized the payment, the merchant generates a digitally signed purchase response then sends it to the cardholder.

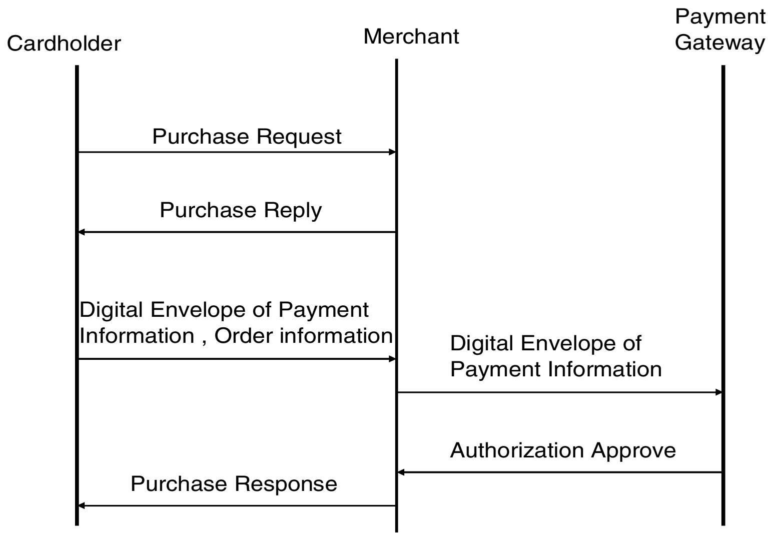


Fig. 3. Purchase Request in SET Protocol

3 Agent-Based Secure Payment Protocol with Non-repudiation Mechanism: SETNR/A

We improve SET/A by considering Broker as some trust verification center; on the other hand, adopting non-repudiation mechanism to SET/A. This new protocol is called SETNR/A. We adopt an efficient non-repudiation protocol proposed by Zhou and Gollmann where TTP acts as a lightweight notary (we name it ZGP) [12].

3.1 Agent-Based SET Protocol with Non-repudiation Mechanism (SETNR/A)

SET/A [5] is an implementation of the SET protocol on top of the mobile agent semantic. Yi et al. [13] suggested mobile agents sent to trusted verification center (for example, Broker) for signature verification to reduce the computational loading for agents.

SET/A can be improved for non-repudiation mechanism. Adding some trusted third party for dispute resolution in the improved protocol is one major consideration. The basic idea here is to embedding existing on-line TTP of ZGP for evidence generation before merchant sends PG payment information (more specifically, authentication request).

We now design the SETNR/A protocol for credit card payment system with non-repudiation mechanism. The notations are as follows.

PRequest	purchase request issued by C
PI:	payment information
OI:	order Information

PO:	purchase order being sent from C to M, which is equal to PI‖OI
DS:	dual signature for PO signed by C's signature private key
□ K:	symmetric key generated by C.
□ Cs(M):	signature certificate of M's
□ C_k(PG):	key-exchange certificate of PG
□ Env_{PG}(PI, K):	digital envelope of PI and K
CPO:	commitment for purchase order PO which is equal to e_K(PO) (e_K represents encryption by key K).
□ sS_C(PO):	signature of PO signed by C's private key.
□ L=H(PO,K):	a label linking PO and K (H represents a hash function).
□ f_i:	flag indicating the purpose of a signed message.
e_{TTP}(.),e_{BR}(.), E_{PG}(.):	encryption by TTP's, Broker's and PG's public keys respectively
EOO_CPO:	evidence of origin of CPO, which is equal to $sS_C(f_{EOO}$, PO, L, CPO).
EOR_CPO:	evidence of receipt of CPO, which is equal to $sS_M(f_{EOR}$, C, L, t_M, CPO).
*sub*_K:	authenticator of receipt of CPO, which is equal to $sS_C(f_{SUB}$, M, L, t_C, K, EOO_CPO).
*con*_K:	evidence of confirmation of K issued by the TTP with time stamp T, which is equal to $sS_{TTP}(f_{CON}$, C, M, L, T, t_C, t_M, K, EOO_CPO, EOR_CPO).

t_C is a time span defined by agent A(C) indicating that sub_K will be kept in TTP's private directory for t_C time units; t_M is a time span defined by merchant M indicating that TTP will keep EOR_CPO in its private directory for t_M time units. T is the time stamp indicating the actual time TTP generate key confirmation con_K and make it public. SETNR/A is as follows.

1.	C →Broker→$_{A(C)}$ M :	PRequest, e_{BR}(PO‖DS‖tc)
2.	M→A(C):	Cs(M), C_k(PG)
3.	A(C) →Broker :	Cs(M), C_k(PG)
4.	Broker→A(C):	verification response
5.	A(C) →M :	Cs(C), OI, Env_{PG}(K, PI)
6.	M→PG:	E_{PG}(K, PI)
	M→A(C):	Prequest_correct
7.	A(C) →M:	f_{EOO}, C, L, PO, EOO_CPO, t_C
8.	A(C) →TTP:	f_{SUB}, C, L, t_C, K, EOO_CPO, sub_K
9.	M→TTP:	f_{EOR}, C, L, t_M, EOO_CPO, EOR_CPO
10.	PG→ M, TTP : authorization	
11.	M→A(C):	response, Cs(M)
12.	TTP ← A(C), M:	f_{CON}, C, M, L, T, t_C, t_M, K, EOR_CPO, con_K

"C →Broker→$_{A(C)}$ M: message" means C sends message to broker, then broker will generate an agent A(C) for C; message will be carried by this agent to M; "TTP ←M" means M fetches messages from TTP. The basic idea is that C is able to send K, sub_K to TTP in exchange for con_K; on the other hand, M sends EOO_C, EOR_C and t_S to TTP. We describe each step in details as follows.

- In step 1, cardholder generates a request (PRequest, the same elements as in the original SET) which is sent to Broker using Broker's public key; Broker decrypts PRequest, generates A(C) which is delegated to TPE of M with carried message.
- In step 2, M then sends certificates Cs(M), C_k(PG) to A(C) in TPE.
- In step 3, A(C) returns to Broker's TPE and Broker verifies these two certificates.
- In step 4, Broker returns verification results.
- In step 5, A(C) generates digital envelope of PI and K by using PG's (key-exchange) public key, then sends it along with OI and Cs(C) to M.
- In step 6, M verifies the certificate and the dual signature on OI; if it is correct, then M forwards this digital envelope and request verification to PG and A(C) respectively.
- In step 7, A(C) sends M the message "f_{EOO}, C, L, PO, EOO_CPO, t_C" within the TPE of M.
- In step 8, M needs to verify EOO_CPO by retrieving C's (signature) public key from the corresponding CA's repository. EOO_C is saved as an evidence of origin for M. On the other hand, A(C) sends TTP the message "f_{SUB}, C, L, t_C, K, EOO_CPO, sub_K " within TPE of the TTP's.
- In step 9, TTP keeps sub_K in its private directory after the verification of digital signature of C; TTP will delete it after t_C time units or until con_K is generated and published. M sends "f_{EOR}, C, L, t_M, EOO_CPO, EOR_CPO" to TTP. After receiving EOO_CPO, EOR_CPO and t_S from M, TTP needs to verify EOR_CPO using M's (signature) public key and compare EOO_CPO with the one sent by A(C) in step 7. If either one is not true, TTP concludes that at least one party is cheating and it will not generate con_K. We call {f_{CON}, C, M, L, T, t_C, t_M, K, EOR_C, con_K} the evidence of this purchase order PO.
- In step 10, TTP requests authorization from PG; then PG sends authorization to M and TTP
- In step 11, M sends authorization response to A(C) within its TPE, A(C) then verifies the certificate Cs(M) and the corresponding signature.
- In step 12, A(C) can return to C's mobile device from TTP with carried evidence. M can also start evidence collection from TTP. C obtains K and con_K; M fetches con_K from TTP to prove that K is available for M.

4 Security Issues

In this section, we analyze security of SETNR/A protocol from cardholder's point of view, namely, payment security and non-repudiation security.

4.1 Payment Security

In a mobile payment system, one major concern is to protect the mobile client's sensitive data such as credit card information, purchase information, etc; this is exactly what SET protocol has achieved. On the other hand, PI, which includes credit card information, will be encrypted by a symmetric key generated by cardholder's agent and protected by digital envelope along with this key. This digital envelope is

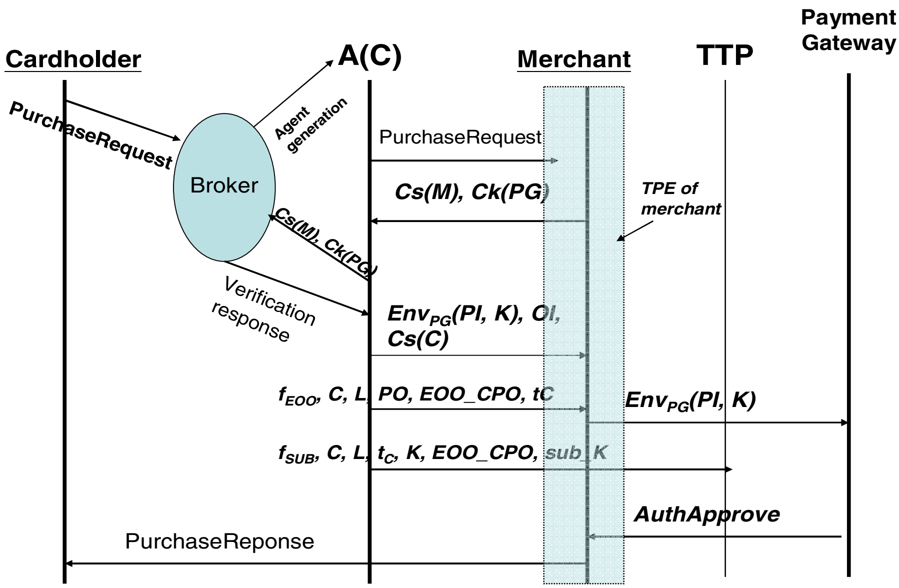


Fig. 4. Purchase request Transaction in SETNR/A.

encrypted by PG's public key-exchange key. Therefore, only PG can decrypt this envelope and then decrypt PI by using K.

SETNR/A, similar to SET/A, suggests running agents in TPEs, which are tamper-proof environment or secure coprocessors to protect agents against malicious merchants. The following table shows us information revelation to participants.

Table 1. Information revelation to participants

	PI	OI	EOO_CPO	EOR_CPO	K
Cardholder	YES	YES	YES	YES	YES
Broker	NO	NO	YES	YES	NO
Agent	NO	NO	YES	YES	NO
Merchant	NO	YES	NO	YES	YES (after NR is complete)
Payment Gateway	YES	NO	NO	NO	YES
TTP	NO	NO	YES	YES	YES

4.2 Security of Non-repudiation Mechanism

The most important security issue of a non-repudiation protocol is the dispute resolution. We analyze the generated evidences of step 12 in the above SETNR/A protocol, dispute resolution mechanisms of C and M to see whether non-repudiation can be reached. A trusted arbitrator will help solve disputes according to submitted evidences.

Dispute of Cardholder

When C denies having sent purchase order PO to M, M may present EOO_CPO, EOR_CPO and con_K to the arbitrator in the following way:

M→ arbitrator: EOO_CPO, EOR_CPO, con_K, sS_C(EOO_CPO, EOR_CPO, con_K), L, K, PO, CPO

The arbitrator first verifies the signature of C, sS_C(EOO_CPO, EOR_CPO, con_K); if the verification is positive, the arbitrator checks the following five steps:

step 1: if EOO_CPO is equal to $sS_C(f_{EOO}$, M, L, CPO).
step 2: if EOR_CPO is equal to $sS_M(f_{EOR}$, C, L, t_M, CPO).
step 3: if con_K is equal to $sS_{TTP}(f_{CON}$, C, M, L, T, t_C, t_M, K, EOO_CPO, EOR_CPO).
step 4: if L is equal to H(PO,K).
step 5: if PO is equal to dK(CPO).

If step 1 is checked positive, this arbitrator concludes that C has sent M the encrypted purchase order CPO. If step 2 is checked positive, arbitrator concludes that M has sent all the correct payment information to TTP. For all 5 steps being checked positive, this arbitrator finally concludes that C has sent M the purchase order PO, which is encrypted by K and presented to be CPO.

Dispute of Merchant

When M denies receiving the purchase order PO from C, C may present EOO_C, EOR_C, con_K to the arbitrator in the following way:

C→ arbitrator : EOO_CPO, EOR_CPO, con_K, sS_C(EOO_CPO, EOR_CPO, con_K), L, K, PO, CPO

The arbitrator first verifies the signature of C, sS_C(EOO_CPO, EOR_CPO, con_K); if the verification is positive, the arbitrator checks all five steps same as those in the dispute of origin. For all five steps being checked positive, arbitrator concludes that M has received PO, which is encrypted by K and presented to be CPO.

5 Conclusions

We propose an improved mobile agent-based credit card payment system, namely, SETNR/A, by "embedding" a non-repudiation mechanism to SET/A. Mobile payment transactions need time information included in evidences for dispute resolutions. The broker generates a mobile agent for cardholder which carries this encrypted purchase order to the merchant. The advantage of this agent-based payment protocol is to provide a convenient way for mobile clients to reach non-repudiation for mobile payment transactions.

References

1. Zhou, J., Deng, R., Bao, F.: Evolution of Fair Non-repudiation with TTP. In: Pieprzyk, J.P., Safavi-Naini, R., Seberry, J. (eds.) ACISP 1999. LNCS, vol. 1587, pp. 258–269. Springer, Heidelberg (1999)
2. Stallings, W.: Cryptography and Network Security, 3rd edn. Prentice Hall, Englewood Cliffs (2003)
3. Lee, T., Yip, Y., Tsang, C., Ng, K.: An Agent-Based Micropayment System for E-Commerce. In: Liu, J., Ye, Y. (eds.) E-Commerce Agents. LNCS (LNAI), vol. 2033, pp. 247–263. Springer, Heidelberg (2001)
4. WPKI Implementation, Initial stage, Testing and Experiments (Internal Report and Discussion), Chunghwa Telecom Lab, Taiwan (2004)
5. Romao, A., Silva, M.: An Agent-based Secure Internet Payment System for Mobile Computing
6. ITU-T, Recommendation, X.813: Information Technology-Open Systems Interconnection-Security Frameworks in Open Systems. Non-repudiation Framework (1996)
7. Li, B., Luo, J.: On Timeliness of a Fair Non-repudiation Protocol. In: InfoSecu 2004, November 14-16, pp. 99–106 (2004)
8. Esparza, O., Munoz, J., Soriano, M., Forne, J.: Secure Brokerage Mechanisms for mobile Electronic Commerce. Computer Communications 29, 2308–2321 (2006)
9. Pagnia, H., Vogt, H., Gartner, F., Wilhelm, U.: Solving Fair Exchange with Mobile Agents. In: Kotz, D., Mattern, F. (eds.) MA 2000, ASA/MA 2000, and ASA 2000. LNCS, vol. 1882, pp. 57–72. Springer, Heidelberg (2000)
10. Wilhelm, U., Staamann, S., Buttyan, L.: On the Problem of Trust in Mobile Agent Systems. In: Symposium on Network and Distributed System Security, pp. 114–124. Internet Society (March 1998)
11. Esparza, O., Munoz, J., Soriano, M., Forne, J.: Host Revocation Authority: A Way of Protecting Mobile Agents from Malicious Hosts. In: Cueva Lovelle, J.M., Rodríguez, B.M.G., Gayo, J.E.L., Ruiz, M.d.P.P., Aguilar, L.J. (eds.) ICWE 2003. LNCS, vol. 2722, pp. 289–292. Springer, Heidelberg (2003)
12. Zhou, J., Gollmann, D.: A Fair Non-repudiation Protocol. In: Proceedings of 1996 IEEE Symposium on Security and Privacy, Oakland, California, pp. 55–61 (May 1996)
13. Yi, X., Wang, X., Okamoto, E.: A Secure Agent-based Framework for Internet Trading in Mobile Computing Environment. Distributed and Parallel Database 8, 85–117 (2000)

A Key Management Scheme Combined with Intrusion Detection for Mobile Ad Hoc Networks

Yingfang Fu[1], Jingsha He[2], Liangyu Luan[3], Guorui Li[1], and Wang Rong[2]

[1] College of Computer Science and Technology Beijing University of Technology Beijing 100022, China
{fmsik, liguorui}@emails.bjut.edu.cn
[2] School of Software Engineering Beijing University of Technology Beijing 100022, China
jhe@bjut.edu.cn, godtroop@emails.bjut.edu.cn
[3] College of Applied Science, Beijing University of Technology, Beijing 100022, China
luanly@bjut.edu.cn

Abstract. A mobile Ad Hoc network is a multi-path autonomous system comprised of many mobile nodes with wireless transmission capability. Although there is a trend to adopt ad hoc networks for commercial uses due to their unique properties. One main challenge in design of these networks is their vulnerability to security attacks. In this paper, we first review current research in key management in mobile Ad Hoc networks. Then by considering the characteristics of such networks in which free movement of mobile nodes can lead to frequent topological changes, especially network separation and convergence, we propose an key management mechanism based on a combination of techniques, such as hierarchical topology structure, virtual CA (certification authority), offline CA, certificate chain, and proactive measurement, and show that the proposed scheme could improve the security of mobile ad hoc network in the areas of expandability, validity, fault tolerance and usability.

Keywords: Ad Hoc networks, hierarchical topology structure, distributed intrusion detection, key management.

1 Introduction

A mobile Ad Hoc network is a wireless network with the characteristics of self-organization and self-configuration so that it can quickly form a new network without the need for any wired network infrastructure. Mobile Ad Hoc networks not only play an important role in military applications, but also have a variety of commercial applications such as disaster recovery. On the other hand, compared with the wired and normal wireless networks, mobile Ad Hoc networks are more prone to security threats and attacks, e.g., passive interception, data interpolation and replaying, identity forgery, and denial of service. That is because mobile Ad Hoc networks usually operate in a wide open space and their topologies change frequently, which makes such networks lack of centralized mechanisms for protection and clear lines for defense. Therefore, Security is crucial for mobile Ad Hoc networks. Current research in security for mobile Ad Hoc networks is mostly focused on routing protocols, key management

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 584–593, 2008.

and intrusion detection. On one hand, while these mechanisms are not sufficient by themselves, e.g., many routing protocols for Ad Hoc networks are designed with the assumption that all communication between mobile nodes is secure, which is not true in general because there are a lot of potential threats in such a wireless environment, e.g., modification of routing protocols and forgery of IP addresses. Consequently, key management plays a central role in network security for ensuring confidentiality, integrity and non-repudiation of information and intrusion detection provides the necessary means to fight against attacks launched by compromised nodes. On the other hand, these mechanisms developed for wired networks cannot be directly applied to mobile Ad Hoc networks due to different characteristics of these networks. In this paper, we propose a new key management scheme using such techniques as hierarchical topology structure, virtual CA, offline CA, certificate chain and proactive measurement. We show that the proposed scheme would improve the security of mobile ad hoc network in the areas of expandability, validity, fault tolerance and usability.

Several results related to the key management described in this paper have already been published in [1], but the paper is only loosely related to the key update and revocation problems addressed in this paper. The rest of this paper is organized as follows. In the next section, we review some related work in key management for mobile Ad Hoc networks and the hierarchical cluster algorithm. In Section 3, we present a new key management scheme along with a thorough analysis. Finally, in Section 4, we conclude this paper with a discussion on our future work.

2 Related Work

We review some previous work in key management in mobile Ad Hoc networks in this section. In general, we don't assume that there is a trustworthy node in a mobile Ad Hoc network. Therefore, we don't presumably have a single node to act as the Certificates Authority (CA) in key management. In this section, we describe some key management schemes and point out their limitations.

2.1 Partial Distributed Key Management Scheme

Zhou proposed a threshold key management scheme based on public key encryption [2]. The method uses a trusted offline organization and the (k, n)-threshold scheme to protect the private key. The offline organization would issue a certificate to a mobile node when it joins the network and generate a public key and a private key for the system, in which the public key is distributed to all the mobile nodes in the network. The secret of the private key is, however, shared by n serving nodes, which are randomly selected in the network by the offline organization. The n serving nodes would then manage the certificates according to the (k, n)-threshold scheme [2]. Therefore, the network can withstand the situation in which at most k-1 serving nodes are compromised without compromising the system security with respect to key management. Each serving node has the public key of every other mobile node in the network and can thus communicate with all the nodes securely. But this scheme has the following shortcomings: (1) Since the method randomly selects n nodes as the serving nodes, it could make the system less stable [3, 4]. (2) Since each serving node stores the public keys of all the other mobile nodes, it would need a large amount of memory space. (3) Some distributed CA schemes don't provide the means for revoking certificates. (4) The method doesn't have corresponding means to deal with network separation caused by the mobility and loss of nodes. Therefore, the method can hardly support network expandability. (5) When the network gets very large, it would be difficult, if not impossible, for the serving nodes to get the public keys of all the other nodes. Consequently, this scheme is not suitable for large networks.

2.2 Self-organizing Key Management Scheme

The self-organizing key management scheme proposed by Hubaux et al [5, 6] uses a certificate chain graph to implement key management. There is no centralized CA and every node creates its own private key. The distribution of the public key certificate is purely the business of the node itself. For two nodes, say, nodes A and node B, if A trusts B, A could issue a certificate to B that contains B's public key encrypted with A's private key. In addition, there is no centralized certificate directory and every node has its own certificate directory that contains a list of certificates including those that the node has sent to other node, those that the node has received from other nodes, and those that have been created using the Maximum Degree Algorithm [5]. The trust relationship between nodes can be described using a certificate chain graph. In the graph, a vertex represents the public key for a node while an edge represents a certificate issuance for the public key. Assuming that K_A and K_B are the public keys of node A and node B, respectively, if A has issued a certificate to B, there would be a directed edge from A to B in the certificate chain graph. If two nodes want to certify each other, they would merge the two certificate directories and look for the certificate chain in the certificate chain graph that connects A and B. The method doesn't have any special requirement on the mobile nodes and every node is completely independent. Therefore, it is more suitable for Ad Hoc networks with the characteristics of self-organization. Nonetheless, the method has the following shortcomings: (1)The certificate chain that connects two nodes may not be found. (2)Without a trusted organization in the system to provide guarantee, the method would make it less trustable for a long certificate chain. (3)Since the method depends on the trust relationship between nodes, vicious node may destroy the network security through forging large numbers of certificates. (4)The method doesn't provide the means for revoking certificates.

2.3 Identity-Based Key Management Scheme

The identity-based key management scheme proposed by Khalili et al [7] uses node identification and system public key to implement key management, in which the system public key is generated and broadcast in the network by n special nodes, namely the PKG nodes. The PKG nodes also issue certificates using the (k, n)-threshold secret sharing scheme [5]. Node identification relies on unique IDs such as node names, postal addresses or Email addresses. A node doesn't have a specific public key and would obtain a private key from PKG nodes. In order to obtain a private key, a node needs to contact at least k PKG nodes, provides its identity and related information, gets partial private key from each such node, and compose its own private key. The node identity consists of a random string of characters, which would make it unpredictable, and the PKG nodes issue an exclusive private key for every node. The method could reduce the complexity of computation as well as communication cost because each node would not have to create its own public key and broadcast it in the network. But the method has following two shortcomings: (1)A malicious node may pretend to be a PKG node and forge a public key to a new node, while possessing a corresponding private key. (2) The method doesn't provide any means to deal with key update.

2.4 Hierarchical Cluster Algorithm

To deal with node mobility and improve the stability of cluster frame, we could assign a weight value to each node based on its mobility. The more frequently a node moves, the lower its weight value will be. A cluster head is the node with the largest weight value. In the hierarchical cluster algorithm [8], node mobility is expressed in terms of relative mobility by comparing the strengths of two signals that the node receives from its neighbors. Define the relative mobility index of node y relative to node x using the formulas (1) below: in which $Rxpr^{new}_{x\rightarrow y}$ represents the power that node y

$$M_y(x) = 10\log\frac{Rxpr^{new}_{x\rightarrow y}}{Rxpr^{old}_{x\rightarrow y}} \tag{1}$$

receives from node x and $Rxpr^{old}_{x\rightarrow y}$ represents the power that node y received from node x last time. If $M_y(x)<0$, it means that the distance between two nodes becomes longer, otherwise it becomes shorter. Through computing the mean value of the absolute values of the mobility between node y and all the other neighboring nodes x_i ($i\in m$) , The local mobility value of node y can be obtained using the formulas(2).

$$M_y = \frac{\sum_{i=1}^{m} |M_y(x_i)|}{m} \tag{2}$$

The less the value M_y is, the lower the relative mobility of the node in relative to all the other neighboring nodes.

3 The Key Management Scheme for Mobile Ad Hoc Networks

3.1 The Network Model

The whole network consists of one or more sub-networks. We call each such sub-network a cluster. Each cluster will have a head node, n-1 PKG nodes and many other ordinary cluster nodes. The cluster heads in turn form the next layer of the network, and so on. Nodes can roam form one cluster to another. Nodes in a lower layer network, e.g., in a cluster, communicate within a relatively shorter range and those in a

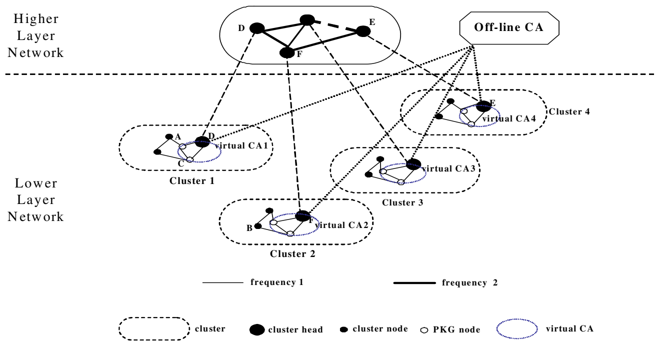


Fig. 1. The Network Model

higher layer network communicate within a relatively longer range. Let's assume without loss of generality that cluster nodes communicate with each other in frequency 1, the cluster heads communicate with each other in frequency 2, and so on. There is an off-line CA in the network that only connects to the network under the condition that it is notified of the existence of an attacker, whether the attacker is a cluster node, a PKG node or a cluster head. One of the main functionalities of the off-line CA is to appoint the cluster heads using the hierarchical cluster algorithm [8]. The CA also performs functions such as keeping track of the status of the network and collecting reports from cluster heads For each cluster, the cluster head selects n-1 cluster nodes as the PKG nodes that are generally high performance nodes in terms of security, data storage and computation power. These PKG nodes would then form a virtual CA along with the cluster head. The cluster head is responsible to detect topology change within the cluster as nodes join and leave the network or are destroyed. When necessary, the cluster head would make a connection to the off-line CA although most of the time they are not connected Fig. 1 illustrates the network model in our key management scheme.

3.2 **Key** Creation

Key creation is illustrated in Fig. 2, which involves nine steps:

(1) The off-line CA creates a pair of private and public keys for each cluster and generates a certificate.
(2) Before joining the network, every cluster node must register with the off-line CA.
(3) The off-line CA needs to identify a new cluster node and then assigns it a cluster ID and node ID.
(4) The off-line CA selects one cluster node from among all the cluster nodes according to the hierarchical cluster algorithm and appoints it as the cluster head. The cluster head is also granted the right to access the cluster nodes information database.
(5) The cluster head broadcasts the public key for the cluster to all the other cluster nodes. The cluster head selects n-1 cluster nodes as the PKG nodes. PKG nodes are generally high performance nodes in terms of security, data storage and computation power. The cluster head then forms a virtual CA together with the n-1 PKG nodes.

(6) The cluster head and the n-1 PKG nodes manage the keys using the (k, n)-threshold cryptographic method. That is, the private key for the cluster is partitioned and held by the cluster head and the n-1 PKG nodes. To reconstruct the private key, any k out of the n-1 PKG nodes would have to submit partial private keys to the cluster head.
(7) The cluster head could compute a new private key for the cluster based on its own private key, the old private key for the cluster and all the partial cluster private keys submitted to it by the k PKG nodes.
(8) The cluster head would create the private key for a cluster node based on the cluster private key and grants it to the cluster node.

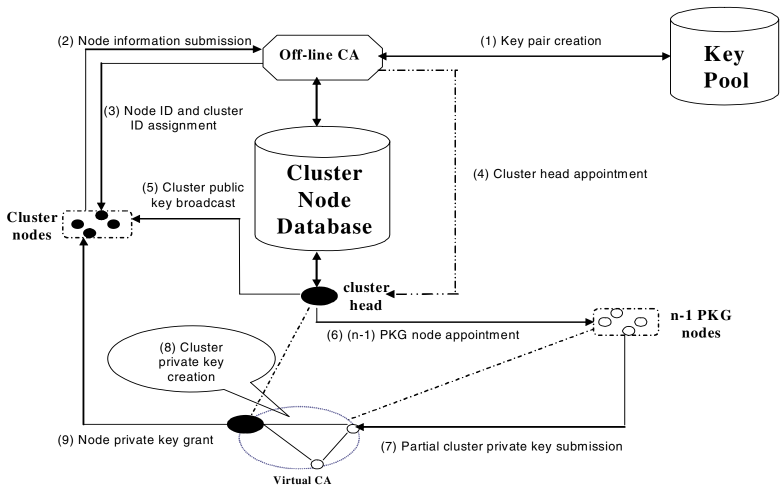


Fig. 2. Key Creation

3.3 Key Update and Revocation

Cluster nodes can roam from one place to another in the network, resulting in the need that an existing key have to be replaced by a new one. There are two reasons for it. One is that when a cluster node joins the network, it should not be allowed to access information transmitted in the past. The other is that when a cluster node drops out of the network, it should not be allowed to access information to be transmitted in the future any further. To detect the abnormal behavior of a node in real time, which occurs when a new node joins the network, an existing node drops out of the network or is destroyed, we use a distributed intrusion detection scheme to monitor and collect information about the status of every node in the cluster. Fig. 3 illustrates the IDS structure in our scheme, in which there are two components for each IDS: IDSBH (Intrusion Detection System based on Host) and IDSBN (Intrusion Detection System based on Network). The IDSBH runs on all ordinary cluster nodes and consists of six modules: Data Collection (DC), Local Association and Aggregation Engine (LAAE), Local Detection

Engine Based on Pattern (LDEBP), Local Detection Engine Based on Abnormality (LDEBA), Host Response (HR) and Local Security Communication (LSC). For the six modules, DC, LDEBP and LDEBA are always running and the other three can be in two operational modes: running and waiting. When LDEBP or LDEBA detects an intrusion, the modules in the waiting state will be activated. That is to say, they will change from the waiting state to the running state. The IDSBN also has two operational modes: running and waiting. If IDSBH cannot determine an intrusion easily, it will run on the n-1 PKG nodes and the cluster head node. The IDSBN also consists of six modules: Distributed Voting (DV), Global Response (GR), Global Security Communication (GSC), Global Association and Aggregation Engine (GAAE), Global Detection Engine Based on Pattern (GDEBP) and Global Detection Engine Based on Abnormality (GDEBA). In our scheme, intrusion detection relies on the host as well as the network modules. While the former employs detection techniques based on pattern and abnormality, the latter adopts distributed voting scheme in addition to techniques used by the former. Fig. 4 illustrates the ten steps for the detection of an intrusion.

(1) The DC module collects two types of audit data: network data and journal information (such as grouping and routing information). The audit data will then be sent to both the LDEBP and the LDEBA modules for analysis and intrusion detection. There may be multiple DC modules in each IDSBH to monitor different data sources.
(2) The LDEBP and LDEBA modules will detect the abnormal behavior of nodes based on the audit data they receive from the DC modules. While the former detects abnormal behavior of nodes by comparing the behavior of nodes with patterns in the database of attacks, the latter does so by comparing the behavior of nodes with the descriptive normal behaviors of nodes. For example, we use the Drop_Bandwidth and the Send_Bandwidth thresholds to detect the transmission of abnormal packets [9]. If a node takes a bandwidth which is lower than the Drop-Bandwidth threshold to drop packets or one which is less than the Send-Bandwidth threshold to transmit copied packets, we record the node in the normal behavior database.
(3) The LAAE module will aggregate the results from the LDEBP and the LDEBA modules with appropriate algorithms and will then send the combined results to the HR module.

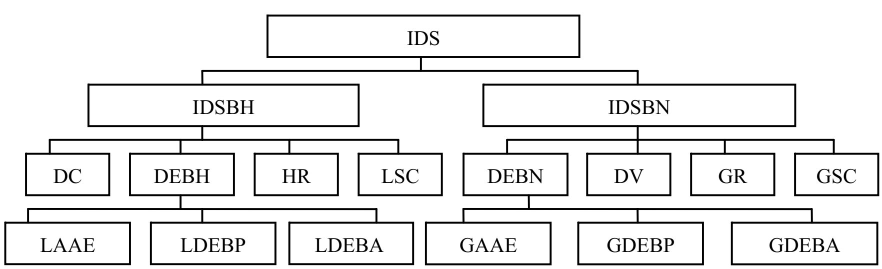


Fig. 3. The IDS structure

(4) If a node in question continues to behave abnormally, the HR module will make a unilateral decision to disassociate itself with the abnormal node and also send the intrusion detection results to the LSC module.
(5) The LSC module will the encrypt intrusion detection results with the identity-based cryptosystem scheme [6] and send it to the GAAE module in the cluster head. If the node suspects that its cluster head is an attacking node, it will only send the alarm information to the GAAE module of its nearest PKG node.
(6) The GAAE module in the cluster head will aggregate the detection results received from all the other hosts in the same cluster and send the combined results to the GDEBP and the GDEBA modules.
(7) The cluster head will invoke the GDEBP and the GDEBA modules to determine the type of intrusion and the result will be sent to the GR module. If the intrusion detection results cannot be easily determined, the cluster head will initiate the voting scheme in (7').
(7') If the intrusion detection results of a node is sent to the GAAE module of its nearest PKG node, the PKG node will initiate the voting scheme in which all PKG nodes will aggregate its own detection results with those from other PKG nodes to determine the type of intrusion based on results from the GAAE module. If a majority of the PKG nodes determine through voting that a node is an attacking node, the PKG node that is closest to the attacking node will send the intrusion detection result to the GR module.
(8) The GR module that receives the final intrusion detection result will keep the attacking node away from the network and will also send the intrusion detection result to the GSC module.
(9) The GSC module will encrypt the information about the attacking node with identity-based cryptosystem scheme [8] and broadcast it to all the other nodes in the same cluster as well as to the offline-CA.
(10) When the off-line CA receives any information regarding nodes' joining or dropping out of the network, or intrusion results, it would connect itself to the network and update the keys in different scenarios. In the first scenario, the change is caused by an ordinary cluster node. The off-line CA would then update the key pair for the cluster with the cluster head. The cluster head would then announce the newly joined node to all the cluster nodes and update the private key for the cluster as well as that for each cluster node together with the PKG nodes. In the second scenario, the change is caused by a PKG node. Then, in addition to updating the private key for a cluster, the cluster head must select new PKG nodes from the remaining cluster nodes and would then update the private keys accordingly. In the third scenario, the change is caused by the cluster head. Then, the off-line CA not only updates the key pair for the cluster, but also selects a new cluster head from the remaining cluster nodes. The newly selected cluster head would then select n-1 PKG nodes from the remaining cluster nodes and update the keys accordingly. To reconstruct the private key, the cluster head would have to work with at least k PKG nodes. To prevent attacks, the off-line CA, together with the virtual CA, would update the private key for each cluster as well as for each cluster node periodically. Therefore, each and every private key has a timestamp associated with it, which limits the validity of the private key.

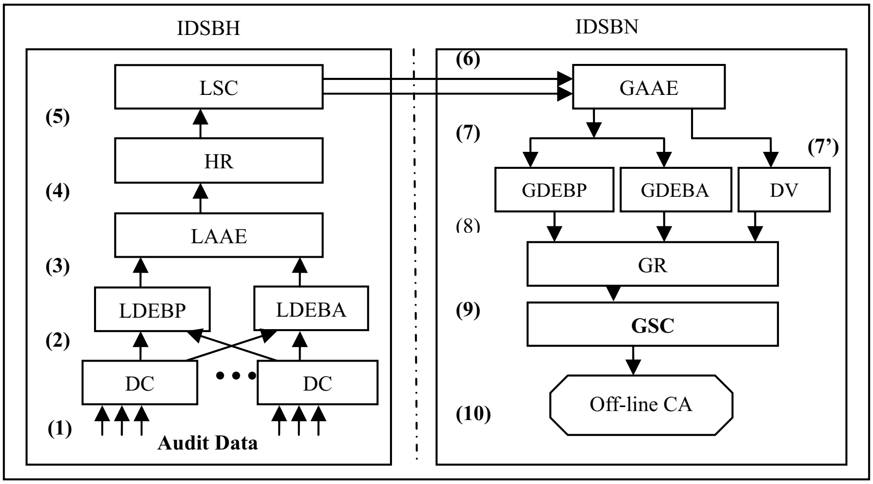


Fig. 4. The intrusion detection process of IDS

4 Conclusion

In this paper, we represent a key management combined with intrusion detection for mobile Ad Hoc networks. By analyzing and identifying the advantages and limitations of existing key management approaches, we can see that our scheme exhibits the following advantages:

(1) Compared to the scheme presented in [2], our scheme uses a hierarchical topology structure that can be easily extended to deal with mobile Ad Hoc networks of any sizes. By using identity-based cryptographic means and the threshold scheme, cluster nodes don't need to create and broadcast their own public keys in the network, which helps to improve performance by taking less network bandwidth and storage space. The cluster heads and cluster nodes keep track of and collect status of the clusters with proactive measurement techniques, which could improve key management in dealing with continuous topological changes. PKG nodes are generally high performance nodes in terms of security, data storage and computation power, which could make the system more stable.

(2) Compared to the scheme presented in [2,5,7], in our scheme, every cluster has its own public key that is known only to all the cluster nodes in the same cluster and all the other cluster heads; A new cluster private key is computed based on the old cluster private key and the partial private keys created by the k PKG nodes. The cluster public key, the cluster private key and the private key of cluster nodes all have the timestamp in them. These characteristics help to improve the validity and security in key management.

(3) The off-line CA connects to the mobile Ad Hoc network only under certain circumstances, which improves not only the trustworthiness of key management, but also its security. When cluster heads communicate with each other, keys can be managed with the certificate chain technique and the off-line CA, which improves key management in the areas of fault tolerance, serviceability and availability.

(4) For a small network in which cluster nodes trust each other, the certificate chain technique would make the cluster heads communicate more conveniently. For a large network in which cluster nodes don't necessarily trust each other, the off-line CA not only improves the trustworthiness among the cluster heads, but also helps to overcome the limitation of the certificate chain.

The scheme that we presented in this paper only represents our initial effort for the development of a sophisticated key management for mobile Ad Hoc networks. As the future work towards achieving our eventual goal, we plan to develop and improve the key management. In addition, although we have argued qualitatively that our security scheme is more advantageous over most previous schemes in terms of message exchange overhead and performance, we will perform more quantitative analysis through simulation or measurement to further justify our claim.

References

1. Fu, Y., He, J., Li, G.: A Composite Key Management Scheme for Mobile Ad Hoc Networks. In: Proc. OTM Workshops 2006. LNCS, vol. 427, pp. 575–584. Springer, Heidelberg (2006)
2. Zhou, L., Haas, Z.J.: Securing Ad Hoc Network. IEEE Networks 13(6), 24–30 (1999)
3. Yi, S., Kravets, R.: Key Management for Heterogeneous Ad Hoc Wireless Networks. In: Proc. 10th IEEE International Conference on Network Protocols, pp. 202–203 (2002)
4. Yi, S., Kravets, R.: MOCA: Mobile Certificate Authority for Wireless Ad Hoc Network. In: Proc. 2nd Annual PKI Research Workshop (2003)
5. Hubaux, J.P., Buttyan, L., Capkun, S.: Self-Organized Public-Key Management for Mobile Ad Hoc Networks. Transactions on Mobile Computing, 52–64 (2003)
6. Hubaux, J.P., Buttyan, L., Capkun, S.: The Quest for Security in Mobile Ad Hoc Networks. In: CACM, pp. 146–155 (2001)
7. Khalili, A., Katz, J., Arbaugh, W.A.: Toward Secure Key Distribution in Truly Ad-Hoc Networks. In: Proc. Symp. on Applications and the Internet Workshop, pp. 342–346 (2003)
8. Zheng, S.R.: The Technique of Ad Hoc Network. Posts & Telecom Press (2004)
9. Lin, X., Zen, X.: Mobile Ad Hoc Network-Self-Organization Wireless Network Technology. Publishing House of Electronics Industry, Beijing, China (2006)

Agent Based Formation and Enactment of Adaptable Inter-organizational Workflows in Virtual Organizations

Habin Lee[1], Hyung Jun Ahn[2], Hanseup Kim[3], John Shepherdson[4], and Sung Joo Park[3]

[1] Brunel Business School, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK
[2] Waikato Management School, Waikato University, Hamilton, New Zealand
[3] Graduate School of Management, KAIST, Seoul, South Korea
[4] Intelligent Systems Research Centre, BT, Ipswich, UK
Habin.Lee@brunel.ac.uk, hjahn@waikato.ac.nz, Hanseup.Kim@samsung.com, john.shepherdson@bt.com, sjpark@business.kaist.ac.kr

Abstract. This paper proposes an agent-based framework for the composition and handling of change requests in dynamic inter-organizational workflows (IOWs) in virtual organization context. The framework adopts the software component technology to enable agents in an IOW to handle ad-hoc change requests in the run-time of the IOW. A change propagation model that guides the coordination process of the agents to handle a change request is proposed and a prototype system of the framework is detailed.

Keywords: Agent, inter-organizational workflows, eMarketplace, change handling.

1 Introduction

An eMarketplace is an Internet-based open place wherein different organizations with heterogeneous information systems may continuously form virtual organizations (VOs) [1] temporarily for business opportunities and dissolve it when they complete their goals.

While the openness of the eMarketplace is a very desirable feature as it can attract more organizations which do not belong to any legacy supplier-buyer networks and can achieve optimal efficiency via complete competition among the organizations, at the same time it poses a huge challenge of implementing flexible information systems to enable such open eMarketplaces. Due to the heterogeneity of the workflow management systems (WfMSs) of the participating organisations, much effort has been expended to enable the integration of disparate WfMSs [2][3][4]. While such research has mainly concerned itself with the transfer of process and application data between heterogeneous WfMSs, the coordination of the process participants to handle change requests in IOWs has not been addressed sufficiently.

The issue of handling changes on customer requirements becomes more complicated due to the nature of a VO. The participating organizations may need to have a series of negotiations to change a partner, change colour of ordered products at the cost of a member organization, and so on. Apart from that, the openness of

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 594–603, 2008.

eMarketplaces makes it difficult to have a shared mechanism to handle change requests due to the heterogeneity of information systems and participating organizations. The problem becomes even more complicated with flexible entrance of new organizations into eMarketplaces.

This paper proposes a framework, called MADIOF to develop a WfMS to automate agile IOWs, where autonomous and intelligent agents play a central role in coordinating IOWs and handling change requests. In MADIOF, each participating organisation is represented by an agent, and the execution of an IOW is coordinated by an agent that routes process and application data from one organization to another, according to a dynamically composed IOW. The handling of change requests is done via dynamically provided coordination components which encapsulate all the details of negotiation protocols and can be plugged into the agents when they join a VO.

This paper is organized as follows. Section 2 reviews related works and section 3 presents the suggested adaptive coordination framework for IOWs. Section 4 explains the execution and change management in dynamically composed IOWs and section 5 explains a prototype system which implements the framework. Finally, section 6 concludes the paper.

2 Literature Review

Agent technology has been used to enable VOs mainly for dynamic formation and operation of a VO [5][6][7][8]. The autonomy and intelligence of software agents are regarded very suitable for representing the individual interests of organizations in an open environment. Their focus, however, has been mainly on the mediation of organizational agents for finding their partners for a business opportunity, and not enough attention has been paid to the adaptive execution of IOWs within a VO context. The study tries to expand the flexibility and adaptability of agents by allowing them to use ad-hoc interaction protocols, while also maintaining the basic strengths of autonomy and intelligence in coordinating VOs.

While there have been a significant number of studies on handling exceptions in workflows ([9][10][11] for instance), very few have challenged handling change requests in IOW executions. For example, Luo et al. [12] used the case-based reasoning technique to reuse exception handling knowledge in IOW context. This study focuses on rather static relationship between organizations where the system can monitor, accumulate, and reuse exception handling knowledge over a long period of time. Some studies used agent technology to enable adaptive supply chain collaboration [13][14]. They proposed conversation models, or interaction protocols, that can be dynamically installed and interpreted by agents, and that can lead to highly flexible collaboration among agents. The models allow new organisations that use ad-hoc conversation policies to participate in existing VOs dynamically by providing the models to the existing participants in the VOs. Although the studies were not developed for change handling of IOWs, the key ideas can be applied to the change handling problem as well in a similar manner.

3 MADIOF: A Multi-agents Based ADaptive IOw Coordination Framework

MADIOF is an eMarketplace framework in which intelligent agents form a VO and execute inter-organizational workflows autonomously with minimum intervention of human users. The overall architecture of MADIOF is shown in Fig. 1. MADIOF consists of several sub components for the support of the VO lifecycle. The major components of MADIOF are VO management service and member organizations. VO management service provides generic functionality for the operation and maintenance of virtual organizations in an eMarketplace such as registry of member organizations (registry service), matchmaking partner organizations (market service), and transaction support (coordination service). A new organization becomes a member organization by registering to the VO management service (VOMS). In MADIOF, the lifecycle of a VO has been slightly modified to be used under eMarketplace context into 5 phases: preparation, candidate pooling, organizing, execution, and dissolution. In preparation phase, a hosting organization prepares a call-for-bid (CFB) to form a VO for a business opportunity. The CFB contains an IOW schema that defines the participant roles and the interfaces for the interaction between the roles as well as the general information with regard to the negotiation process to determine the actual member organizations in charge of the roles in the IOW including the due date/time for bids. The CFB is registered at the market service of the VO management service. The market service may notify the new registry to subscribed organizations based on matchmaking between the registered content and the business interests of the subscribed organizations. The candidate pooling is a phase wherein the hosting organization prepare a list of candidate organizations for each role in the IOW based on the bids made by the participating organizations. In formation phase, the host organization negotiates with the candidate organizations to determine the partner organizations. The negotiation protocol may different for different role and the candidate organizations can contact coordination service of the VOMS to find the coordination protocol service which was registered by the host organization. This allows the participating organizations do not have to know all the coordination protocols before they join the VO. The operation phase enacts the IOW formed at the formation phase. Actual data and control is routed based on the IOW scheme of the VO and coordination protocols are used to handle any exceptions or change requests if necessary. Finally, the host organization declares the dissolution of the VO when the mission of the VO has been achieved or turned out not achievable. All participating organizations performs their dissolution processes to finalize any output from the IOW.

In MADIOF, autonomous and intelligent agents automate most of the processes during the VO lifecycle. The agents in MADIOF are largely classified into two categories: administrational agents (AAs) and organizational agents (OAs). Administrational agents are the major players providing VO management services. An OA performs all the activities on behalf of its representing organization in an eMarketplace. There are four different AAs in MADIOF: registry agent, mediation agent, audit agent, library agent, and coordination agent. A registry agent maintains a registry that keeps all the information of the organizations participating in the eMarketplace including organization information and the identifiers of initiator and/or respondent agents representing the organization. A mediation agent is responsible for the mediation service in which an OA registers a call for bid (CFB) to form a VO. The mediation agent notifies a new registration of a CFB to OAs that subscribed their interests on the mediation agent. The mediation agent maintains a repository to maintain the details of the CFBs. Finally, a coordination agent is a special agent that coordinates organizational agents in a VO in case a change on the IOW instance has been requested. A coordination agent is a mobile agent that is transferred to the organization which issued a change request and initiates a coordination process to handle the change request based on a change handling rule.

Depending on who initiates a VO creation, an organization agent can be a host or participant in the VO. Firstly, a host agent is assigned to the organization which wants to form a VO for a business opportunity. The host agent executes all the activities necessary for the creation of a VO and execution of the IOW required for the business opportunity. The activities include preparing a CFB, registering the CFB at the VOMS via the interaction with registry agent, accepting and evaluating bids from responding organizations, awarding winner organizations, composing an IOW and executing the IOW by collaboration with participating organizational agents. A participating agent is assigned to an organization that wants to join to the VO advertised by a host agent. Its mission is to monitor any business opportunity in the eMarketplace and make a bid for an opportunity to win a contract and execute the IOW given by the host agent. On being notified by the mediation agent in the VOMS, a participating agent prepares a bid based on predefined rules if the CFB satisfies the conditions that are set up by its representing organization.

The communication between organizational agents for the composition and change request handling is performed via C-COM (cooperative component) in MADIOF. A C-COM (Lee et al., 2003) is an executable software module that performs message-based communication between two or more role-components according to a predefined interaction protocol, in order to achieve a coordination goal.

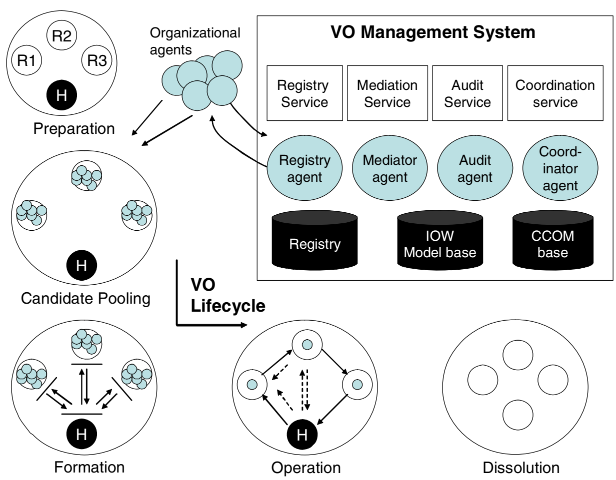


Fig. 1. The services and the VO lifecycle in ADIOF

The main feature of a C-COM is that it is divided into two or more role-components. There are two generic role-components for each C-COM - Initiator and Respondent. In particular, the Initiator component can be plugged dynamically into an agent. From an agent's point of view, the Initiator component is a black box, which hides the details of the interaction process (with its designated a Respondent component) from its master agent, only exposing an interface which specifies the input and output data needed to execute the role component. A C-COM enables a group of agents to communicate (via an interaction protocol previously unknown to them) by dynamically installing one or more role components, provided by the C-COM, which implement the new interaction protocol.

4 IOW Execution and Change Management in MADIOF

The progress of an IOW instance is made along with the distributed execution of the local workflows of the participating organizations in a VO via their organizational agents. As the IOW schema is shared by all the participating agents, each agent interprets the schema and decides next organizational agent to which workflow control is handed over based on the IOW specification. When an OA takes the control of an IOW, the OA interacts with a WfMS via its API to initiate a local workflow instance and receives an execution result from it. The detail of the workflow instance is encapsulated within the local WfMS, and is accessible only to the OA.

The C-COM approach allows a coordinator agent and OAs to communicate with each other using unknown interaction protocols to handle new types of change requests. This functionality is very important where the types of change request in an IOW cannot be determined in advance. If a new type of interaction between a coordinator agent and OAs is required to handle a new change request, this interaction is implemented in a C-COM. The Initiator component of the C-COM is installed into a coordinator agent and the Respondent role-component is installed in one or more OAs.

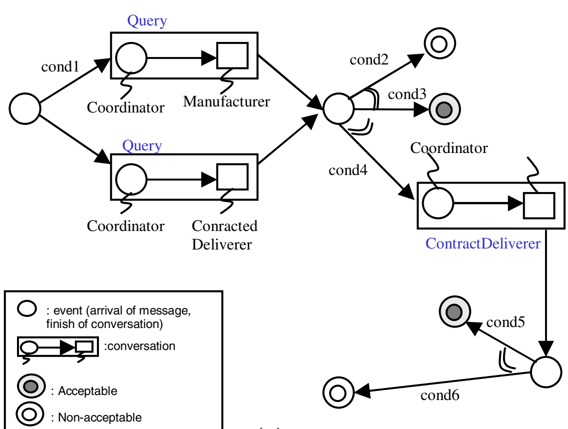
The process of handling a change request is controlled by a coordination agent that is instantiated at the VOMS when an organization issues a change request and migrates to the organization. The reason the coordination agent migrates to the organization is to minimize the computational load of VOMS which easily can become the bottleneck of the whole system if it processes all the change requests from participating organizations. Furthermore, the coordination for handling change requests is separated from the OA to separate the focus of concerns between normal IOW execution and exception handling.

The change request handling process is performed through two phases: acceptance checking and commitment. In the acceptance checking phase, a coordinator agent identifies the change propagation scope which is a set of OAs affected by a change request and communicates with the identified OA(s) to evaluate the acceptability of the change request. The main tool used to accomplish this is the change propagation model (CPM). A CPM is used to specify the conversation policy for coordination among a coordinator agent and OAs to determine the acceptability of a change request based on the current state of the target IOW instance. The provision of an interpretable CPM to a coordinator agent, rather than hard coding the handling process of a change request into the coordinator agent, increases the flexibility of MADIOF to adapt to new kinds of change requests dynamically, as the latter may require novel patterns of communication amongst participating agents.

In MADIOF, a CPM is represented as an XML file to allow easy interpretation of a CPM by coordinator agents and OAs. Fig. 2 (b) shows the part of the DTD of a CPM.

The main modelling constructs of a CPM are events, C-COMs, and links between an event and a C-COM. An event represents an arrival of a change request (a trigger event), a final state of a conversation (a mediator event), or a final state of an entire conversation (a terminator event). A CPM starts with a trigger event. One or more links connect the initiator event to one or more C-COMs. The links may be attached with one or more conditions that constrain the transfer of control from the initiator event to the C-COM. A condition is represented as a desirable state of the ontology used in an IOW, IOW instance, participant attributes, change types, or a result of any previous conversations. A transition from an event to a C-COM triggers execution of a conversation with one or more OAs via a C-COM. Execution of a C-COM may lead to a mediator event or a terminator event in a CPM. If it is a mediator event, then a coordinator agent follows the links from the event until it is confronted with a terminator event. A terminator event ends the change request handling process by marking the acceptability of a change request as either 'Accepted' or 'Rejected'.

Once an acceptance-checking phase is completed, a coordinator agent initiates a change commitment phase. In this phase, the IOW which was halted (pending the outcome of a change request) is restarted, aborted, or re-configured according to the result of the acceptance checking phase. If the coordinator agent evaluates the change request as unacceptable (for instance, the manufacturer in the example IOW cannot meet the new delivery date), the coordinator agent sends an abort message to each OA participating in the IOW, which enables them to restart their halted process instances. On the other hand, if the change request is found to be acceptable (for instance, both the manufacturer and delivery company can meet the new delivery date), the coordinator agent sends a commit message to each OA.



(a)

<!ELEMENT cpm (event+, ccom+, link +)>
<!ELEMENT event (id, name, event_ type, pre_cond)>
<!ELEMENT event_type(trigger | mediator | terminator)>
<!ELEMENT ccom (id, class_ locator, input_ontology+, output_ ontology+)>
<!ELEMENT link (pre_node, post_node, cond)>*
<!ELEMENT cond (cond_type, attr_type, attr_name, value)>
<!ELEMENT cond_type(iow_status | change_type | order_value | participant_attribute | ...)>

(b)

Fig. 2. (a) A CPM Diagram (b) The DTD of the CPM

The commit or abort message is processed by OAs using a dynamic software component that is provided along with a CPM.

5 Illustrative Example: Furniture on Demand

Suppose there is an electronic shopping mall that provides a ‘manufacturing on demand’ service to its online customers. A customer can order items of furniture according to his/her preference such as the colour of a specific item, the number of drawers in a wardrobe, the number of chairs in a dining table set, required delivery date, and so on. Once the order is issued, the shopping mall contacts manufacturing and delivery companies to find those that can manufacture and deliver the ordered furniture meeting the customer specified date. Once the order details have been agreed between a customer and the shopping mall, order forms are created and sent to the corresponding manufacturers and a delivery company. The manufacturers start their processes and notify the delivery company when the items are ready. The delivery company collects the ordered components from the various manufacturers and delivers them to the customer on the specified date.

The CPD in Fig. 2 (a) shows the process that handles a request for changing the delivery date regarding a customer order. Once the request arrives from an OA that is responsible for a shopping mall, the coordinator agent is supposed to initiate one or two conversations based on the current state of the IOW instance that is processing

```
<cpm>
<ontology>manufacturing.furniture
</ontology>
<event>
<id>e001</id>
<name>change_requested</name>
<type>trigger</type>
</event>
<ccom>
<id>query_manufacturer</id>
<class_locator>madiof.app.fod.ccom.
changes.date.initiator.
QueryManucaturer_initiator
</class_locator>
<input_ontology>NewDeliveryDate
</input_ontology>
</ccom>
<link>
<id>lnk001</id>
<pre_node>e001<pre_node>
<post_node>query_manufacturer
</post_node>
<cond><cond_type>iow_status
</cond_type>
<attr_type>task</attr_type>
<attr_name>manufacture</attr_name>
</attribute><value>started</value>
</cond>
</link> … </cpm>
```

(a)

```
(defrule load-a-cpm
  ?f1 <- (event (id ?t-evt)(type trigger)
               (status activated))
  => (retract ?f1) (load-cpm ?t-evt))
(defrule event-activated
  (event (id ?evt))
  (link
(pre_node ?evt)(post_node ?successor)
  (cond ?cond))
  ?f1<-(condition (id ?cond))
  => (evaluate-cond ?f1 ?successor))
(defrule execute-ccoms
  (ccomqueue)
   =>
  (execute-ccom ?ccom ?fileloc ?context))
(defrule increase-token
  ?f1<-(ccom (id ?cid)(status responded))
  (link (pre_node ?cid)(post_node ?eid))
  ?f2<-(event (id ?eid)(cur_token ?ct))
  =>
  (retract ?f1)  (retract ?f2)
  (bind ?nt (+ ?nt 1))
  (assert (?f2 (cur_token ?ct))))
(defrule activate-m-event
  ?f1<-(event (id ?eid)(no_token ?nt)
                     (cur_token ?ct))
  (test (eq ?nt ?ct))
  => (retract ?f1)
  (assert (?f1 (status activated))))
```

(b)

Fig. 3. (a) A CPM specification example. (b) A part of the rules for the evaluation of a CPM.

the original order. If the local workflow 'Manufacture' is still in progress, the coordinator agent initiates a conversation with the manufacturer (via the manufacturer OA) that is executing the local workflow to check the acceptability of the change request via the Query C-COM. At the same time, the coordinator agent executes a C-COM to communicate with the contract delivery company (via the delivery OA) to check whether the new delivery due can be met. The execution result of the two C-COMs leads to a mediator event, which conditionally connects to an Accepted event, Rejected event, or another C-COM depending on the responses from the OAs. If the responses from both the manufacturer and delivery OA are positive, the link from the mediator event to the Accepted event is activated and the Acceptability checking phase is finished. If the manufacturer agent responds with a negative message, the control is transferred from the mediator event to the Rejected event by completing the process. Finally, if the manufacturer agent responds with a positive message and the deliverer agent responses with a negative message, the coordinator agent must execute another C-COM to find an alternative delivery company which can meet the new delivery date. The execution result of the C-COM, then, leads to either an accepted or rejected event. A part of the CPD in Fig. 2 (a) can be specified as follows.

6 Implementation

A prototype system has been implemented to show the practicality of MADIOF by integrating a multi-agent platform JADE (Bellifemine et al., 2003) with a JAVA™ based expert system shell JESS (2003).

Fig. 4 shows the internal architecture of the coordinator agent and organizational agent. A change request is processed by the components in Fig. 4 as follows. First, if a coordinator agent receives a message containing a change request, the message is passed to a FactExtractor component which parses the content of the input message and transforms it into JESS facts, which are inserted into the FactBase component. This triggers a system rule that loads a ruleset that handles the change request into the RuleBase component. At the same time, the FactExtractor component gets the state data from the local process instance that forms part of the IOW instance that is the target of the change request. The state data is also transformed into JESS facts and inserted into the FactBase component. This initialisation of the FactBase and RuleBase components triggers the evaluation of the CPM to handle the change request. If it is necessary to execute a C-COM to enable communication with OAs during the evaluation process, a Coordinator component is requested by the Engine component with input data such as the name of C-COM and the input arguments needed to execute it. The Coordinator component now searches a C-COM library to find the named component and, when found, executes it with the given input arguments. The result of the communication with the OAs is passed to the FactExtractor component to be inserted into the FactBase component, which triggers the continuation of the evaluation process of the CPM. A part of the rules used by the coordinator agent is shown in Fig. 3 (b).

Fig. 4 (b) shows the internal architecture of an OA. One of its key elements is the component registry that maintains the list of installed C-COM Respondent role-components (that interact with the Initiator role-components of a coordinator agent) and a set of change handlers. A change handler is a software module that implements interfaces for checking the acceptability of a change request and handling a change commit or abort command from a coordinator agent. A change handler is customised to access a specific legacy WfMS API for each organisation that wants to participate in an IOW. A message that queries the acceptability of a change request from a coordinator agent is handled by one of Respondent role-components. The Respondent role-component then contacts the Component registry to locate a handler for the change request. Once located, the Respondent role-component dynamically instantiates the change handler (a JAVA™ class) and executes a predefined method.

7 Conclusion

This paper has proposed an agent-based approach for handling change requests in IOWs. The handling of change requests in the middle of the execution of an IOW instance is considered difficult because of the inherent uncertainty of possible changes that may occur due to new participants, new business models, or re-configuration of the virtual organization. MADIOF has been proposed as a component based framework for the creation and execution of adaptable IOWs.

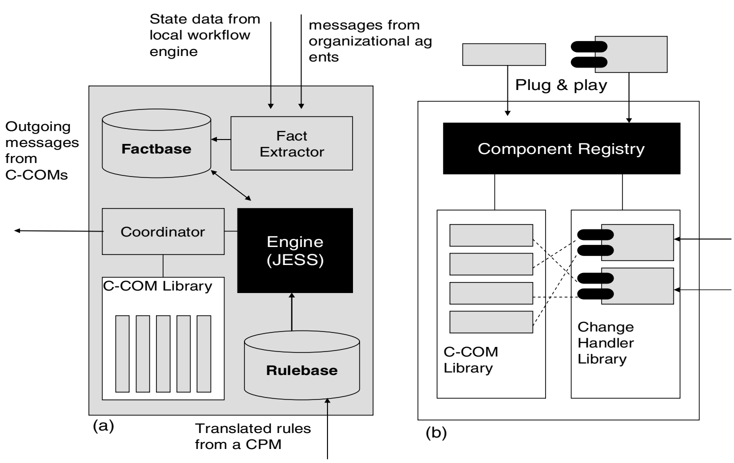


Fig. 4. The internal architecture of a coordinator and organizational agent

In MADIOF, cooperation components that encapsulate the negotiation logics are dynamically plugged into participating agents that use the components in the negotiation processes for composing VOs and handling ad-hoc change requests. CPM has been proposed to allow organizations to specify the negotiation processes to handle ad-hoc change requests.

References

1. Strader, T.J., Lin, F.R., Shaw, M.J.: Information infrastructure for electronic virtual organization management. Decision Support Systems 23, 75–94 (1998)
2. Blake, M.B., Gomaa, H.: Agent-oriented compositional approach to services-based cross-organizational workflow. Decision Support Systems 40, 31–50 (2005)
3. Chebbi, I., Dustdar, S., Tata, S.: The view-based approach to dynamic inter-organizational workflow cooperation. Data & Knowledge Engineering 56, 139–173 (2006)
4. Van der Aalst, W.M.P., Kumar, A.: XML-based schema definition for support of interorganizational workflow. Information Systems Research 14(1), 23–46 (2003)
5. Norman, T.J., Alun, P., Chalmers, S., Nicholas, R.J., Luck, M., Dang, V.D., Nguyen, T.D., Deora, V., Shao, J., Gray, W.A., Fiddian, N.J.: Agent-based formation of virtual organizations. Knowledge Based Systems 17, 103–111 (2004)
6. Rabelo, R.J., Camarinha-Matos, L.M., Vallejos, R.V.: Agent-based brokerage for virtual enterprise creation in the moulds industry. In: E-business and Virtual Enterprises, pp. 281–290. Kluwer Academic Publishers, Boston (2000)
7. Jain, A.K., Aparicio, I.V.M., Singh, M.P.: Agents for process coherente in virtual enterprises. CACM 42(3), 62–69 (1999)
8. O'Leary, D.E., Kuokka, D., Plant, R.: Artificial Intelligence and Virtual Organizations. CACM 40(1), 52–59 (1997)
9. Hwang, S.Y., Tang, J.: Consulting past exceptions to facilitate workflow exception handling. Decision Support Systems 37(1), 49–69 (2004)
10. Casati, F., Ceri, S., Paraboschi, S., Pozzi, S.: Specification and implementation of exceptions in workflow management systems. ACM Transactions on Database Systems 24(3), 405–451 (1999)
11. van der Aalst, W.: Generic workflow models: How to handle dynamic change and capture management information? In: Lenzerini, M., Dayal, U. (eds.) Proceedings of the Fourth IFCIS International Conference on Cooperative Information Systems (CoopIS 1999), pp. 115–126. IEEE Computer Society Press, Los Alamitos (1999)
12. Luo, Z., Sheth, A., Kochut, K., Arpinar, B.: Exception Handling for Conflict Resolution in Cross-organizational workflows. Distributed and Parallel Databases 13, 271–306 (2003)
13. Iwao, T., Wada, Y., Okada, M., Amamiya, M.: A framework for the exchange and installation of protocols in a multi-agent system. In: Klusch, M., Zambonelli, F. (eds.) CIA 2001. LNCS (LNAI), vol. 2182, pp. 211–222. Springer, Heidelberg (2001)
14. Ahn, H., Lee, H., Park, S.: A flexible agent system for change adaptation in supply chains. Expert Systems with Applications 25, 603–618 (2003)

The Impact of Information Sharing Mechanism to Geographic Market Formation

Xingang Xia, Hua Sun, Guo Liu, and Zhangang Han*

Department of Systems Science, School of Management, Beijing Normal University, Beijing, 100875, People's Republic of China
zhan@bnu.edu.cn

Abstract. The importance of information to any kinds of social behaviors is needless to waste any words. Different information sharing mechanisms and their functionalities to social systems are not well studied. In this paper, with agent-based model, we investigated information sharing mechanisms that leads to different patterns in the macro-level spatial market formation. We proposed three mechanisms. The first is a localized global information sharing mechanism where pheromone is used. The second is one that agents use only information out of their own memory. The third mechanism is one that agents can share information within a group of friends. Simulation results show that centralized markets are formed while agents use localized global information. The distribution of market size in this case follows a power-law distribution. When agents use individual information, no big markets can be formed. In this case, people cannot easily make a successful trade. When agents communicate, they get more information from others but increase uncertainty when communication is not sufficient.

Keywords: Market formation, information sharing mechanisms, spatial pattern.

1 Introduction

Information utilization has a significant impact on people's behavior in economic or social systems. Some research efforts have been devoted to address the importance of information in a system. George Akerlof firstly introduced the definition of asymmetric information in a second-hand car market [1]. Y. C. Zhang [2] found out that with the increase number of participants in a matching game, and improvement of their discernment, people could obtain more information from environment so as to improve the matching quality. Information can influence people's expectations and market pricing. M. Marsili [3] derived an approximation for the stochastic dynamics of a simple model that emulates phenomenon of information aggregation and people herd.

In a spatial system, information sharing is usually modeled via pheromone, which has strong ecological evidence. In an ant society, pheromone trails laid by

* Correspondence Author: Tel. 86.10.58807876. This research is supported by National Science Foundation of China under grant No. 60774085.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 604–613, 2008.

foraging ants serve as a positive feedback mechanism for the sharing of information about food sources [4,5,6]. David J. T. Sumpter [4] investigated how colonies of ant distribute their workforce when offered a choice of two food sources of differing energetic value. He found out that the ability of ants to choose reliably the better of two food sources does not require individual ants to compare food sources, but is simply the result of a slower build-up of the pheromone to the less profitable food source. This work illuminates how pheromone provide a mechanism whereby ants can efficiently allocate their workforce among the available food sources without centralized control. Communicating through pheromone is localized global information sharing (will be addressed later in this paper). This communication-via-pheromone feature finds good application in optimization problems like traveling sales man's problem [7].

Spatial structure can emerge even without information sharing between agents. Spatial integration and segregation are of interest to researchers due to T.C. Schelling's [8] analysis on the segregation in social environments. He defined a model in which agents divided in two species move on a checkerboard according to a given utility function. He showed that even if people only have a very mild preference for living with neighbors of their own type, as they move to satisfy these preferences, complete segregation would occur. This result is considered surprising and has generated a large literature. The structure of the segregated areas is known to be related to the free space available and the exact specification of the rules that govern how individuals move. Interest for Schelling's results has grown among mathematicians [9] and statistical physicists [10]. Michael Batty [11] focuses on advanced spatial analysis of city system. In order to find solution of packing many people into small spaces and not letting them crush each other to death, and developing a quality environment that minimizes crime, he used agent-based modeling. An agent-based modeling in global pattern formation and ethnic/cultural violence published in Science magazine [12] showed that characteristic size of spatial clustering of different ethnic groups is the underlying cause for ethic violence. Violence arises at boundaries between regions that are not sufficiently well defined. The authors modeled cultural differentiation as a separation of groups whose members prefer similar neighbors, with a characteristic group size at which violence occurs. This study points to imposed mixing or boundary clarification as mechanisms for promoting peace.

All these studies are interested in spatial social behaviors and/or the influence of information in a complex system. Out of theses studies, an important issue, information sharing mechanisms, needs to be thoroughly studied. Does it make much difference for an agent if it shares or does not share information with other agents? Is there a middle stage between a pheromone-like information sharing mechanism and a state with no information sharing at all? What is the collective behavior difference in the macro-level if the micro-level information sharing mechanism is not the same?

We study the impact of information sharing in the scope of market formation. Market is a very important mechanism in social life [13,14,15,16]. A market place is where people gather and trade. Many researches have been done in the fields of market formation and people gathering in economic and social systems. Michael W. Macy and Yoshimichi Sato [17] found that social mobility may be the explanation for the difference between Japan and U.S. in trust, cooperation and market formation. With moderate mobility, people can take advantage of better opportunities outside the neighborhood and navigate through a global market. Individuals with similar ideas, habits or preferences tend to create cliques and cluster in communities, which results in segregation at the scale of the society. But no research has been done to address the geographic formation process.

In this paper, we focus on what is the market pattern formed due to the underlying information sharing mechanism. People in a real economy go to a market and trade. This successful or unsuccessful trade experience is memorized by each trader and may influence the decision of the trader in the future to go to this place again or not. People may also decide to go to a market due to the buildings left there. They may select a market with big and good-looking buildings that may hint a potential successful trading site. People may also ask their friends for suggestions of good trading sites. This corresponds to three mechanisms we propose in this paper. The first is a localized global information sharing mechanism, the *pheromone model*, where trading history is marked as pheromone left on the trading site. The second is *individual-memory model*, where agents use only their own memory. The third is the *friend information sharing mechanism*, where agents use memorized information both of their friends and their own.

The contribution of this research may lie in the fact that this is the first research attempt to distinguish information sharing mechanisms and use it to examine the impact of information sharing to spatial integration/segregation. The model is explained in the next section. Simulation results and conclusions are presented in later sections.

2 The Evolutionary Model

All experiments assume a population of N agents, which are initially distributed at random on a square lattice of $L \times L$ with a periodic boundary. When an agent moves out from one side of the space, it enters from the other side. Due to the periodic boundary, there are 9 distances between two agents. We calculate and select the shortest one. The total set of N agents is categorized into two sub-sets with equal number of agents. Half of these agents are buyers and half are sellers. Agents in the buyer sub-set can only trade with agents in the seller sub-set and vice versa. Trade in this paper is only a virtual process which marks the information that the two agents traded. Moore neighborhood is employed. Each agent has a sight range r. At each time step, with probability $1 - P_R$, each agent chooses a site within his sight range and moves to it, where P_R is called the random walk probability. Next, the agent randomly selects a trader (buyer finds seller, seller finds buyer) within his sight range, if there is still one not being selected.

Pheromone Model: In this model, when a trade happens, buyer and seller agents leave trading information (pheromone) on the trading location. Agents use this pheromone to find where to go and their partner.

At each time step, agent finds the site where the amount of pheromone is larger than other sites within its sight range and then jumps to that site. If another agent occupies this site, the agent randomly selects a site within its sight range and jumps there. Next, the agent randomly selects a trader in his sight range, if there are still any agents from the other sub-set left. After trading, the buyer agent (or the seller agent) leaves the amount of H_0 pheromone on the site where it trades. Pheromone evaporates. We introduce a recency factor. Pheromone amount H will become $H(1-\alpha)^n$ after n iterations, with α the evaporate rate (the recency factor).

In this case, agents do not communicate with each other and do not use information from their own trading history. They detect only pheromone around them. This is *localized global* information sharing. It is *global* information sharing because all agents are allowed to get access to the information. It is *localized* because only those agents come close to that site can share that information.

Individual-memory Model: In the individual-memory model, instead of using localized global information, agents use their own trading memories. Each agent has a memory of his own trading experience. When an agent trades, it memorizes the location of this trade and adds S_0 on this location in its memory. S is the memory strength. There is also a recency factor so that S becomes $S(1-\beta)^n$ after n iterations, with β is the forget rate (the recency factor). At each time step, within its sight range, an agent finds the site with the largest S value and jumps there. If that site is occupied, it jumps randomly within its sight range. In this case, agents do not share information with any other body.

Friend Information Sharing Model: The models above are either global information sharing or no information sharing. A natural question arises: is there an information sharing mechanism in between? Agents in the real world may not isolate themselves to use only their own knowledge about the world nor they have the ability to travel all around the world to get global information. They usually sharing information within a group of friends, they communicate. Thus, each agent initially chooses k friends of same type randomly. Incorporating the individual memory mechanism just mentioned in the above model, when each time an agent tries to move, it asks his k friends' suggestions. The k friends will suggest largest value positions, within the asking agent's sight range, in their memories. Considering both his own and his friends ideas, agent moves to the best position unless the position is occupied. Agents have their own memory but do not memorize their friends' suggestions.

3 Measure Clusters

In this research, we define a market by agents gathering together geographically. There are lots of well-developed clustering methods. But some of them are quite time consuming. Here we use DBSCAN (density based spatial clustering of applications with noise) which was proposed by Ester et al [18]. The basic idea of the algorithm DBSCAN is that for each point of a cluster, the neighborhood of a given radius ε has to contain at least a minimum number of $MinPts$ points where ε and $MinPts$ are input parameters. If neighborhood of point p contains at least $MinPts$ points, then create a cluster with p is core point. Then, retrieve all points that are density-reachable from the core point and put them into the cluster. The algorithm terminates when no new point will be put into any clusters. The average run time complexity of DBSCAN is $O(n \log(n))$. In our work, we set $\varepsilon = R_c$ and $MinPts = 1$. $MinPts = 1$ means all points in the neighborhood of core point p will be added into this cluster, reiteratively, we add these neighbors' neighbors into this cluster. Detail of the clustering process is shown below:

1. Create a LocationList to store the lattice information.
2. Create a MarketList.
3. If LocationList is Null, terminate the algorithm.
4. Delete the first element in LocationList. If there is agent in this position, put this agent into MarketList, else go to 3.
5. Find neighbors of first unmarked element in MarketList from LocationList. Delete these neighbors from LocationList and put them into MarketList.
6. Mark the first unmarked element in MarketList. If it is not the last element in MarketList, then go to 5.
7. Add a null element at the end of MarketList to show the whole set is complete and mark the list in order to terminate the algorithm. Go to 4.

Furthermore, we define $C(t) = \sum_{i=1}^{N} \frac{N_i}{N}$ as cluster degree to measure clustering. N_i is the number of agents in the given agent i's sight range r. Trading ratio is defined as $D(t) = \frac{N_d}{N/2}$, which indicates the market trading efficiency, where N_d is the amount of trades in time t and $\frac{N}{2}$is the maximum trading time in one iteration.

4 Simulation Results

First, we investigate spatial market formation for the three information sharing mechanisms put forward above with same initial parameter values (See Fig. 1). We find that agents aggregate tightly together and there are several large markets and a few small markets in *pheromone model*. In the *individual-memory model*, there are lots of small markets spreading around the world and large markets cannot be formed. In *friend information sharing model*, several moderate size markets exist, while there are still a lot of small size markets. That is to say, centralized markets are formed while people using localized global information. Large markets cannot be formed if people use their own experiences without communicating each other. Communication enables people cluster tighter than agents use individual information.

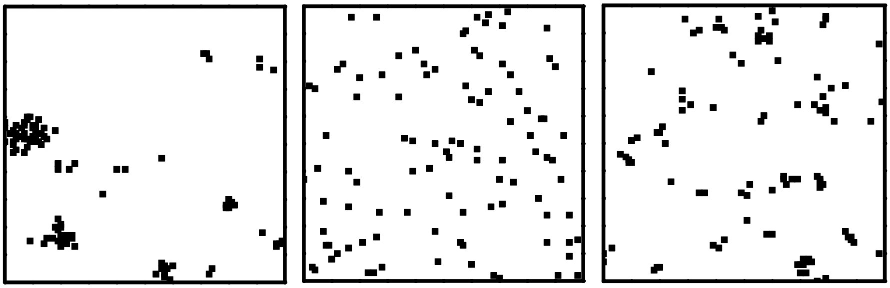


Fig. 1. Spatial market formation with various information sharing mechanisms. From left to right, *pheromone model*, *individual-memory model*, and *friend information sharing model*. The parameter values are given as follows: $L = 100, N = 100, P_R = 0.2, H_0 = S_0 = 100, \alpha = \beta = 0.01, r = 2, R_c = 5, k = 8$.

Convincing results are arrived at while we test the cluster degree. (See Fig. 2). For the *pheromone model*, cluster degree fluctuates from 6 to 16, while the value is about 3 in *individual-memory model*. We also noticed that in *friend information sharing model*, the cluster degree is a little higher than that in the *individual-memory model*.

With the *individual-memory* mechanism, we also test the impact of sight range. Even if there is no limit for agent's sight range, an agent can see all the sites, centralized market cannot be formed (See Fig. 3). Thus, when people use only individual information, it impossible to form a big market.

In order to study market efficiency, we look at the trade ratio. In an efficient market, agents can easily find partners to trade. The average value of trade ratio in *pheromone*, *individual-memory* and *friend information sharing models*

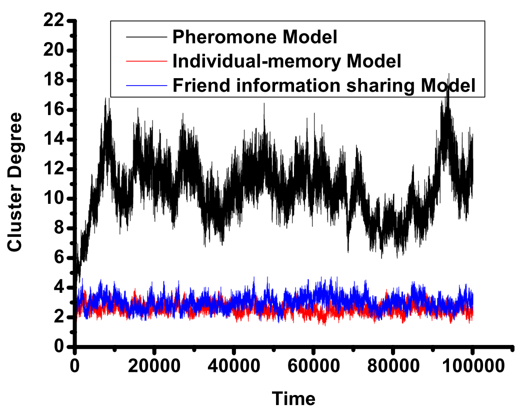


Fig. 2. Cluster degree of three information sharing mechanisms. The average value of cluster degree in *individual-memory model* is 2.62 while it is 3.02 in *friend information sharing model*.

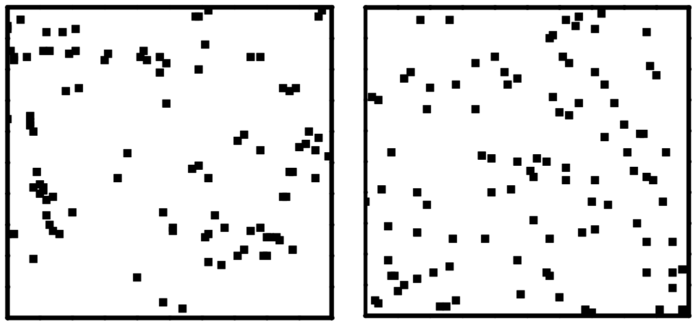


Fig. 3. Test the impact of sight range in the *individual-memory* mechanism. Left is the case that agents are restrained by sight range, right is the case that agents have sight range restrain. In either case, no central market can form.

are 0.57, 0.43 and 0.43 respectively. Compared with individual-memory model, the cluster degree is higher but trade ratio is the same in friends model.

From Fig. 4, we can find out why communication makes trade ratio almost as same as individual-memory at $k = 8$. With the number of friends increase, cluster degree increases almost linearly. Trade ratio, however, drops at first and then increases with number of friends increasing. That is to say, communication enables people to get more information, but increases uncertainty when communication is not sufficient. But with people communicating more, market efficiency grows.

Next, we considered the distribution of market size (See Fig 5). In *pheromone model*, the market size follows power-law distribution. While in *individual-memory model* and in *friend information sharing model*, market size follows e-exponential distribution.

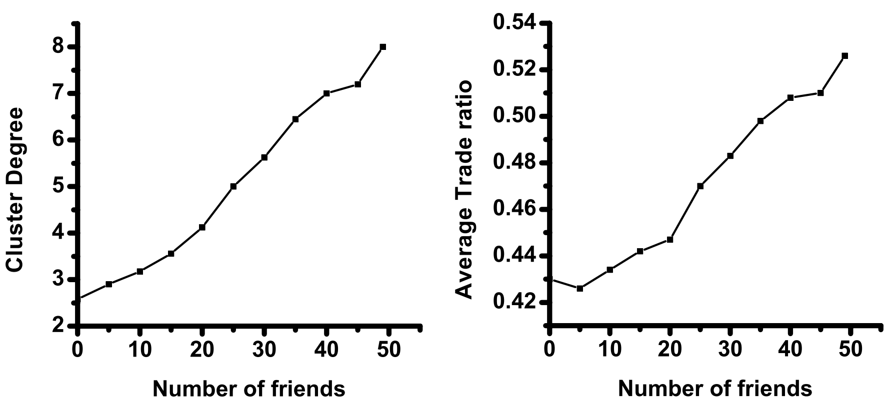


Fig. 4. The change of cluster degree and trade ratio vs number of friends. Points in the curves are the average value of 10 times run.

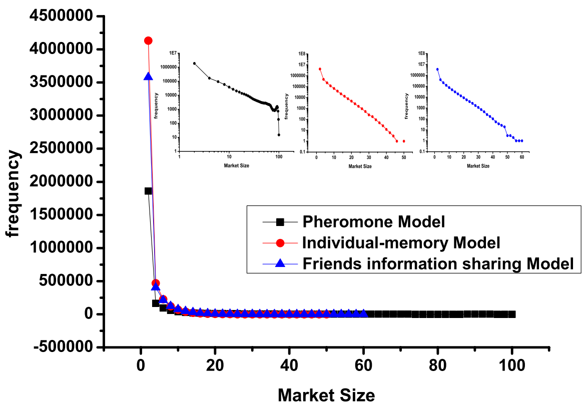


Fig. 5. Market size distributions of these three models. The insets show the distribution of market size on a log-log scale for the *pheromone model* and log-linear scale for the other two. The frequency of market size is calculated 100000 time steps after system goes stable.

Power-law distributions have been observed in a diverse range of research fields. Some well-known examples exhibiting "scaling" behavior are city sizes, word frequencies, sizes of business firms, numbers of links of web pages, interactions of proteins, sizes of earthquake and fire, which cover many research fields such as biology, economics, sociology, engineering, and physics. Some of those mechanisms that exhibit growth and preferential attachment give power-law distribution [19]. In *pheromone model*, agents tend to go to places with the larger values of pheromone. This process is similar to preferential attachment. But, the world size and population in our model are fixed, and leads to the drop of the distribution curve at the right end (inset, left).

5 Conclusion and Discussions

In summary, in this paper, we investigate the impact of information sharing to geographic integration/segregation (market formation). We propose three information sharing mechanisms, the pheromone model, the individual-memory model, and the friend information sharing model, which agents use information in a public or private or communicated way. Centralized markets are formed when agents use localized global information and the distribution of market size in this case follows a power-law distribution. When agents use only individual information, no big size markets can be formed and agents have a low trade ratio, hence the efficiency of the market is low. Communication enables agents to get more information, but increases uncertainty to some extent. When agents have enough friends, communication promotes bigger market to emerge.

This paper tries to study information sharing mechanisms in a comprehensive way. This research work wants to integrate information sharing mechanisms with three gradually changing approaches, from no information sharing to information sharing in groups of agents to localized global information sharing. The results of this paper may shed some light on studying functionality of information instead of counting information in bits and bytes.

The study makes several simplifying assumptions that might usefully be relaxed in follow-up studies. We will try to integrate the three information sharing mechanisms into one and use control parameters to change gradually from one to the other. Agents in this study are heterogeneous in the way that they are buyers or sellers. We will try to introduce mechanisms that allow even the same kind of agents use mixed strategies such as using public information as well as private information. Agents in the next step study may also learn from their trading experiences and make better decisions. We plan to test strategies that agents can add or reduce friends based on friends' suggestions and their own revenues.

References

1. George, A.: The market for Lemons: Quality Uncertainty and Market Mechanism. The Quarterly Journal of Economics 84, 488–500 (1970)
2. Zhang, Y.C.: Happier world with more information. Physica A 299, 104–120 (2001)
3. Veglio, A., Marsili, M.: Stochastic analysis of an agent-based model. arXiv0705.4025v1 (2007)
4. Sumpter, D.J.T., Beekman, M.: From nonlinearity to optimality: pheromone trail foraging by ants. Animal Behaviour 66, 273–280 (2003)
5. Bonabeau, E., Deneubourg, G.: Self-organization in social insects. Trends in ecology and evolution 12, 188–193 (1997)
6. Pratt, S.C., Sumpter, E.B.: Quorum sensing, recruitment, and collective desision-making during colony emigration by the ant Leptothorax albipennis. Behavioral Ecology and Sociobiology 52, 117–127 (2002)
7. Chu, S.-C.: Ant colony system with communication strategies. Information Sciences 167, 63–76 (2004)
8. Schelling, T.C.: Dynamic Models of Segregation. Journal of Mathematical Sociology 1, 143–186 (1971)
9. Pollicott, M., Weiss, H.: The Dynamics of Shelling-Type Segregation Models and a Nonlinear Graph Laplacian Variational Problem. Advances in Applied Mathematics 27, 17–40 (2001)
10. Stauffer, D., Solomon, S.: Ising, Schelling and Self-Organising Segregation. arXiv:physics/0701051 (2007)
11. Batty, M.: Modeling cities from the Bottom-Up using agent-based models
12. Lim, M., Metzler, R., Bar-Yam, Y.: Global Pattern Formation and Ethnic/Cultural Violence. Science 317, 1540–1544 (2007)
13. Tassier, T., Menczer, F.: Emerging Small-World Referral Networks in Evolutionary Labor Markets. IEEE Transactions on Evolutionary Computation 5, 482–492 (2001)
14. Nicolaisen, J., Petrov, V., Tesfatsion, L.: Market Power and Efficiency in a Computational Electricity Market With Discriminatory Double-Auction Pricing. IEEE Transactions on Evolutionary Computation 5, 504–523 (2001)
15. LeBaron, B.: Empirical Regularities From Interacting Long-and Short-Memory Investors in an Agent-Based Stock Market. IEEE Transactions on Evolutionary Computation 5, 442–455 (2001)

16. Izumi, K., Ueda, K.: Phase Transition in a Foreign Exchange Market-Analysis Based on an Artificial Market Approach. IEEE Transactions on Evolutionary Computation 5, 456–470 (2001)
17. Macy, M.W., Sato, Y.: Trust, cooperation, and market formation in the U.S. and Japan. PNAS 99, 7214–7220 (2002)
18. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd international conference on knowledge discovery and data mining, pp. 226–231 (1996)
19. Albert, R., Barabasi, A.-L.: Statistical Mechanics of Complex Networks. Reviews of Modern physics 74, 48–97 (2002)

Agent System for Online Ticket Resale

Jae Hyung Cho[1], Hyun Soo Kim[2], Hyung Rim Choi[2], and Jae Un Jung[2]

[1] School of International Business and Area Studies, Pusan University of Foreign Studies
55-1 Uam-dong, Nam-gu, Busan 608-738, South Korea
chojh@pufs.ac.kr
[2] Division of Management Information Science, Dong-A University
840 Hadan2-dong, Saha-gu , Busan 604-714, South Korea
{hskim, hrchoi}@dau.ac.kr, ace@goodplayer.kr

Abstract. This study has tried to suggest a new model that can effectively redistribute the tickets in the online ticket resale market, while suggesting a new allocation mechanism based on an agent negotiation. To this end, this study has analyzed and simulated the secondary ticket market through System dynamics. As a result of these simulations, it has been proved that the price stability of ticket resale market leads to an increase in revenue. An agent negotiation helps to stabilize the ticket prices that are usually inclined to rise at auction, benefiting all the participants in the negotiations, consequently showing a Pareto solution.

Keywords: **A**gent negotiation, Allocation problem, System dynamics, Ticket resale.

1 Introduction

This study has tried to make a model for ticket redistribution in the secondary market, also suggesting an effective negotiation methodology for it. The tickets for performing arts, sports games, and movies are generally sold online on the Internet, and the market of this online ticketing is rapidly expanding year after year. Because of this, the ticket resale market, i.e. a secondary ticket market, in which the booked tickets bought online on the Internet can be resold, is also expected to steadily expand. In the case of America, the market size of the secondary ticket market is estimated to be about US$25 billion, and now the well-known Internet specialists such as Tickeknow.com and Stubhub.com have appeared[6].

If a ticket holder cannot participate in the event, in general, he will be able to be refunded by the initial seller, or he will try to resell it in the secondary ticket market. But in the former case, he has to pay the charge for refund or cancellation, frequently causing friction. In the latter case, he himself sells his ticket on the Internet auction site. In this case, if he sells it above its face value, he will be able to gain a profit. But if the secondary ticket market lacks transparency and reliability, it can be controlled by the negative forces seeking only a capital gain such as a ticket-scalper. This will have a negative impact on the primary market, eventually causing damage to many related persons[2].

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 614–623, 2008.

On the other hand, if a rigid regulation such as resale prohibition is imposed on the market, it will rather have a negative impact on market, thus shrinking existing markets, damaging the positive function of a market, and consequently causing a loss to the buyers.

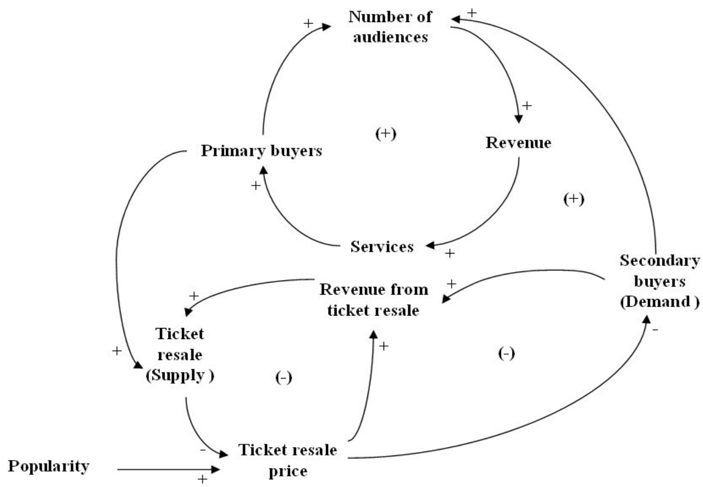


Fig. 1. Ticket resale's impact on the market

Let's take a look at the above Fig. 1 described in the cause and effect diagram of System dynamics. First, the increase of primary buyers means the increasing number of audiences, and this will increase revenue of primary ticket sellers. In this case, better services(including performance, auxiliary facilities and services) will be able to be provided to the audiences. That is, the increase of primary buyers leads to a virtuous cycle. Meanwhile, if the number of ticket resale from among the primary buyers increases, the ticket resale price will generally fall. However, if the popularity and perception of the event is high, and the number of ticket resale is limited, the resale price will be put above its face value. In this case, the revenue from ticket resale will increase, and the supply of ticket resale will also expand.

What matters here is that if the buyers seeking only a capital gain from the ticket resale market increase, and so if the secondary market becomes larger than the primary market, a lot of buyers will suffer losses. Also, the heightened price of ticket resale is to reduce the demand of secondary buyers, consequently having an adverse impact on the primary market. For instance, special stands or ringside tickets can be sold by negative forces that don't see the event. So many people who want to watch the event have to pay more money, or give up. The Fig. 1 shows that the transparency and reliability of the secondary market leads to the activation of its primary market. In particular, how the prices of ticket resale are determined in the secondary market is very important. The reason is that it has a direct influence on the resale demand. In the next chapter, this research will be made on the model that can bring efficient and transparent ticket redistribution in the secondary market, while suggesting an agent-based price determination mechanism.

2 Seller-Based Ticket Resale Model

2.1 Modeling

The ticket resale model suggested in this study is based on the “Customer-driven reservation adjustment (CRA) mechanism” of Kawamura et. al (2004)[4]. In the CRA mechanism, the secondary buyer who wants to buy the ticket of a specific date suggests to the system a ticket price higher than its face value. Then the system sends e-mail to the primary buyers, and selects at random from the primary buyers who respond to this. The selected primary buyer sells to the secondary buyer the ticket. If the primary buyer wants to buy another ticket, the system can provide the ticket of a different date. The ticket price will be at a discount in this case.

This study has added an agent negotiation to the CRA mechanism, thus trying to suggest a more practical model and system. First of all, the key point of this model is that ticket seller company is directly operating the model. That is, the primary buyers don’t sell their tickets in the secondary market, but original ticket seller(company) does the job of ticket resale. For this reason, this model is named a “Seller based Agent Ticket Resale (SATR)” model.

By using this model, the original ticket seller can offer the ticket of a different date to the primary buyer. By using this model, the primary buyers can avoid the risk of being unsold or being sold at a much-discounted price. More importantly, this system can eliminate a counter productive black market, while making it possible to secure in advance the buyers who can buy the ticket of different dates. This will make a booking market more brisk and active, eventually leading to stable revenue increase. Another key point of SATR model is that the seller(primary buyer) and the buyer(secondary buyer) directly determine their ticket price through an agent negotiation and redistribute their tickets. The basic structure of this model is as follow.

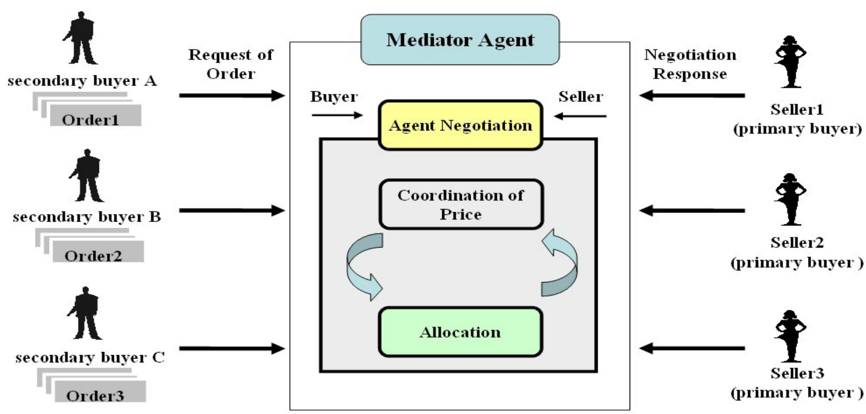


Fig. 2. The process of SATR

Now, let’s compare this SATR model with existing auction models. The differences are shown in the below <table 1>.

Table 1. Comparison between existing auction models and SATR model in the ticket resale market

	Existing auction model	SATR model
Final number of customers (buyers)	1	More than 1 (Advanced ticket sale by suggesting the ticket of different date to the primary buyer who wants to resell his first ticket)
Resale revenue	More than face value	More than face value
Resale beneficiary	Primary buyer(Reseller), Auction company	Primary buyer, Second buyer, Original ticket company
Price determination method	Negotiations between sellers(primary buyers) and buyers(secondary buyers), or at an appointed price	Negotiations between sellers(primary buyers) and buyers(secondary buyers)
Negotiation method (simultaneous negotiation)	One to one, or one to many(buyer)	Many to many

But this model has the following considerations. First, if resale is impossible and the primary buyer decides to cancel, the original ticket seller can suffer the loss of original ticket price. Second, in case that resale revenue is less than an existing auction model, is this model equipped with another proper system to guide the primary buyer?

In order to eliminate these two risk factors, this study has also suggested a negotiation model based on an agent.

2.2 Agent Negotiation

The principle of an agent negotiation hangs on the efficiency of ticket redistribution. To this end, the agent negotiation plays the role of restraining the excessive increase of ticket resale prices. As shown in the Fig. 1 of the cause and effect diagram, if the ticket price increases, the supply of ticket resale will increase, thus increasing the danger of making the secondary market a negative black market. Moreover, if the ticket resale price rises sharply, the demand of secondary buyers will diminish, consequently shrinking the whole ticket market. Therefore, this model has been designed in the way that the prices in the ticket resale can be determined by the negotiations between buyers and sellers, thus restraining the prices from rising sharply.

In general, the sharp increase of prices in the auction market comes from the following reasons: the first is the scarcity of auction items, and the second is the competition between buyers. In the case of a ticket, the scarcity of it in the auction market means its popularity or amusement. In this case, it is to be determined in advance before negotiations, and so it cannot be determined by a negotiation. The key function of this agent negotiation is to cool down overheated competition coming from multiple buyers. To this end, this negotiation is based on the premise that there are many buyers and many sellers.

In order to explain the mechanism of this agent negotiation, the following IP(Integer Programming) model is suggested:

$$Min \sum_{p \in M} \sum_{j \in N} c_p^j x_p^j + \sum_{j \in N} \sum_{e \in E} o_e^j z_e^j$$

$$\sum_{j \in N} x_p^j = 1 \qquad for\ p \in N \qquad \text{(C1)}$$

$$z_e^j \le x_p^j,\ z_e^j \le x_q^j \qquad for\quad e=(p,q) \in E \qquad \text{(C2)}$$

$$z_e^j \ge x_p^j + x_q^j - 1 \qquad for\quad e=(p,q) \in E \qquad \text{(C3)}$$

$$x_p^j \in \{0,1\} \qquad for\quad p \in M,\quad j \in N \qquad \text{(C4)}$$

$$z_e^j \in \{0,1\} \qquad for\quad j \in N,\ \ e=(p,q) \in E \qquad \text{(C5)}$$

- Set of Buyers: M = {1,2, …., m}
- Set of Sellers = {1,2, …., n}
- c_p^j : The lowest ticket price in which Seller j∈N can carry out Buyer p∈M
- E: Set of Buyer p∈M and q∈M(p≠q) which have the same seller and an overlapping process period
- o_e^j : The additional ticket price when Seller j∈N negotiates with Buyers (p,q)∈E simultaneously
- x_p^j : if a Seller j∈N carries out a Buyer p∈M , it becomes 1, otherwise 0
- z_e^j : if a Seller j∈N carries out a Buyer p∈M and q∈M(p≠q) simultaneously, it becomes 1, otherwise 0

In the above model, o_e^j is the additional price increase by the seller(j), which has been caused by overheated competition between the buyers(p, q). In order to restrain the price increase, if the buyer considers the seller's price to be higher than expected, he will move to another seller. In this case, the problems of which buyers move to which sellers and of how to compensate the profits and losses coming from the price changes are to be settled by renegotiations among buyers[5]. At this point in time, the renegotiations among the buyers are based on the following premises:

First, there can be another ticket cheaper than o_e^j. This means that the prices of ticket resale can vary according to the seat class, performance round, and date. It also means that buyers can trade with other sellers anytime. Secondly, because of specific buyers' movement, the competition can be reduced, and as a result, the prices(o_e^j) of ticket resale will be lowered. It is so because the ticket price is very sensitive to the performance time, and also it is determined by supply and demand.

Based on these premises, the core logic of renegotiations among buyers is as follows:

- Step 0: More than two buyers choose the best(cheapest) seller.
- Step 1: One buyer(negotiation leader) selects the best seller, and the other buyers(negotiation partners) move to the second best seller.

- Step 2: If a loss takes place owing to the movement to the second best, the loss will be compensated by the negotiation leader.
- Step 3: If the compensation makes the negotiation leader's gain negative, it will move to step 4, but if not, it will move to step 8.
- Step 4: The negotiation leader moves to the second best, and the other buyers(negotiation partners) are left to take the best seller.
- Step 5: If losses take place due to the movement of negotiation leader to the second best, the leader will be compensated.
- Step 6: If the negotiation leader's gain negative after the compensation, it will move to step 7, but if not, it will move to step 8.
- Step 7: Comparing with the compensated result of Step 3, higher gain will be selected.
- Step 8: In this situation, if there is no the second best to move, or each buyer selects the ticket of each seller, the negotiation will be terminated, but if not, it will move to step 1.

[The price change in the ticket resale according to the negotiation situation]

① The ticket price when more than two buyers select the best seller: $(c_p^j + o_e^j)$

② The ticket price when one buyer selects the best seller: c_p^j (j=1)

③ The ticket price when one buyer selects the second best seller: c_p^{j+1}, if $(c_p^j + o_e^j - c_p^{j+1}) < 0$, a loss takes place.

④ The ticket price when more than two buyers select the second best seller: $(c_p^{j+1} + o_e^{j+1})$, if $(c_p^j + o_e^j - c_p^{j+1} - o_e^{j+1}) < 0$, a loss takes place.

3 Experiment

In this chapter tests will be made to check the validity of an agent negotiation for the secondary ticket market. First, in order to simulate the cause and effect diagram of the ticket resale market shown in the Fig. 1, remodeling has been done as in Fig. 3.

As a result of simulation, in the Fig.4, the balance between cumulative buyers(demand) and cumulative sellers(supply) in the resale market has been maintained until the thirty performance(the number of event). But after the sixth performance, the gap between demand and supply has been widened.

As illustrated in the Fig. 5, the gap between demand and supply has a positive influence on the negotiation price. Therefore, the larger the gap, the sharper the fluctuation of negotiation price(resale price) becomes. That is to say, if demand is larger than supply, then the negotiation price(resale price) will rise, but if supply is larger than demand, then the resale price will fall. Therefore, the fluctuation of negotiation prices hangs on the gap between supply and demand.

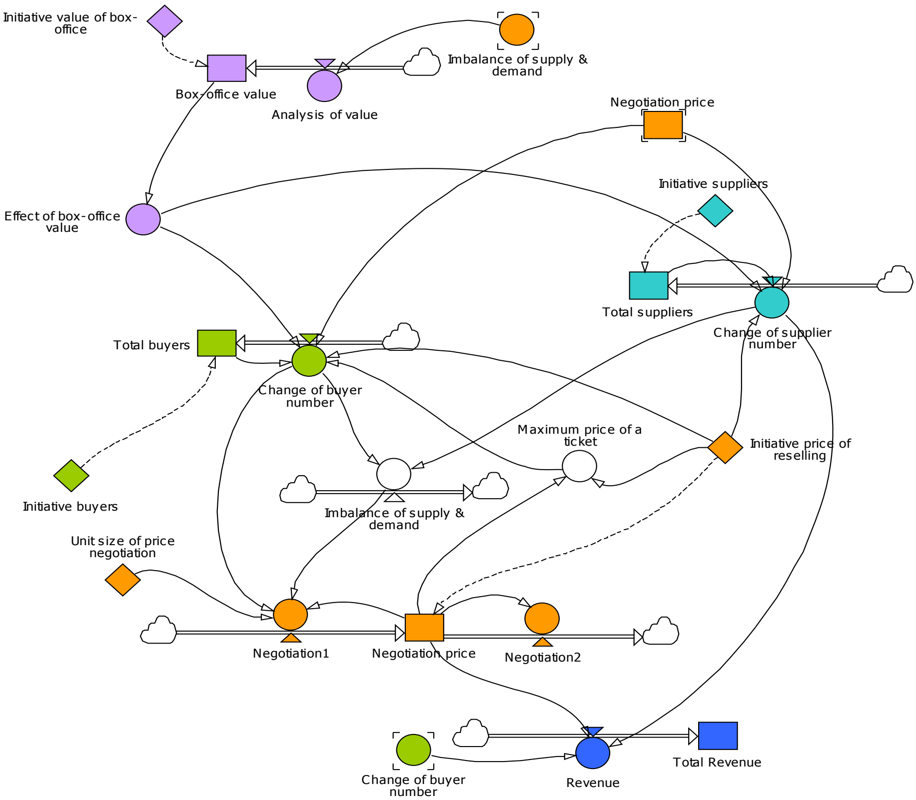


Fig. 3. Simulation model of a ticket resale market

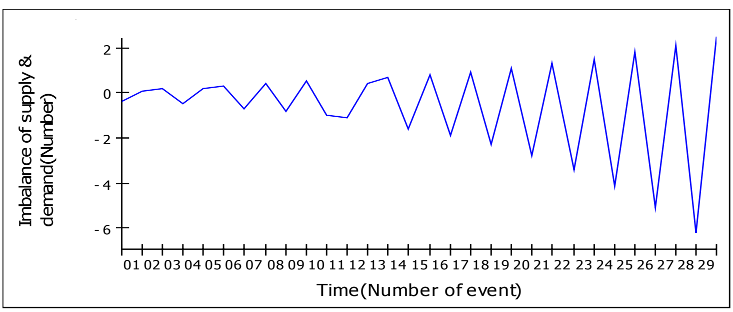


Fig. 4. The change of imbalance of demand and supply in the ticket resale market

As shown in Fig. 6, the negotiation prices for the fifth, sixth and seventh performance in the Fig. 5 are raised, and accordingly, the sale revenue in the secondary market has been slightly increased. But in case that there is little gap between demand and supply, that is to say, in case of a balance between demand and supply, the sale revenue is generally higher than an imbalanced case.

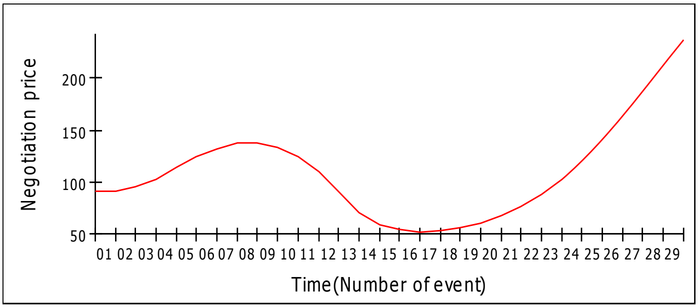


Fig. 5. The pattern of negotiation prices change

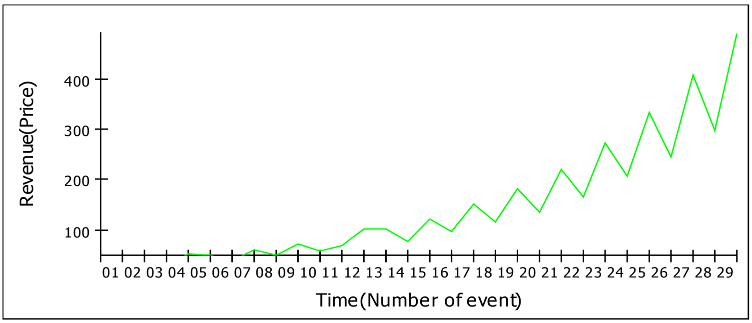


Fig. 6. The change of sale revenue in the ticket resale market

As a result of these simulations, this study has found out that the stability in price determination leads to an increase of sale revenue. In order to stabilize price determination, now let's apply an agent negotiation to the more concrete but simple scenario. As illustrated in the Table 2, three sellers are selling their tickets(class A and class R) to the multiple buyers. Naturally, many buyers will flock to the cheapest seller A(best measure). In this case, heated competition among multiple buyers will raise the price of $100. However, if the negotiation price rises above $105, the buyers will move to the seller B(second best). But if multiple buyers simultaneously flock to the seller B, it will again raise the ticket price. Because of overheated competition and ensuing price surge, it could be better if buyers go to the seller C or choose a different class (class R). In the real market seeking the maximization of one's profit, however, this kind of forecasting is very difficult. Likewise, it is also not easy to an agent that is fully reflecting the real world [7]. After all, multiple buyers' fierce competition makes it difficult to forecast the price movement in the secondary market.

Table 2. Scenarios: The initial ticket price from each seller

Seller		Seller A	Seller B	Seller C
Ticket Price	class A	$100	$105	$107
	class R	$110	$112	$115

When we apply this scenario to agent negotiation, its result is as follows. It shows how the sellers have been selected and how the prices of ticket resale have been determined, along with the allocated number of buyers according to the increasing number of buyers. In particular, in case that the tickets of two different classes are being sold at the same time, the number of buyers of each ticket class will be determined by agent negotiations. Therefore, an agent negotiation helps maintain the balance between buyers and sellers in the market, thus determining the proper price of ticket resale.

Table 3. The result of ticket allocation through agent negotiation in the resale market

Scenario (1)	Seller / Ticket Price (class A)		Seller A ($100)	Seller B ($105)	Seller C ($107)
	Number of Buyers	4	*($105, 2)	($105, 1)	($107, 1)
		10	($129, 4)	($132, 3)	($133, 3)
		25	($161, 9)	($164, 8)	($168, 8)
Scenario (2)	Seller / Ticket Price (class A, class R)		Seller A ($100, $110)	Seller B ($105, $112)	Seller C ($107, $115)
	Number of Buyers	4	class A:($100, 1) Class R:($110, 1)	class A:($105, 1)	class A:($107, 1)
		10	class A:($113, 3) Class R:($110, 1)	class A:($111, 2) class R:($112, 1)	class A:($114, 2) class R:($115, 1)
		25	class A:($145, 6) Class R:($153, 4)	class A:($151, 5) class R:($154, 3)	class A:($155, 5) class R:($148, 2)

* (Final sale price of tickets, final number of buyers)

The problem of ticket allocation through agent negotiation is NP-hard problem in that the number of sellers are "m", the number of orders are "n", and the number of total combinations are $(m)^n$[1, 3, 8]. The results of agent negotiations cannot ensure an optimal solution for ticket resale market, but it brings a Pareto solution in that each and every participant doesn't suffer a loss[5, 7]. Accordingly, if this agent negotiation is added to SATR model in the ticket resale market, its usage will be doubled.

4 Conclusion

The SATR model and agent negotiation suggested in this study contribute to enhancing the transparency and reliability of a ticket resale market. Although it is unable to maximize one's profits, it can get rid of the possibility of causing a loss, consequently improving a social value and common profit. SATR model and agent negotiation complement each other. The success of SATR model necessitates the ticket reallocation mechanism by means of an agent negotiation, and at the same time an agent negotiation is based on the political establishment of infrastructure through SATR.

From now on, further studies will be made on the diverse price determination mechanism that takes place at the auction, so that it can make this study more practical. Also, by using the real data, the results of agent negotiations will be simulated.

Acknowledgments. "This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)" (The Regional Research Universities Program/Research Center for Logistics Information Technology).

References

1. Bertsekas, P.D.: The Auction Algorithm: A Distributed Relaxation Method for the Assignment Problem. Annals of Operations Research 14, 105–123 (1988)
2. David, V.H.: e-Commerce economics. Thomson Learning (2002)
3. Fischer, K., Jorg, P., Pischel, M.: Cooperative Transportation Scheduling: an Application Domain for DAI. Journal of Applied Artificial Intelligence, Special Issue on Intelligent Agents 10(1), 1–31 (1996)
4. Kawamura, H., Yamamoto, M., Ohuchi, A.: Multiagent Analysis of Customer-Driven Reservation Adjustment System. In: APIEMS Conference, 26.4.1–26.4.14 (2004)
5. Kim, H.S., Cho, J.H., Choi, H.R., Hong, S.G., Kang, M.H.: Optimal Supply Chain Formation Using Agent Negotiation in a SET Model-Based Make-To-Order. In: Harper, R., Rauterberg, M., Combetto, M. (eds.) ICEC 2006. LNCS, vol. 4161, pp. 579–583. Springer, Heidelberg (2006)
6. Korea Institute of Science and Technology Information: Briefing on global developments (2007)
7. Michael, P.W., Willian, E.W.: Auction Protocols for Decentralized Scheduling. Games and Economics Behavior 35, 271–303 (2001)
8. Russ, C., Vierke, G.: The Matrix Auction: A Mechanism for the Market-Based Coordination of Enterprise Networks. German Research Center for Artificial Intelligence, Research Report RR-99-04 (1999)

Application of Ubiquitous Computing Technology to the Steel-Plate Piling Process of Ship Construction*

Hoon Oh and Jeong Seok Heo†

Dept. of Computer Engineering and Information Technology, University of Ulsan, Korea
{hoonoh, heojs}@ulsan.ac.kr

Abstract. A gigantic ship is constructed by assembling various types of ship blocks, each of which is made by the cut and paste of the steel-plates. The steel-plate piling process as the very initial stage of ship construction sorts and manages the steel-plates according to the ship blocks that the steel-plates are used to make. This process poses some problems such as process delay due to piling errors, safety vulnerability due to the handling of extra heavy-weight objects, and the uncertainty of work plan caused by lack of information in the pile spaces. We constructed a steel-plate piling process system by employing the ubiquitous computing technology to resolve such problems. The system was experimented on a work simulator that can simulate the steel-plate piling process. Workers can receive an appropriate or intelligent service through the context information of the smart work space that is managed in real time.

Keywords: Smart space, context, ubiquitous, manufacturing, ship construction.

1 Introduction

The interest in applications of the ubiquitous computing (hereafter referred to as UbiCom) technology to various industrial fields including manufacturing plants and construction facilities has been increasing. One application is the bridge monitoring system that monitors the safety status of a bridge and gives a timely warning against lapses in regular maintenance [1]. We apply the ubiquitous computing technology to the steel-plate piling process, the initial stage of ship construction that has been known as a bottleneck degrading ship construction productivity severely.

The UbiCom environment advocated by Mark Weiser [2] was defined as "the computing and information environment where human beings receive an intelligent service according to the current situation, and work efficiency and knowledge sharing are improved by realizing the integration of computers, objects, human beings, and information through the computing elements pervaded into the physical objects spread in our daily life" [3]. The UbiCom environment includes a parsing agent that interprets the meaning of environmental data, a context-aware agent that produces a

---

* This research was supported by the Ministry of Information and Communication, Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment)(IITA-2007_C1090-0701-0039)).

† Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 624–633, 2008.

context, the collection of information resulted from the analysis and synthesis of the data obtained from environment, and a service providing agent that provides an intelligent service by exploiting the contexts [4]. Although there have been researches related to the UbiCom technology, MS EasyLiving [5, 6], iRoom [7], AwareHome [8], HP CoolTown [9], and MIT Oxygen [10] were constructed to conduct fundamental or experimental studies about the UbiCom enabling technologies; there is virtually no case the UbiCom technology was applied to the manufacturing fields.

This paper describes the Steel-plate Piling Process System (U-SPPS) that was built for the intelligent guidance of the steel-plate piling process. The ship is constructed by going through the eight stages: (1) unloading steel-plates from the cargo ship, (2) moving the steel-plates into the temporary pile (shortly *t-pile*) spaces, (3) moving the steel-plates from the t-pile spaces and sorting them into the pile spaces, (4) moving the sorted steel-plates to the pre-processing stage, (5) pre-processing the steel-plates, (6) cutting the steel-plates and making ship blocks, (7) moving the blocks to the docks, and (8) constructing a ship by putting the blocks together. Fig. 1 shows a partial view of the bay consisting of several *t-pile* spaces and several tens of *pile* spaces. A forklift moves the steel-plates unloaded from the cargo ship to the t-pile spaces in the front of the bay. The steel-plates are sorted into the pile spaces such that those piled in one pile space are used to make the same type of blocks. The sorted steel-plates are moved on request, one by one, onto a conveyor belt to be delivered to the pre-processing stage. The steel-plate piling process lies across the third and fourth stages.

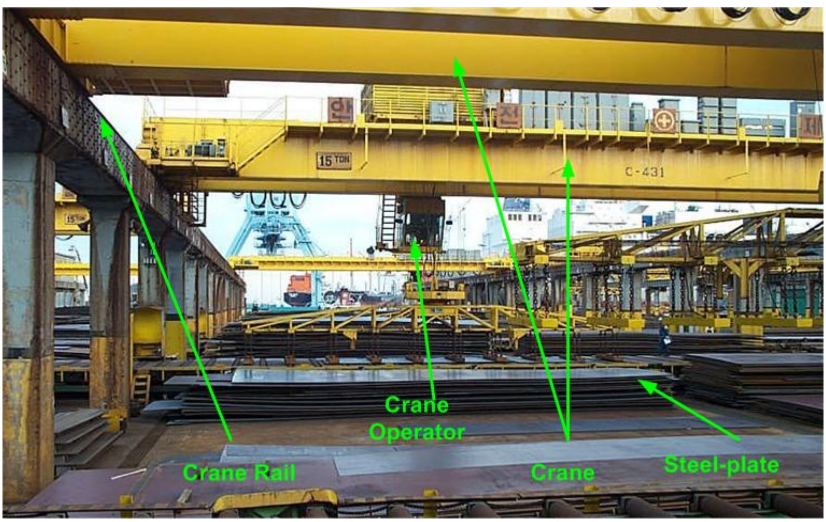


Fig. 1. View of steel-plate pile spaces (bay)

Piling errors in Stage (3) are the major cause of productivity degradation according to the comments of field engineers. In addition, the process poses safety vulnerability because the cranes deal with heavy steel-plates. The work plan becomes uncertain since stack information of the steel plates within pile space is not managed. So, if a steel-plate is lost, discovering it from the steel-plate heap is time-consuming. To resolve the problems, we constructed a smart space on the simulator that can simulate various activities of the process. We implemented a context-aware operation model that provides workers with intelligent services pertinent to the working situation. The system ensures safety for workers by managing the contexts of workers as well as heavy objects. It also improves work efficiency by enabling an effective work plan with the management of the steel-plate stack information in the pile spaces.

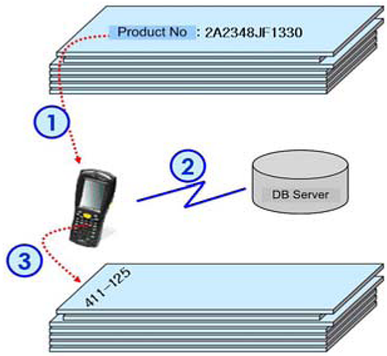
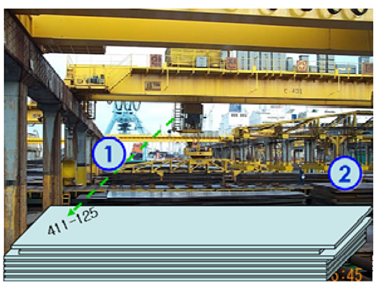
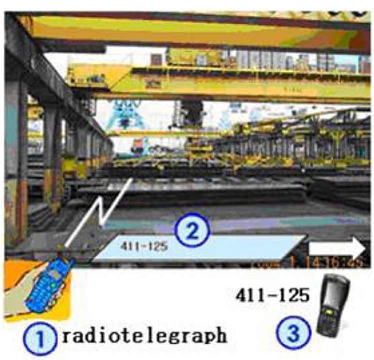
Workers		Actions
Pile-in Manager		① Read the *plate product number* (2A2348JF1330) on the stencil mark and input it into PDA ② Receives *pile space address* (411) and *serial number* (125) from DB ③ Writes *pile space address* and *serial number* in a chalk on the steel-plate
Piling Manager (Crane Operator)		① Confirms the pile space address on the steel-plate by eyes from the *crane* seat ② Lifts the plate, move it to the *corresponding* pile space, and put it down in it.
Pile-out Manager		① By using a radiotelegraph, notifies crane operator of a pile space address whose plates are to be delivered to the preprocessing stage ② Moves the plates onto conveyor belt one by one ③ Types in pile space address and serial number that was moved out to update DB

Fig. 2. Current piling process

2 Background

2.1 Steel-Plate Piling Bay and Its Components

A steel-plate pile bay measures 37 meters by 250. It consists of about 8 t-pile spaces and several tens of pile spaces, multiple conveyor belts, and multiple overhead cranes moving along the rails. Every pile space is assigned its own unique address. A t-pile space holds steel-plates waiting to be sorted into pile spaces. Each pile space holds steel-plates used to make a particular ship block. Each steel plate measures about 3 meters by 15 with a thickness of 15 mm and weighs up to 3,500 kg.

Table 1. Problems in the steel-plate piling process

Steps	Problems
pile-in	- Typing error of product number. - Typo error when writing on a steel-plate with a chalk.
piling	- Misidentification of the pile space address by the crane operator because of reading it from a long distance. - Piling in a wrong place by careless mistake.
pile-out	- Waste of manpower since a pile-out manager performs a simple operation to confirm a pile-out completion.

2.2 Current Steel-Plate Piling Process and Problems

The steel-plate piling process is divided into *pile-in*, *piling*, and *pile-out* transactions which are performed by a pile-in manager, a crane operator, and a pile-out manager, respectively. The current piling process is illustrated in Fig. 2.The pile-in manager reads a stencil (that includes product number, quality, mark number, ship number and so on) printed on each steel-plate by eyes and sends its product number as a search key to a DB server by using a PDA. The DB server searches for the detailed information (e.g., manufacturer, quality, weight, size, manufacturer, etc), *pile space address* to pile the steel-plate, and *serial number* that is a sequential number of the steel-plate when it is piled within the selected pile space, and then replies to the PDA with the retrieved information. The pile-in manager confirms information on the PDA and hand-writes pile space address and serial number on the steel-plate with a chalk. A crane operator drives the crane toward the t-pile spaces to read the numbers from the crane seat by eyes. Note that he may read them erroneously since the crane seat is located about 7~8 miters up in the air. He lifts the steel-plate, moves to the corresponding pile space address, and puts it down in the space. These three transactions take place independently of each other's one.

The manager of the pile-out transaction notifies a crane operator of the list of pile space addresses whose steel-plates need be delivered to the pre-processing stage. The crane operator moves steel-plates onto the conveyor belt one by one after selecting one pile space from the list until it processes all steel-plates in the pile space. He repeats the above procedure. Whenever each steel-plate is put down on the conveyor belt, a pile-out manager inputs both pile space number and serial number into the PDA that transmits the data to the DB server to finish the pile-out of the steel-plate. The problems that occur in the current piling process are summarized in Table 1.

2.3 Work Simulator

It may not be reasonable to develop a system in the field because of the following difficulties. It is not easy to stop on-going work even for a moment due to a tight work schedule in a shipyard. It is difficult to execute work scenario that takes into account various abnormal or erroneous conditions. It is not easy to test applicability of various required H/W components such as special sensors and tags. Finally, developers do not prefer staying in the field for a long time.

For these reasons, we contrived the steel-plate piling simulator as shown in Fig. 3 that can simulate various transactions pertinent to the steel-plate piling process of a shipyard. We could develop and test system components and prove the functional correctness of the integrated system by using this simulator.

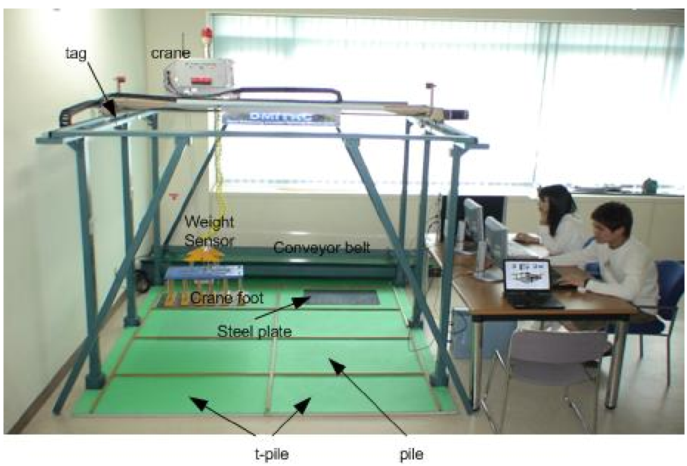


Fig. 3. Work simulator

This simulator of size 3 x 2.25 x 2 (m^3) is comprised of one conveyor belt and 8 pile spaces (2 t-pile spaces and 6 pile spaces), but can simulate the various transactions required in the steel-plate piling process. Two parallel overhead frames and the suspended overhead frame have rails installed on them. So, the crane on the suspended overhead frame can move back and forth or right and left along the rails. The crane operated by a remote controller can lift the steel plate and put it down after being moved, and also carry the steel plate from a pile space to a conveyor belt. If it puts the steel-plate down on the conveyor belt, the conveyor belt can sense the steel plate and carry it outside (i.e. to the pre-processing stage in real field) automatically.

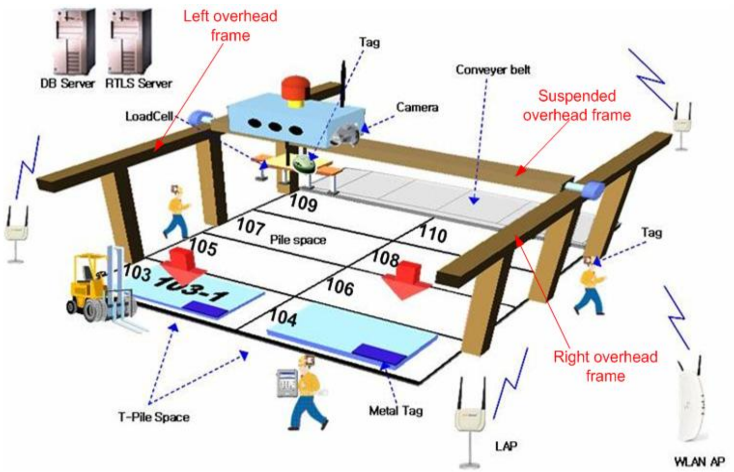


Fig. 4. Graphics screen corresponding to the work simulator (screen capture)

3 U-SPPS

The simulator in Fig. 3 contains various components to build a smart space so that any context change in the simulator is updated in the context DB in a real-time manner and the movements of the physical objects in the space are synchronized with those of the virtual objects on the 3D graphics screen in Fig. 4.

3.1 Components of U-SPPS

A PDA equipped with an RFID reader reads the steel-plate ID from the metal RFID tag attached onto each steel-plate and sends it to the DB server; the metal tag was used for the testing purpose only because of its technical limitations such as its thickness and the fragility for a big impact. We used a weight sensor named *load cell* that measures the weight of a steel-plate and sends it to the system via 2.4 GHz mote interface. The RTLS (AeroScout, Ltd) applied to this simulator consists of 1 LAP, 2 Exciters, 6 RTLS tags and RTLS server. Two Exciters that transmit their ID are attached on the left and right edges of the suspended overhead frame. 5 RTLS tags are attached onto the sections of the overhead frame corresponding to each of 5 columns including the conveyor belt zone while one is attached on the crane head. In order to prevent two or more RTLS tags from getting the same Exciter's ID, these RTLS tags are attached onto two overhead frames in a zigzag form, the first tag onto the first column zone of the left overhead frame, the second one on the second column zone of the right one, and so on. By parsing information received from these tag devices, the system can tell the pile space on which the crane head is located. For example, if the crane head is positioned left, the RTLS tag on it receives the Exciter's ID on the left side of the suspended frame and sends its ID, its location, and its received Exciter's ID to the RTLS server. We also used the ULS (Ultrasonic Location System) to track the location of a mobile object more precisely (max. error of 5cm) within the simulator of small dimension; we can use RTLS in the field in which error range (maximum error of 3 m) is relaxed. Wireless camera was used to show the moving picture of a lifted steel-plate and its property information (pile space, serial number, mark number, ship number, weight, size, etc) in overlap. So, the operator can see the properties of the steel-plate in real time. Lastly, voice and 3D graphics interfaces were used. The crane moves a steel-plate from a t-pile space to a pile space or from a pile space onto a conveyor belt. A ULS receiver is attached on crash helmet of ground worker to keep track of their location. The components are summarized in Table 2.

Table 2. Key components and their uses

Key Components		Specification	Use
Metal RFID Tag		Frequency: 900MHz Max. Reading Distance: 1 m	- Steel-plate identification
Load Cell Weight Sensor + Sensor Mote		Mote interface: RS-232 Mote OS: Tiny OS Mote Frequency: 2.4GHz Max. Trans. range: 50 m	- Measures the weight of a steel-plate
Real-Time Location System (RTLS)	LAP (Location Access Point)	Radio: 2.4GHz DSSS 802.11b Outdoor range: up to 200m Maximum error range: 3m	- Receives a message from RTLS tags and relays it to RTLS server
	Exciter	Programmable trans. range: 1.5~2m	- Transmit ID periodically - RTLS tag receives the ID when it is within Exciter's transmission range
	T2 Tag	Outdoor Range: <= 200m Read range(125KHz): 0.5~7 m Trans. period: 6.5 ms ~ 3.5h	- Transmit its MAC address, location information, and the received Exciter ID if any to the LAPs
RTLS Server		System: Desktop PC Software: RTLS software	- Manage RTLS information
DB Server		System: Desktop PC DB: MS SQL	- Manage information of steel-plate ordering, context in work space and steel-plate stack in each pile space
ULS (Ultrasonic Location System)	1 Receiver 4 Transmitters 1 Sync Generator	Frequency: 40 KHz RF Sync: 315MHz Coverage range: 7 m Maximum error range: 5cm	- Keep track of the location of ground worker and steel-plates

3.2 Context-Aware Model

Fig. 5 shows the context-aware system structure used in the U-SPPS whose key components are context library, context manager, and service provider. Fig. 6 shows the example data structure of context library. The context library is the collection of contexts that expresses service knowledge in the form of 5W1H (who, when, where, what, how, why) [11]. The operation knowledge context defines the knowledge related to the pile-in and pile-out transactions. The situation knowledge context defines a situation that can occur and indicates the conditions under which the situation can take place. The service knowledge context defines service functions suitable for situations represented by a current context.

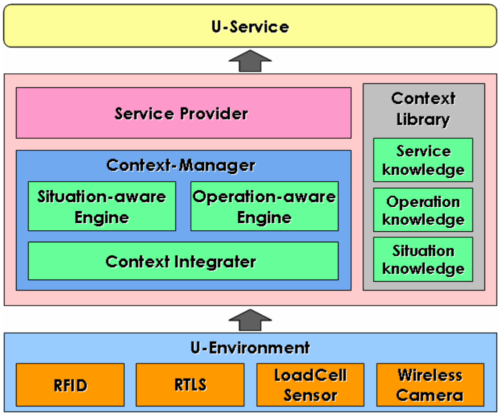


Fig. 3. Context-aware system structure

	Who	When	Where	What	How	Why
Operation Knowledge	crane	*	t-pile space	*	lift	piling
	crane	*	conveyor belt	*	put-down	pile-out
	•••					
Situation Knowledge	worker	T ± 3s	(a±2, b±2)	*	*	*
	•••					
Service Knowledge	crane	*	t-pile space	ObjectNum	lift	piling
	InputGuideSound(ObjectNum), DisplayUpdate(), SendMsg(SPPS* msg)					
	•••					

T : Location Measurement Time of Crane, (a, b) : Crane's Location, * : Any Value

Fig. 4. Context library data structure

The context manager analyzes, synthesizes, and processes the context data that the tags or sensors provide. The context integrator represents context information in the form of 5W1H by grasping and extracting correlations from the received data. However, a complete context information cannot be made with only the data from sensors since there is no "Why" information that allows us to be aware of a worker's intention. We infer the "Why" information from the comparison of the incomplete context and the operation knowledge contexts of the context library. The complete context is stored in the DB server and accessed by the service provider. The service provider provides an intelligent service suitable for situations by exploiting the completed context. An appropriate service function is invoked by comparing the completed context and the service knowledge contexts.

The example in Fig. 7 illustrates four stages for context generation, management, and service. First, the situation data are compiled from the UbiCom space as shown on the top of the figure. In the second stage, the compiled data is transformed into a 4W1H form of "a crane lifted steel-plate 103-1 at 16:31 at the coordinate of (12, 15)", which is an incomplete context because it does not have "Why" information. The context manager in Fig. 5 is aware of the piling request corresponding to the "Why" information by comparing this incomplete context with the operation and situation knowledge contexts, and then creates a completed context as in the third stage of Fig. 7. Lastly, the service provider fires the service function corresponding to the context by comparing it with the service knowledge of the context library.

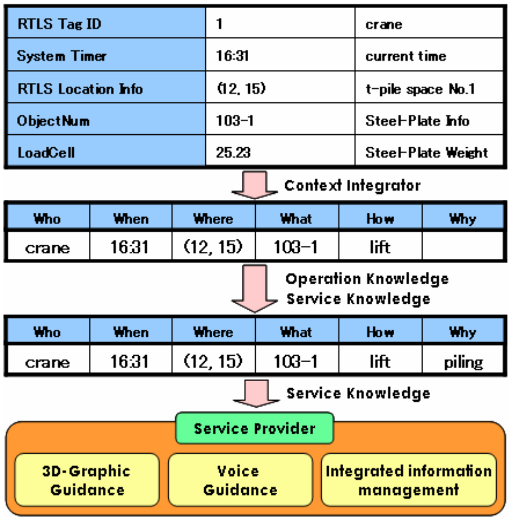


Fig. 5. Example of context compilation management and service

3.3 Steel-Plate Piling Process in U-SPPS

A pile-in manager approaches the metal tag attached on the steel-plate and reads the steel-plate product number by using an RFID reader attached to PDA. When he presses [Send] button, his current location is identified by the ULS receiver on his crash helmet and becomes the location of the steel-plate to be piled in. The location is recorded in the context DB. So, the current steel-plate whose piling is requested and the pile space to pile are shown in the 3D screen of a crane operator by two flickering red arrows. Fig. 4 requires the crane operator to move the steel-plate in t-pile address 103 to address 106. If a pile-in manager moves to another steel plate and types the product number in PDA, two steel plates to be piled are listed on the screen.

If a crane operator picks a steel-plate up, load cell sensor senses the weight and displays it on the front LCD panel, and then sends it to the context DB. The system compares the measured weight and the one stored *a priori* in the DB and can know that the crane lifted a correct steel plate. If it happens to lift two steel-plates by mistake, an error (voice) message is given. As a crane operator carries the steel-plate to the pile space indicated by the other red arrow, its changing pile space location is updated in the context DB in real time. So, if he tries to put it down in a wrong pile space, an error (voice) message is also given.

A pile-out manager sends a list of pile spaces to be processed to a crane operator. The crane operator selects one from the list and moves a steel-plate from the pile space to the conveyor belt one by one. Whenever he moves one, he clicks the steel-plate number on the screen. Then, the context DB is updated immediately for each pile-out completion. If the crane operator puts the steel-plate down on a conveyor belt, the system guides the crane operator to next transaction by flickering red arrows with voice guidance. A pile-out manager is no more needed.

In this way, the U-SPPS detects piling errors and provides workers with intelligent services by exploiting the context information updated in real-time. The system also gives warning signs to workers if they enter dangerous spots.

4 Conclusions and Further Research Directions

Applying new technology in manufacturing fields is cumbersome because of the limited accessibility to the field. To overcome this limitation, we made a simulator that can simulate the steel-plate piling process. Based on the simulator, we constructed the steel-plate processing system that guides workers for next action depending on the current situation and then examined our system by applying various operation scenarios. We are working on the construction of an unmanned U-SPPS system that can remotely operate the piling work and monitor field situations.

References

1. Yeongjong BMS, http://www.yeongjongbridge.com/english/budae/10-budae.asp
2. Weiser, M.: Hot Topic-Ubiquitous Computing. IEEE Computer 26, 71–72 (1993)
3. Weiser, M.: The Computer for the 21st Century. Scientific American, pp. 94–104 (1991)
4. Oh, Y., Jang, S., Woo, W.: User Centered Context-Aware Smart Home Applications. Journal on Korea Information Science Society: Software and Applications 31(2), 111–125 (2004)
5. Shafer, S., Brumitt, B., Meyers, B.: The EasyLiving Intelligent Environment System. In: CHI Workshop on Research Directions in Situated Computing (2000)
6. Shafer, S., Krumm, J., Brumitt, B., Meyers, B., Czerwinski, M., Robbins, D.: The New Easy-Living Project at Microsoft Research. In: DARPA/NIST Workshop on Smart Space (1998)
7. Fox, A., Johanson, B., Hanrahan, P., Winograd, T.: Integrating Information Appliances into an Interactive Workspace. IEEE Computer Graphics and Applications 20(3), 54–65 (2000)
8. Dey, A.K., Abowd, G.D., Salber, D.: A Context-based Infrastructure for Smart Environments. In: Proceedings of the 1st International Workshop on Managing Interactions in Smart Environments (MANSE 1999), pp. 114–128 (1999)
9. HP CoolTown, http://www.cooltownstudios.com
10. MIT Oxygen, http://oxygen.lcs.mit.edu/overview.html
11. Abowd, G.D., Mynatt, E.D.: Charting Past, Present, and Future Research in Ubiquitous Computing. ACM Trans. on Computer-Human Interaction 7(1), 29–58 (2000)

The Choquet Integral Analytic Hierarchy Process for Radwaste Repository Site Selection in Taiwan

Chen Lin

Department of Business Administration,
Kainan University,
No. 1 Kai-Nan Rd., Luchu, Taoyuan 338, Taiwan
lin.chen@mail.knu.edu.tw

Abstract. The Radwaste Repository Site Selection In Taiwan have received considerable continuous improvement by the fuzzy Choquet integral aggregation operator. It allows expressing scenario that are either hosted externally by the repository safety assessment or by the hosting facility which requires a secure zone for its radioactive waste repository.

Radioactive waste repository site selection is typical of society-based applications Taiwanese expert analysis has been conducted on different strategic courses of action concerning governmental and society-approved solutions. AHP method makes it possible to evaluate several strategies and to synthesize qualitative and quantitative factors in the decision-making process. This article is a contribution with a Choquet Integral aggregation for optimal improvement. Based on this study it is possible to draw conclusions concerning the best strategic course for a Taiwanese radwaste repository site selection.

Keywords: Radioactive waste repository , Choquet Integral, AHP method.

1 Introduction

Previous study[2] use original AHP in investigating site selection. This report can provide a Choquet Integral aggregation for optimal improvement to integrate the previous one in order to find a better method if possible. The AHP was used as a decision-aiding tool during a time of wide public debate in Taiwan on the future of nuclear policy. The subject of the debate was whether or not a new nuclear power plant should be built. A decision analysis of the issue was carried out to clarify the differences between anti-nuclear and pro-nuclear opinions for the public [7]. The AHP was conveniently used in solving a sitting location problem [8].

The site location choice is actually a complex decision-making problem. An AHP can effectively deal with both qualitative and quantitative factors in multiple criteria decision environments. AHP is an important decision tool because of its ability to synthesize multi-attributed scenarios and provide diagnostic information, which enables decision makers to better understand the behavioral processes underlying certain choices. The AHP is capable of handling multiple conflicting goals inherent in

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 634–646, 2008.

site location; consequently, it helps the policy maker to analyze various tradeoffs between competing criteria such as costs versus environment, especially if the government should shift its priorities in site evaluation. Through sensitivity analyses, the AHP enables the policy maker to evaluate 'what-if' scenarios associated with changes in the government policy [5][14].

Radioactive waste disposal is a scientific and social issue of highly-charged significance amongst today's many global challenges. As other concerns rise to the forefront of public consciousness, however, the problem of radioactive waste appears at times to be pushed to the periphery of public concern. While issues like global warming and the oil crises tend to be headline-grabbers and typically garner most of the public interest at any given moment, the serious threats posed by radioactive waste remain a central concern for experts and specialists alike who are all too aware of the dangers involved to both local communities and the global village at large. This paper takes the position that radioactive waste disposal ought to occupy a pivotal position in governmental and societal concerns. As a consequence, it seeks to offer a viable solution to the problem of disposing of radioactive waste in the best safe and efficient manner possible, and to recommend such a procedure be seriously considered by the appropriate governmental bodies. While Taiwan is chosen as the example nation to be analyzed, such a policy as advocated in this paper could have applications in another of other world nations.

Radioactive waste was not always considered a problem. The disposal of radioactive waste became a controversial scientific and social issue in Taiwan only after the Atomic Energy Commission (AEC), charged with the development, regulation and promotion of nuclear technology, tried to implement a program of disposal technology in 2016. Before this time, the issue occupied a very low-level position of priority on the table of contemporary social issues. The present report is a result of a feasibility study of site-selection model conducted in 2003 out of personal interest. It traces the scenario consequences of this controversy as part of the long-term development in radioactive waste disposal technology. Problem definitions and technical designs that underlie this technology are constructed, and it is thus possible to show how some definitions received attention and others not, and how some became, and remained, dominant. This is a political process, and by heuristically using scenario deployment from political science and recent technology studies we can pave the way for future action.

Radioactive waste disposal in every case encounters many serious problems. It is more complex and risky than sanitary landfill, and many factors influencing its decisions conflict with one another [6]. Effective radioactive waste disposal site planning not only must deal with a host of quantitative and qualitative variables but also must consider multiple conflicting factors.

2 Disposal of LLRW Management in Taiwan

Low-level radioactive waste (LLRW) presently stored on the Lan-Yu Site have to be permanently disposed in a safe manner. The disposal site can then be freed from institutional control. Taipower Company(TPC) estimate about 910,000 LLW drums until 2049, of which over 97,000 drums are on Orchid Island. According to AEC, there are no laws prohibiting TPC from disposing of nuclear waste at a foreign site. TPC alone is responsible for conditioning and disposal of its waste. AEC must approve the utility's plans before they can be implemented. The Regulations also point out a set of site requirements for the final disposal program. When selecting final disposal sites, natural and social environmental characteristics like geology, hydrology, geochemistry, population and social development shall all be considered, which shall not influence the safety of the disposal facilities. AEC has clearly specified in " Law on Site Selection for Low Level Waste Final Disposal" that the disposal facilities must avoid the following areas:

1. Area where active faulting or geological conditions could endanger the safety of the disposal facility.
2. Area where the geochemical conditions are unfavorable for effectively suppressing the diffusion of radioactive nuclides, and it is likely to endanger the disposal facilities.
3. Area where the hydrologic conditions of surface water or groundwater are likely to endanger the disposal facilities.
4. Area of high population density.

The program for LLRW disposal will be carried out five stages of regulatory control :

1. Site selection: Select potential sites through regional survey and initial site investigation, and then select one or more candidate sites from the list. During the selection, shall communicate and negotiate with local government and people, finally, the results of local negotiation and conclusion of environmental impact evaluation shall be submitted to the Executive Yuan for approval.
2. Environmental impact assessment: To prevent and mitigate environmental influence from the development of disposal facilities, the environmental impact analysis shall be carried out on candidate sites. Only after the review and approval by the competent authorities of environmental protection, can the development of the disposal sites be started.
3. Construction: TPC shall submit the safety analysis report of the disposal facility to AEC for construction license application. The work shall be started after the construction license is issued. During the construction, the Fuel Cycle and Materials Administration (FCMA) will dispatch inspectors to ensure quality of the construction.
4. Operation: Upon completion of the construction, the operator shall first submit a test run plan to AEC for approval. Upon completion of the test run, the operator shall submit the latest safety analysis report, facility operation technical specifications, test run report and incident response plan to AEC for operation license application. During the operation, FCMA will send inspectors for safety inspection and environment monitoring to ensure safety of the waste disposal.
5. Closure and post-operation monitoring: When the disposal capacity is full, the operator shall submit the closure plan and the institutional control plan and shall implement the two plans after they have been approved by AEC. Once the radioactivity is decayed to harmless level,

the operator shall apply to AEC for approval of the re-utilization or the exemption from institutional control of the land where the facility is located.

TPC has set up guidelines on the financial benefits for communities that volunteer to host LLRW disposal facilities. According to the guidelines, once a disposal site has been given the government go-head, TPC will offer negotiated benefits totaling one hundred million US dollars to the hosting and affected areas. A number of communities voiced their interest. However, after disclosure by the media, every community, without exception, withdrew their letters of consent as a potential host site for LLRW at the last minute before signing an agreement with TPC. A voluntary siting program thus seems to have been proven unsuccessful. In its place, a state-wide technical screening program has being launched. TPC has illuminated the selection process and criteria to the public, and has also drafted a communication program.

To promote the safe disposal of low-level radioactive waste, the AEC has set up a panel of specialists to oversee the implementation plans of radwaste final disposal. Government policy also mentions regional co-operation (i.e. finding a site outside Taiwan as an option). However, in recognition of the technical, economic, political and environmental problems associated with this option, we are currently continuing to pursue a strategy of disposal within Taiwan.

3 Choquet Integral AHP

3.1 Choquet integral

In this section we present our Choquet integral AHP.

Definition I: Let X={ $x_1, x_2, \ldots, x_n$ } denote the attribute set in a traditional AHP. We define the interaction between attributes as a $n \times n$ matrix C = $(\alpha_{ij})_{1 \le i \le n, 1 \le j \le n}$,

$$C = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix}$$

Where $\alpha_{ij} \in [-1,1]$ is the interaction coefficient between attributes i and j. It represents the degree of conflict between attributes i and j. We call the matrix C the Interaction Matrix.

If $\alpha_{ij} > 0$ the attributes i and j enhance each other or "they are cooperative"; if $\alpha_{ij} < 0$ then the attributes i and j are in conflict or in "competition"; if $\alpha_{ij} = 0$ then the attributes i and j are mutually neutral.

Assume now that the traditional AHP [7] pair-wise comparison matrix P= $(w_{ij})_{1 \le i \le n, 1 \le j \le n}$ is given. The Choquet Integral AHP is characterized by the triplet (X, P, C), where X and C are introduced in Definition 2.1 and P is the pair-wise comparison matrix given above. Now we will show how the attribute weights are calculated in Choquet Integral AHP. Let w_i represent the normalized weight of attribute i in the traditional AHP [7] and let $h(x_i) = w_i$, i= $\overline{1..n}$ Let H^1 = { x_1 }, H^2 ={ x_1, x_2 },..., H^n = { x_1, x_2,..., x_n }=X and $h(x_1) \geq h(x_2) \geq .. \geq h(x_n)$, if this ordering doesn't hold, one can just reorder the attributes accordingly.

In order to take into account the interdependence of attributes we will use the Choquet Integral to calculate the new weights of attributes. Let us assume, for the time being, that the λ-fuzzy measure g describing the interdependence of attributes is known (identification will be done in the next section). Now, according to [17], the Choquet Integral of the function $h(x_i)$ is

$$\int h \; dg = h(x_n)[g(\mathrm{H}_n)-g(\mathrm{H}_{n-1})]+h(x_{n-1})[g(\mathrm{H}_{n-1})-g(\mathrm{H}_{n-2})]+\ldots+h(x_1)g(\mathrm{H}_1)$$

This Choquet Integral gives an aggregated evaluation of the effect of interdependence of attributes on the weights $h(x_i) = w_i$, i = $\overline{1,n}$ given by the traditional AHP where attributes are assumed independent. Then based on the above Choquet Integral, it is natural to consider the new weight of any attribute i as h(x_i)[g(H^i)-g(H^{i-1})], that is, the corresponding term in the Choquet Integral. Furthermore, the coefficient [g(H^i)-g(H^{i-1})] can be interpreted as the effect of the interdependence of attributes on the weight $h(x_i) = w_i$ of the attribute i. Let us consider the vector W^f = (w_1^f, w_2^f ,..., w_n^f) with the following components (1):

$$\begin{aligned} w_1^f &= h(x_1)\, g(\mathrm{H}_1), \\ w_2^f &= h(x_2)\, [g(\mathrm{H}_2)-g(\mathrm{H}_1)], \\ &.., \\ w_n^f &= h(x_n)\, [g(\mathrm{H}^n)-g(\mathrm{H}^{n-1})] \end{aligned} \tag{1}$$

Definition II: We define the vector W$^{f'}$ = ($w_1^{f'}$, $w_2^{f'}$,..., $w_n^{f'}$) of weight attributes in Choquet Integral AHP as the normalized vector of W^f . Thus, we have

$$w_1^{f'} + w_2^{f'} + \ldots + w_n^{f'} = 1,$$

$$\text{where } w_i^{f'} = \frac{w_i^f}{\sum_{j=1}^{n} w_j^f}, \; i = 1..n \tag{2}$$

Next we give a procedure that determines the weights by our Choquet Integral AHP.

3.2 Procedure

Step 1. Ask the decision maker to provide the traditional AHP pair-wise comparison matrix P, the interaction matrix C (definition I) and the value of λ (additivity degree).

Step 2. Determine the density of the λ-fuzzy measure by solving the optimization problem (5). It is to be noted that the mathematical expression of the constraint $g(\{x_1, x_2, ..x_n\}) = 1$ can be determined using (4).

Step 3. Use formula (4) to calculate g(H^i), $i = \overline{1..n}$, then calculate the components of the vector W^f by (1), where $h(x_i) = w_i$, $i = \overline{1..n}$ and w_i represents the normalized weight of attribute i in the traditional AHP [7].

Step 4. Normalize the vector W^f to obtain the vector $W^{f'}$, which is also the vector of attribute weights in our Choquet Integral AHP$.

We will apply the Procedure to the three levels of hierarchy of the decision making problem: Level 1, Level 2 and Level 3. The classical AHP weights have been already calculated, we need to determine interaction matrix C (definition I) and the value of λ (additivity degree) for each of the three levels, C_1, C_2, C_3 and $\lambda_1, \lambda_2, \lambda_3$ respectively.

4 The Decision Aid Process

Decision aids and analytical models can be very useful in resolving public policy issues if utilized early in the decision-making process. Members of radwaste experts, regulators, and the public in Taiwan have to evaluate the desirability of a radwaste repository site. Saaty's AHP was chosen for the analysis because it uses the pairwise comparison method to both judge the different criteria and to rate the options with respect to these criteria. It has been used in multi-criteria decision making (MCDM) to release the restriction of hierarchical structure, and has been applied to project selection [10][12], product planning, strategic decision [9][16], optimal scheduling [11], and so on.

The key criteria were: 1. Economic consideration of repository selection, 2. Environmental impact, health hazard and risk concerns, and 3. Political concerns about possible selection. The decision options were: 1. main island,. 2. off-island, and 3. regional cooperation.

People make 3 general types of judgments to express importance, preference, or likelihood and use them to choose the best among alternatives in the presence of environmental, social, political, and other influences. They base these judgments on knowledge in memory or from analyzing benefits, costs, and risks. From past knowledge, sometimes standards of excellence and poorness can be developed and used to rate the alternatives one at a time. The AHP includes both the rating and comparison methods. Rationality requires developing a reliable hierarchic structure or feedback network that includes criteria of various types of influence, consequences, and decision alternatives to determine the best choice.

The decision hierarchy (Figure 1) was developed in cooperation with the leading radwaste disposal policy committee in Taiwan . The overall site selection criteria is the focus of the hierarchy. The focus is divided into three aspects which serve as elements of a second level, that is socio-economy, environmental impact, and finally, political consideration. These three broadly stated main criteria are each in turn broken down into three sub-criteria, which serve as criteria for rating the decision alternatives.

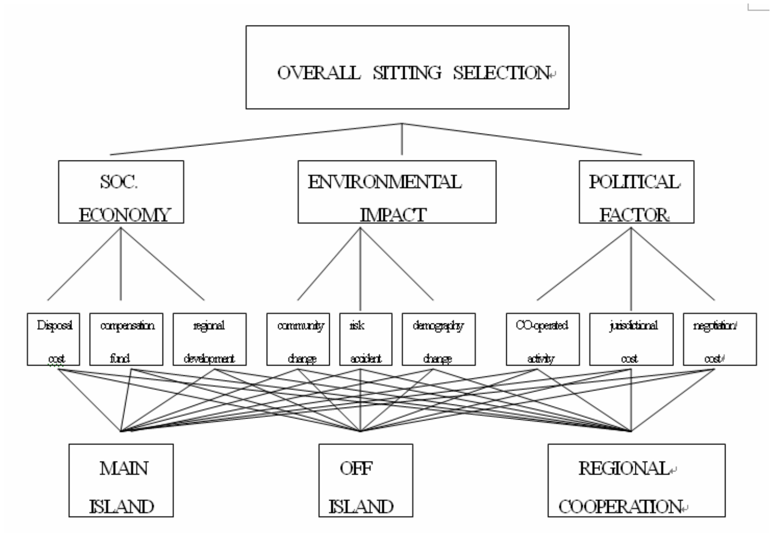


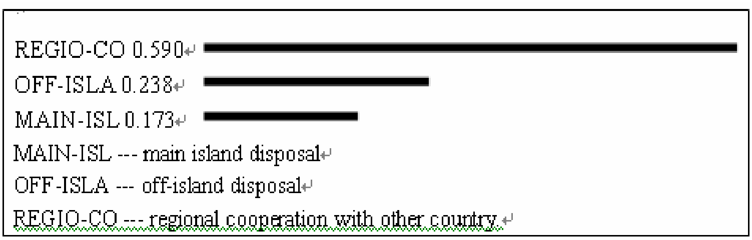
Fig. 1. The radwaste disposal policy decision hierarchy in the site-selection process

The three decision alternatives, which comprise the lowest level of the hierarchy, are the following: (1) Main island, main Taiwan island will be selected for radwaste disposal site-selection. (2) Off-island, small islands near Taiwan will be selected for radwaste disposal sitting-selection. (3) Regional cooperation, foreign country will be selected for radwaste disposal site-selection. The alternatives were pair-wise compared with respect to the goal. These comparisons were prepared by the radwaste disposal policy committee and, Table 1 shows the final weights of the each level with respect to the goal. These results were obtained using the AHP tool “Expert Choice”[15]. The AHP analysis showed that regional cooperation is the top choice overall with other alternatives. The results of the analysis as shown in Table 2.

Table 1. Sorted Details For Synthesis Of Leaf Nodes With Respect To Radwaste Repository Site Selection

LEVEL 1	LEVEL 2	LEVEL 3
ECONOMY =0.498		
	COST =0.329	
		REGIO-CO =0.329
		OFF-ISLA =0.076
		MAIN-ISL =0.023
	REGIONAL =0.135	
		REGIO-CO =0.135
		OFF-ISLA =0.030
		MAIN-ISL =0.010
	FUND =0.033	
		REGIO-CO =0.033
		OFF-ISLA =0.019
		MAIN-ISL =0.007
POLITICA =0.367		
	CO-OPERA =0.205	
		REGIO-CO =0.205
		OFF-ISLA =0.117
		MAIN-ISL =0.040
	JURISDIC =0.117	
		MAIN-ISL =0.117
		OFF-ISLA =0.048
		REGIO-CO =0.019
	SERVICE =0.045	
		MAIN-ISL =0.045
		OFF-ISLA =0.025
		REGIO-CO =0.006
ENVIMPAC =0.135		
	RISK =0.093	
		REGIO-CO =0.093
		OFF-ISLA =0.023
		MAIN-ISL =0.008
	SOCIAL =0.022	
		REGIO-CO =0.022
		OFF-ISLA =0.005
		MAIN-ISL =0.002
	DEMOGPHY =0.020	
		REGIO-CO =0.020
		OFF-ISLA =0.005
		MAIN-ISL =0.002

Table 2. Synthesis of Leaf Nodes with respect to Radwaste repository site selection



In the previous section we assumed that the criteria were independent. In most of the real multiple-criteria decision-making problems, the criteria are dependent. In the present section we will use a new method [3bb] based on Choquet Integral to compute the weights of the criteria. This method takes the weights generated by classical AHP method as initial weights, then computes new weights that take into account the interdependence of criteria.

5 Summary

The construction of a radioactive waste repository may not, in the final analysis, be “acceptable.” It is consciously balanced against a perceived good in one sense, but "tolerated" where necessary. In general, people are risk averse, which has led to caution in the acceptance of risk in scenario approaches for radwaste site-selection decision making. Any political action must contend with that fact. In light of these sources of error and uncertainty, very often the only possible solution is the application of expert group judgment. Given this situation, attempts are being made to formalize and rationalize the use of expert group judgment. Its integration with formal scientific methods is not easy. However, an analysis of the use of AHP on radioactive waste disposal site selection planning shows that: 1. AHP is capable of handling the multiple conflicting goals inherent in selection process, 2. AHP enables evaluation of 'what-if' scenarios associated with changes in radwaste policy. 3. AHP has the capability to convert a normative procedure to a decision-making one.

References

1. Bowen, W.M.: A thurstonian comparison of the analytic hierarchy process and probabilistic multidimensional scaling through application to the nuclear waste site selection decision. Socio-economic Planning Sciences 29(2), 151 (1995)
2. Lin, C.: Scenario Deployment of the Analytic Hierarchy Process for A Study to the Radwaste Repository Site Selection in Taiwan. In: 2006 IEEE International Conference on Systems, Man, and Cybernetics., Paper ID: 206784, Taipei, Taiwan, October 8-11 (2006)
3. Chen, Y.W.: Implementing an Analytical Hierarchy Process by Fuzzy Integral. International Journal of Fuzzy Systems 3(3), 493–502 (2001)
4. Chen, Y.W., Larbani, M.: On a Generalization of AHP Method to the Case of interdependent Attributes by λ-Fuzzy Measure, Abstracts of the MOPG 2006 conference (2006)
5. Emshoff, J.R., Saaty, T.L.: Applications of the Analytic Hierarchy Process to Long Range Planning Processes. European Journal of Operational Research 10, 131–143 (1982)
6. Hedge, G.G., Tadikamalla, P.R.: Site selection for a sure service terminal. European Journal of Operational Research 48, 77–80 (1990)
7. Hamalainen, R.P.: A Decision Aid in the Public Debate on Nuclear Power. European Journal of Operational Research 48, 66–76 (1990)
8. Hegde, G.G., Tadikamalla, P.R.: Site Selection for a ’Sure Service Terminal. European Journal of Operational Research 48, 77–80 (1990)
9. Karsak, E.E., Sozer, S., Alptekin, S.E.: Product planning in quality function deployment using a combined analytic network process and goal programming approach. Comput. Ind. Eng. 44(1), 171–190 (2002)
10. Lee, J.W., Kim, S.H.: Using analytic network process and goal programming for interdependent information system project selection. Comput. Oper. Res. 27(4), 367–382 (2000)

11. Momoh, J.A., Zhu, J.: Optimal generation scheduling based on AHP/ANP. IEEE Trans. Systems Man Cybernet.- Part B: Cybernet 33(3), 531–535 (2003)
12. Meade, L.M., Presley, A.: R&D project selection using the analytic network process. IEEE Trans. Eng. Manage. 49(1), 59–66 (2002)
13. Cranwell, R.M., Guzowski, R.V., Campbell, J.E., Ortiz, N.R.: Risk Methodology for Geologic Disposal of Radioactive Waste Scenario Selection Procedure, NUREG/CR-1667, Sandia National Laboratories, Albuquerque, NM (1982)
14. Saaty, T.L.: Homogeneity and clustering in AHP ensures the validity of the scale. European Journal of Operational Research 72, 598–601 (1994)
15. Saaty, T.L.: Highlights and critical points in the theory and application of the Analytic Hierarchy process. European Journal of Operational Research 74, 426–447 (1994)
16. Sarkis, J.: A strategic decision framework for Green supply chain management. J. Cleaner Production 11(4), 397–409 (2003)
17. Sugeno, M.: Theory of Choquet Integral and Its Application, Doctorial Dissertation, Tokyo Institute of Technology (1974)
18. Zimmerman, H.J.: Fuzzy Set Theory and Its Applications. Kluwer-Nijhoff, Boston (1985)

Appendix: Identification of the λ-Value in Fuzzy Measure

In this section we show how to identify the λ-fuzzy measure.

Definition: Assume that X={x_1, x_2,…, x_n}. Let g be a fuzzy measure. If g verifies the following property[8]:

$$A, B \in \beta(X) \text{ and } A \cap B = \phi$$
$$\Rightarrow g(A \cup B) = g(A) + g(B) + \lambda\, g(A)\, g(B) \tag{1}$$

Where $\lambda \in (-1, +\infty]$ indicates the degree of additivity of the fuzzy measure, then the fuzzy measure g is called λ-fuzzy measure. It is to be noted that λ g(A) g(B) measures the additivity (dependence) between attribute subsets A and B.

Proposition 1. For any non empty subset A={x_{i_1}, x_{i_2},…, x_{i_p}} $\in \beta(X)$ we have

$$\begin{aligned} g(A) = {} & \sum_{s=1}^{p} g(\{x_{i_s}\}) + \lambda \sum_{s \neq j}^{p} g(\{x_{i_s}\}) g(\{x_{i_j}\}) \\ & + \lambda^2 \sum_{s\neq j \neq t}^{p} g(\{x_{i_s}\})\, g(\{x_{i_j}\})\, g(\{x_{i_t}\}) \\ & + \ldots\ldots\ldots\ldots\ldots \\ & + \lambda^{p-2} \sum_{s_1 \neq s_2 \neq .. \neq s_{p-1}}^{p} g(\{x_{i_{s1}}\})\, g(\{x_{i_{s2}}\}) .. g(\{x_{i_{sp-1}}\}) \\ & + \lambda^{p-1}\, g(\{x_{i_1}\})\, g(\{x_{i_2}\}) \ldots g(\{x_{i_p}\}) \end{aligned} \tag{2}$$

The proof of this proposition is based on the induction. Indeed, the formula is true for any couple of attributes by Definition. Assuming that the formula is true for any subset of cardinal p then one can easily prove that it is true for any subset of cardinal p+1.

Now the only problem remaining is the identification the fuzzy density of g, i.e. g({ x_i })= g_i , i = 1..n. We define the density g as a solution of the following multi-objective optimization problem

$$\begin{aligned} &Max\ \mathbf{C}\times\mathbf{G} \\ &\quad st\ \ g(\{x_1, x_2, ..x_n\}) = 1 \\ &\quad\ 0 \le g_i \le 1, \quad i = 1..n \end{aligned} \tag{3}$$

Where G= (g_1 , g_2 ,..., g_n) and the interaction matrix C is given in context's definition.

To solve the problem (3) we will use the Zimmermann's approach for solving fuzzy linear multiobjective problems [18].

First we construct the membership function of each attribute. We recall that in Zimmermann's approach, the decision maker is given the freedom to choose the pessimistic and optimistic values for the membership function of each objective. Since for each attribute i the objective of the decision maker is to maximize $\sum_{1}^{n} \alpha_{ij}\ g_j$ with $g_i \ge 0, \ \ i = 1..n$, then the values of the fuzzy density for which $\sum_{1}^{n} \alpha_{ij}\ g_j$ $\le$ 0 will not be interesting for the decision maker, therefore, it is rational to take the pessimistic value as 0. Since we have 0 $\le$ $g_i \le 1, \ \ i = 1..n$, $\alpha_{ij} \in [-1,1]$ and $g(X) = 1$, , then for an attribute i we have $\sum_{1}^{n} \alpha_{ij}\ g_j$ $\le$ α_{ii} + $\sum_{j \in I_i} \alpha_{i\,j}$ $\le$ 1+ $\sum_{j \in I_i} \alpha_{i\,j}$, where

I_i = {j/ $j \ne i$ and $\alpha_{ij} > 0$ }. Based on these considerations for each attribute i, we take the optimistic value as 1+ $\sum_{j \in I_i} \alpha_{i\,j}$. This means that the decision maker is completely satisfied when the function $\sum_{1}^{n} \alpha_{ij}\ g_j$ takes a value equal or larger than 1+ $\sum_{j \in I_i} \alpha_{i\,j}$. Hence, we obtain the following membership functions.

$$\mu_i^L\left(\sum_{j=1}^{n}\alpha_{ij}\ g_j\right)=\begin{cases} 0 & if \quad \sum_{j=1}^{n}\alpha_{ij}\ g_j\le 0 \\ \dfrac{\sum_{j=1}^{n}\alpha_{ij}\ g_j-0}{1+\sum_{j\in I_i}\alpha_{ij}-0} & if \quad 0\le\sum_{j=1}^{n}\alpha_{ij}\ g_j\le 1+\sum_{j\in I_i}\alpha_{ij} \\ 1 & if \quad \sum_{1}^{n}\alpha_{ij}\ g_j\ge 1+\sum_{j\in I_i}\alpha_{ij} \end{cases} \qquad i=1,\ldots\ldots,n \tag{4}$$

We obtain the following problem

$$\begin{aligned} & Max \quad \mu_i^L\left(\sum_{j=1}^{n}\alpha_{ij}\ g_j\right) \\ & st \quad g(\{x_1,x_2,..x_n\})=1 \\ & \quad g_i\le 1,\quad i=1..n \\ & \quad g_i\ge 0,\quad i=1..n \end{aligned} \tag{5}$$

Introducing the auxiliary variable γ, problem (57) is rewritten as follows:

$$\begin{aligned} & Max \quad \gamma \\ & st \quad \gamma\le\mu_i^L\left(\sum_{j=1}^{n}\alpha_{ij}\ g_j\right),\quad i=1,2,\ldots,n \\ & \quad g(\{x_1,x_2,..x_n\})=1 \\ & \quad g_i\le 1,\quad i=1..n \\ & \quad g_i\ge 0,\quad i=1..n \end{aligned} \tag{6}$$

According to (4), problem (6) can be written as

$$\begin{aligned} & Max \quad \gamma \\ & st \quad \gamma\le\frac{\sum_{j=1}^{n}\alpha_{ij}\ g_j}{1+\sum_{j\in I_i}\alpha_{ij}},\quad i=1,2,\ldots,n \\ & \quad g(\{x_1,x_2,..x_n\})=1 \\ & \quad g_i\le 1,\quad i=1..n \\ & \quad g_i\ge 0,\quad i=1..n \end{aligned} \tag{7}$$

It is to be noted that the mathematical expression of the constraint $g(\{x_1, x_2, ..x_n\}) = 1$ can be calculated based on (2). Thus, by solving problem (7) we obtain the fuzzy density g({ x_i })= g_i , i = 1..n..

Proposition 2. Based on the fuzzy density g({ x_i })= g_i , i = 1..n and the formulas (2), the set function g can be identified. In addition to this, the set function g is actually a λ -fuzzy measure, assuming that g($\varnothing$)=0.

One can easily prove that ,by construction, the set function g is a λ -fuzzy measure.

Contribution of Hypercontours to Multiagent Automatic Image Analysis

P.S. Szczepaniak[1,2]

[1] Institute of Computer Science, Technical University of Lodz, ul. Wolczanska 215, 90-924 Lodz, Poland
[2] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland

Abstract. Hypercountors are methods for automatic image analysis. They can be interpreted as contextual classifiers that use expert knowledge and operate in supervised or unsupervised mode. In the present paper, it is shown that hypercontours may act as agents cooperating in recognition of objects placed on computer images.

Keywords: Image analysis, agents, machine learning, active contours.

1 Introduction

Within the field of information technology there are diverse definitions of agents. Here we quote the one by IBM as given in [1]: Intelligent agents are software entities that carry out some set of operations on behalf of a user or another program with some degree of independence or autonomy. In so doing, they employ some knowledge or representation of the users goals or desires.

A software agent can be called intelligent only if it implements methods of artificial intelligence; in this case the notion distributed artificial intelligence is used [2]. Here, a number of classic and modern techniques can be applied: neural computing, crisp and fuzzy logic, evolutionary methods, rough computing, etc. In other words, we may argue that agent technology is a way for building the task oriented ensembles.

Diverse tasks in diverse fields may be subject of distributed solution, such as e-commerce, finance, ecology, marketing, robotics and computer vision, to name but a few. The number and variety of tasks understood as formalized problems is not restricted in practice; let us mention planning, control, decision making, recognition, and the like. The complexity of the problem is the reason for looking of assistance in the area of distributed problem solving.

Research areas which to a certain extent are related to agent systems are [3]: artificial life, agent-based simulation, self-organized criticality, ecology of computation, distributed constraint satisfaction problem and autonomy oriented computing.

Here, a concept of using hypercontours as cooperating entities analyzing the image content is proposed. The system should recognize a class of predefined

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 647–653, 2008.

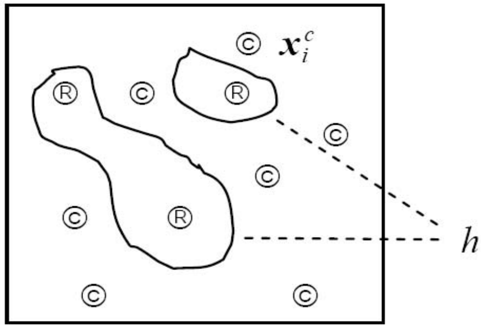


Fig. 1. Hypercontour h consisting of two contours determined by control points {ⓒ,Ⓡ}; binary classification case

shapes. The present paper is organized as follows: in section 2 adaptive potential active hypercontours are briefly described, in section 3 architecture of the system is presented and shape recognition is exemplified. Concluding remarks summarize the papers content.

2 Potential Active Hypercontour

The content of this section is also published in [4]; here it is given for completeness of the presentation.

Potential active hypercontour (PAH) is an implementation of *active hypercontour* introduced in [5]; see also [6,7].

Let X - called feature space - be a metric space with metric $\rho : X \times X$.

The potential hypercontour h is determined by set D^c of labeled *control points* $\mathbf{x}_i^c$ and assigned to them potential functions P_i^c, $i = 1, 2, \ldots, N^c$ (Fig. 1).

Each point $\mathbf{x}_i^c$ is a source of potential (similar to the electrical one) the value of which decreases as the larger the distance from that point increases. There are many ways of defining of P_i^c, for example:

– exponential potential function

$$P_i^c(\mathbf{x}) = \Psi_i \exp\left(-\mu_i \rho\left(\mathbf{x}_i^c, \mathbf{x}\right)\right) \tag{1}$$

– inverse potential function

$$P_i^c(\mathbf{x}) = \frac{\Psi_i}{1 + \mu_i \rho\left(\mathbf{x}_i^c, \mathbf{x}\right)} \tag{2}$$

where real numbers Ψ_i and μ_i are parameters characterizing the potential field of each control point.

Note that the number of control points, their distribution and field parameters fully determine the shape of hypercontour.

Since each control point is labelled, i.e. one dispose of pairs $(\mathbf{x}_i^c, l_i^c)$, then h is in fact a classifier and it can be formally defined as follows:

$$\kappa(\mathbf{x}) = \arg\max_{l \in L} \sum_{i=1}^{N^c} P_i^c(\mathbf{x}_i^c, \mathbf{x})\, \delta(l_i^c, l) \text{ for each } \mathbf{x} \in X \tag{3}$$

where L is set of labels, l_i^c, l denote labels of $\mathbf{x}_i^c$ and the examined one while $\delta(l_i^c, l) = \begin{cases} 0 & l \neq l_i^c \\ 1 & l = l_i^c \end{cases}$

Potential active contour (PAH) possesses an ability to *evolve* with the change of the location or number of control points, and with modification of parameters of potential functions. The search of optimal hyercontour is performed by optimization of some performance index E called energy in the theory of active contours; $E : H \to R^+ \cup \{0\}$ with H being the space of all available hypercontours. It has been proven in [9] that each hypercontour generates the corresponding classification function. This statement is true if the space X is metric. In E almost any type of information can be used assuming that we are able to implement this information in the computer oriented form. We can also decide if the classification is *supervised* or *unsupervised*. In the following we deal with the supervised one because of its more intuitive realization.

As mentioned above, there is certain freedom in defining energy function. In terms of image understanding paradigm, we aim at incorporation of expert knowledge into E and there are many ways to do it. Let for example:

$$E(h) = \sum_{j=1}^{J} \chi_j \left[1 - \delta\left(l_j^{cor}, \kappa(\mathbf{x}_j^{cor})\right)\right] \tag{4}$$

for each $h \in H$ where κ is a corresponding to hypercontour h classifier. Here $\left\{\left(\mathbf{x}_j^{cor}, l_j^{cor}\right)\right\}_{j=1,\ldots,J}$ denotes set of correctly labeled objects (*reference points*), χ_j is a real number introduced by expert and interpreted as significance of the j-th reference point for correct classification, and δ is the Kronecker symbol. Similar concepts of using expert knowledge were proposed in the unsupervised classification problem [8,9].

Another novel approach can be considered here, namely neural networks trained on examples and able to perform evaluation of classifiers or fuzzy expert systems able to use expert knowledge expressed in linguistic form.

The search of the optimal hypercontour may be driven in many ways, e.g. by the use of simulated annealing or genetic algorithm which perform global search and do not use gradient. In the present work, we applied the former method.

Adaptation is another interesting and powerful mechanism. Discrimination ability of given hypercontour is limited and it depends on the number of control points (assuming that other parameters are fixed). Flexibility of the potential active hypercontous (PAH) can be improved if we incorporate the change of the number of control points into the optimization procedure. For example, we can start with a small number of that points and add new ones if necessary. The rate

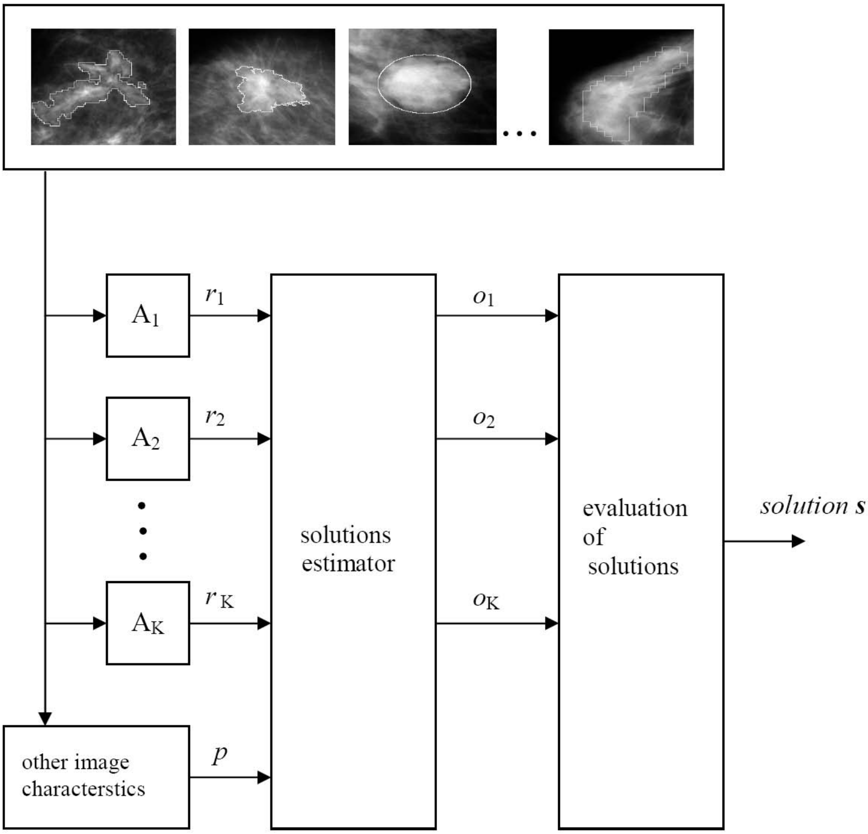


Fig. 2. Architecture of the image analysis system

of misclassification in some area of the space X can be the reason for introducing a few new control points.

3 System Architecture

In the literature (e.g. [11]), two types of implementation of distributed agent systems are distinguished. In the former type, the system is composed of several agents which perform their activities independently. They share the same general goal and can communicate for cooperation purposes. In the latter type, the total solution of the problem is the competition result between propositions of the parallel working agents.

The general goal stated here for the autonomous system (Fig. 2) is the recognition of objects (shapes) which may occur on the images collected in some database. In Fig. 2, $A_1, A_2, \ldots, A_K$ represent agents. The k-th agent is the adaptive active hypercontour method implemented for certain k-th shape. The images from the database are analyzed sequentially, i.e. only one image is analyzed by all A_k and the next image is read when the processing of the previous one has been finished.

Notation:

- E^k – formula of energy used in the agent A_k;
- ν^k – number of iteration within the k-th active contour algorithm ($\nu^k =$ 0,1,2,);
- $E^k_\nu = E^k_\nu(h_\nu)$ – value of energy in the ν-th iteration.

Taking certain form of energy function E^k we decide about the actual form of the contour which we are looking for when using the k-th agent A_k. Due to its activity, the active contour method generates a sequence of contours for each A_k. Each ν^k-th proposed contour h_ν is evaluated by the energy function of the form k. If the quality of current contour is not good enough (E^k_ν is too large), then the contour is modified in such a way that it should better fulfill the expectation closed in the form of the energy in other words, the next possibly better contour is created. The computations are performed for each agent A_k independently. Two criteria for finishing computations in A_k can be applied:

- number of iterations which is common for all K computational processes;
- separate minimization of each k-th energy function according to its specific criterion.

Results $\mathbf{r}_k$ are K contours and K corresponding final values of their corresponding energies. The solutions estimator uses the agents solutions and other available information in the form of vector $\mathbf{p}$ (statistical features, texture, results of discrete transforms applied to images, wavelet analysis, and the like) and determines outcomes $\mathbf{o}_k$ for each pair $(\mathbf{r}_k, \mathbf{p})$. The final solution $\mathbf{s}$ of the system is the decision about the kind of shape that has been recognized under additional conditions or demands $\mathbf{p}$ imposed on the desired image. If the additional demands are not considered, the system works as classifier ensembles performing selection, and it becomes a kind of the multiple classifier system [10].

Note that the presented system is similar to that proposed in [11] for an industrial marketing system. Due to this similarity and type of input/output signals, the solutions estimator can be implemented as fuzzy neural network while evaluation of solutions is a ranking procedure.

To explain the task of shape recognition more concrete, let us assume that all images are prepared in such a way that existing segments of lines have already been detected by the use of some low-level method Fig. 3. The goal is to find given shapes composed by those segments, for example circle, ellipse, triangle, rectangle, etc. For each considered k-th shape, the active contour method uses the relevant form of energy E^k. The form of each energy E^k for each particular shape should assure that its minima give compositions in a form which is similar to the reference concept of circle, ellipse, triangle, rectangle, and so on.

When considering the particular active hypercontours presented in Section 2, we state that the contour is determined by the set of control points $\mathbf{x}^c_i$ ($i =$

a) b)

Fig. 3. Detected shapes: closest to circle (a) and to square (b)

$1, 2, \ldots, N^c$). Looking for analogy in the neural networks theory, we can state that control points play the role of weights, while reference points x_j^{cor} ($j = 1, 2, \ldots, J$) create the learning (or training) set of known correctly labeled objects (l_j^{cor} denotes the corresponding label).

Moreover, it is important to mention that expert knowledge is incorporated in the formula of energy. Consequently, we state that in the active hypercontour approach final decision is based on three sources of information: expert knowledge (reflected in the formula of the energy), information reflected in, χ_j and data reference points.

4 Concluding Remarks

As shown in [5,6] respectively, active contours and active hypercontours are contextual classifiers. Since the form of energy determines the classifier (assuming that the form of the potential function has been fixed) then with each contour another shape can be associated. This allows to construct the proper multiagent system which uses many hypercontours simultaneously. The goal of the system is to decide what is the most possible content of the analyzed image.

It is also worthy to mention that adaptive active hypercontours can be used for tasks other than image segmentation, for example classification or clustering of textual documents.

Acknowledgment

The author wishes express his thanks to Mr Arkadiusz Tomczyk from the Institute of Computer Science of the Technical University of Lodz, Poland, who performed computations.

References

1. Wang, Y., Zhang, Y.-Q.: Fuzzy Web Information Classification Agents. In: Loia, V., Nikravesh, M., Zadeh, L.A. (eds.) Fuzzy Logic and the Internet, pp. 309–325. Springer, Heidelberg (2004)

2. Weiss, G.: Multi-Agent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge (1999)
3. Liu, J., Jin, X., Tsui, K.C.: Autonomy Oriented Computing. From Problem Solving to Complex Systems Modeling. Kluwer Academic Publ., Boston (2005)
4. Tadeusiewicz, R.: Szczepaniak: Basic Concepts of Knowledge-based Image Understanding. In: Nguyen, N.T. (ed.) KES-AMSTA 2008, Incheon, Korea. LNCS(LNAI), vol. 4953, pp. 41–51 (2008)
5. Tomczyk, A.: Active Hypercontours and Contextual Classification. In: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications - ISDA 2005, Wroclaw, Poland, pp. 256–261. IEEE Computer Society Press, Los Alamitos (2005)
6. Tomczyk, A., Szczepaniak, P.S.: On the Relationship between Active Contours and Contextual Classification. In: Kurzyński, M., et al. (eds.) Computer Recognition Systems. Proceedings of the 4th Int. Conference on Computer Recognition Systems CORES 2005, pp. 303–310. Springer, Heidelberg (2005)
7. Tomczyk, A., Szczepaniak, P.S.: Adaptive Potential Active Hypercontours. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 692–701. Springer, Heidelberg (2006)
8. Pedrycz, W., Loia, V.: P-FCM: A proximity-based fuzzy clustering. Fuzzy Sets and Systems 128, 21–41 (2004)
9. Tomczyk, A., Szczepaniak, P.S., Pryczek, M.: Active Contours as Knowledge Discovery Methods. In: Corruble, V., Takeda, M., Suzuki, E. (eds.) DS 2007. LNCS (LNAI), vol. 4755, Springer, Heidelberg (2007)
10. Kuncheva, L.I.: Combining Pattern Classifiers. John Wiley & Sons, Hoboken, New Jersey (2004)
11. Aliev, R.A., Aliev, R.R.: Soft Computing and Its Applications. World Scientific, Singapore (2001)

Bayesian Communication Leading to Nash Equilibrium Through Robust Messages - p-Belief System Case -

Takashi Matsuhisa*

Department of Natural Sciences, Ibaraki National College of Technology Nakane 866, Hitachinaka-shi, Ibaraki 312-8508, Japan
mathisa@ge.ibaraki-ct.ac.jp

Abstract. A communication model in the p-belief system is presented which leads to a Nash equilibrium of a strategic form game through robust messages. In the communication process each player predicts the other players' actions under his/her private information with conditional probability greater than p. The players communicate privately their conjectures through message according to the communication graph, where each recipient of the message learns and revises his/her conjecture. The emphasis is on that each player sends not exact information about his/her individual conjecture but robust information about the conjectures to an accuracy ε.

Keywords: Communication, , p-Belief system, Robust message, Nash equilibrium, Protocol, Conjecture, Non-corporative game.

AMS 2000 Mathematics Subject Classification: Primary 91A35, Secondary 03B45.

Journal of Economic Literature Classification: C62, C78.

1 Introduction

This article presents the communication model leading to a mixed strategy Nash equilibrium for a strategic form game as a learning process through robust messages in the p-belief system associated with a partition information structure. We show that

Main theorem. *Suppose that the players in a strategic form game have the p-belief system with a common prior distribution. In a communication process of the game according to a protocol with revisions of their beliefs about the other players' actions, the profile of their future predictions converges to a mixed strategy Nash equilibrium of the game in the long run.*

* Partially supported by the Grant-in-Aid for Scientific Research(C)(No.18540153) in the Japan Society for the Promotion of Sciences and by BUSAIKU-BUHI Foundation for Scientific Research.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 654–663, 2008.

Recently, researchers in economics, AI, and computer science become entertained lively concerns about relationships between knowledge and actions. At what point does an economic agent sufficiently know to stop gathering information and make decisions? There are also concerns about the complexity of computing knowledge. The most interest to us is the emphasis on the considering the situation involving the knowledge of a group of agents rather than of just a single agent.

In game theoretical situations, the concept of mixed strategy Nash equilibrium (J.F. Nash [13]) has become central. Yet a little is known about the process by which players learn if they do. This article will give a communication protocol run by the mutual learning leading to a mixed strategy Nash equilibrium of a strategic form game from the point of distributed knowledge system.

Let us consider the following protocol: The players start with the same prior distribution on a state-space. In addition they have private information given by a partition of the state space. Beliefs of players are posterior probabilities: A player *p-believes* (simply, *believes*) an event with $0 < p \leq 1$ if the posterior probability of the event given his/her information is at least p. Each player predicts the other players' actions as his/her belief of the actions. He/she communicates privately their beliefs about the other players' actions through messages through robust messages, which message is approximate information about his/her individual conjecture about the others' actions to an accuracy ε. The recipients update their belief according to the messages. Precisely, at every stage each player communicates privately not only his/her belief about the others' actions but also his/her rationality as messages according to a protocol,[1] and then the recipient updates their private information and revises her/his prediction. In addition, the players are assumed to be rational and maximising their expected utility according their beliefs at every stage. When a player communicates with another, the other players are not informed about the contents of the message.

The main theorem says that the players' predictions regarding the future beliefs converge in the long run, which lead to a mixed strategy Nash equilibrium of a game. The emphasis is on the three points: First that each player sends not exact information about his/her individual conjecture but robust information about the actions to an accuracy ε, secondly that each player's prediction is not required to be common-knowledge among all players, and finally that the communication graph is not assumed to be acyclic.

Many authors have studied the learning processes modeled by Bayesian updating. The papers by E. Kalai and E. Lehrer [5] and J. S. Jordan [4] (and references in therein) indicate increasing interest in the mutual learning processes in games that leads to equilibrium: Each player starts with initial erroneous belief regarding the actions of all the other players. They show the two strategies converges to an ε-mixed strategy Nash equilibrium of the repeated game.

As for as J.F. Nash's fundamental notion of strategic equilibrium is concerned, R.J. Aumann and A. Brandenburger [1] gives epistemic conditions for mixed

[1] When a player communicates with another, the other players are not informed about the contents of the message.

strategy Nash equilibrium: They show that the common-knowledge of the predictions of the players having the partition information (that is, equivalently, the **S5**-knowledge model) yields a Nash equilibrium of a game. However it is not clear just what learning process leads to the equilibrium.

To fill this gap from epistemic point of view, Matsuhisa ([6], [8], [9]) presents his communication system for a strategic game, which leads a mixed Nash equilibrium in several epistemic models. The articles [6], [8] [11] treats the communication model in the **S4**-knowledge model where each player communicates to other players by sending exact information about his/her conjecture on the others' action. In Matsuhisa and Strokan [11], the communication model in the p-belief system is introduced:[2] Each player sends exact information that he/she believes that the others play their actions with probability at least his/her conjecture as messages. Matsuhisa [9] extended the communication model to the p-belief system case that the sending messages are non-exact information that he/she believes that the others play their actions with probability at least his/her conjecture. This article is in the line of [9]; each player sends his/her robust information about the actions to an accuracy ε in the Bayesian communication model presented in Matsuhisa [9].

This paper organises as follows. Section 2 recalls the p-belief system associated with a partition information structure, and we extend a game on p-belief system. The Bayesian belief communication process for the game is introduced where the players send robust messages about their conjectures about the other players' action. In Section 3 we give the formal statement of the main theorem (Theorem 1) and sketch the proof. In Section 4 we conclude with remarks. The illustrated example will be shown in the lecture presentation at KES-AMSTA-08.

2 The Model

Let Ω be a non-empty *finite* set called a *state-space*, N a set of finitely many *players* $\{1, 2, \dots n\}$ at least two ($n \geq 2$), and let 2^Ω be the family of all subsets of Ω. Each member of 2^Ω is called an *event* and each element of Ω called a *state*. Let μ be a probability measure on Ω which is common for all players. For simplicity it is assumed that (Ω, μ) is a *finite* probability space with μ *full support*.[3]

2.1 *p*-Belief System[4]

Let p be a real number with $0 < p \leq 1$. The *p-belief system* associated with the partition information structure $(\Pi_i)_{i\in N}$ is the tuple

$$\langle N, \Omega, \mu, (\Pi_i)_{i\in N}, (B_i(*, p))_{i\in N} \rangle$$

[2] C.f.: Monderer and Samet [12] for the p-belief system.

[3] That is; $\mu(\omega) \neq 0$ for every $\omega \in \Omega$.

[4] Monderer and Samet [12].

consisting of the following structures and interpretations: (Ω, μ) is a finite probability space, and i's *p-belief operator* $B_i(*; p)$ is the operator on 2^{Ω} such that $B_i(E, p)$ is the set of states of Ω in which i *p-believes* that E has occurred with probability at least p ; that is,

$$B_i(E; p) := \{\omega \in \Omega \mid \mu(E \,|\, \Pi_i(\omega)) \geq p \}.$$

Remark 1. When $p = 1$ the 1-belief operator $B_i(*; 1)$ becomes knowledge operator in the **S5**-knowledge model.

2.2 Game on *p*-Belief System[5]

By a *game G* we mean a *finite* strategic form game

$$\langle N, (A_i)_{i \in N}, (g_i)_{i \in N} \rangle$$

with the following structure and interpretations: N is a finite set of players $\{1, 2, \dots, i, \dots n\}$ with $n \geq 2$, A_i is a finite set of i's *actions* (or i's pure strategies) and g_i is an i's *payoff function* of A into $\mathbb{R}$, where A denotes the product $A_1 \times A_2 \times \cdots \times A_n$, A_{-i} the product $A_1 \times A_2 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_n$. We denote by g the n-tuple $(g_1, g_2, \dots g_n)$ and by a_{-i} the $(n-1)$-tuple $(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$ for a of A. Furthermore we denote $a_{-I} = (a_i)_{i \in N \setminus I}$ for each $I \subset N$.

A probability distribution ϕ_i on A_{-i} is said to be i's *overall conjecture* (or simply i's *conjecture*). For each player j other than i, this induces the marginal distribution on j's actions; we call it i's *individual conjecture* about j (or simply i's conjecture *about* j.) Functions on Ω are viewed like random variables in the probability space (Ω, μ). If $\mathbf{x}$ is a such function and x is a value of it, we denote by $[\mathbf{x} = x]$ (or simply by $[x]$) the set $\{\omega \in \Omega | \, \mathbf{x}(\omega) = x\}$.

The information structure (Π_i) with a common prior μ yields the distribution on $A \times \Omega$ defined by $\mathbf{q}_i(a, \omega) = \mu([\mathbf{a} = a] | \Pi_i(\omega))$; and the i's overall conjecture defined by the marginal distribution

$$\mathbf{q}_i(a_{-i}, \omega) = \mu([\mathbf{a}_{-i} = a_{-i}] | \Pi_i(\omega))$$

which is viewed as a random variable of ϕ_i. We denote by $[\mathbf{q}_i = \phi_i]$ the intersection $\bigcap_{a_{-i} \in A_{-i}} [\mathbf{q}_i(a_{-i}) = \phi_i(a_{-i})]$ and denote by $[\phi]$ the intersection $\bigcap_{i \in N} [\mathbf{q}_i = \phi_i]$. Let $\mathbf{g}_i$ be a random variable of i's payoff function g_i and $\mathbf{a}_i$ a random variable of an i's action a_i. Where we assume that $\Pi_i(\omega) \subseteq [a_i] := [\mathbf{a}_i = a_i]$ for all $\omega \in [a_i]$ and for every a_i of A_i. i's action a_i is said to be *actual* at a state ω if $\omega \in [\mathbf{a}_i = a_i]$; and the profile a_I is said to be actually played at ω if $\omega \in [\mathbf{a}_I = a_I] := \bigcap_{i \in I} [\mathbf{a}_i = a_i]$ for $I \subset N$. The pay off functions $g = (g_1, g_2, \dots, g_n)$ is said to be *actually played* at a state ω if $\omega \in [\mathbf{g} = g] := \bigcap_{i \in N} [\mathbf{g}_i = g_i]$. Let **Exp** denote the expectation defined by

$$\mathbf{Exp}(g_i(b_i, \mathbf{a}_{-i}); \omega) := \sum_{a_{-i} \in A_{-i}} g_i(b_i, a_{-i}) \; \mathbf{q}_i(a_{-i}, \omega).$$

[5] Aumann and Brandenburger [1].

A player i is said to be *rational* at ω if each i's actual action a_i maximises the expectation of his actually played pay off function g_i at ω when the other players actions are distributed according to his conjecture $\mathbf{q}_i(\cdot\,;\omega)$. Formally, letting $g_i = \mathbf{g}_i(\omega)$ and $a_i = \mathbf{a}_i(\omega)$, $\mathbf{Exp}(g_i(a_i, \mathbf{a}_{-i});\omega) \geq \mathbf{Exp}(g_i(b_i, \mathbf{a}_{-i});\omega)$ for every b_i in A_i. Let R_i denote the set of all of the states at which i is rational.

2.3 Protocol[6]

We assume that the players communicate by sending *messages*. Let T be the time horizontal line $\{0, 1, 2, \cdots t, \cdots\}$. A *protocol* is a mapping $\Pr : T \to N \times N, t \mapsto (s(t), r(t))$ such that $s(t) \neq r(t)$. Here t stands for *time* and $s(t)$ and $r(t)$ are, respectively, the *sender* and the *receiver* of the communication which takes place at time t. We consider the protocol as the directed graph whose vertices are the set of all players N and such that there is an edge (or an arc) from i to j if and only if there are infinitely many t such that $s(t) = i$ and $r(t) = j$.

A protocol is said to be *fair* if the graph is strongly-connected; in words, every player in this protocol communicates directly or indirectly with every other player infinitely often. It is said to contain a *cycle* if there are players $i_1, i_2, \ldots, i_k$ with $k \geq 3$ such that for all $m < k$, i_m communicates directly with i_{m+1}, and such that i_k communicates directly with i_1. The communications is assumed to proceed in *rounds*[7]

2.4 Communication on p-Belief System

Let ε be a real number with $0 \leq \varepsilon < 1$. An *ε-robust Bayesian belief communication process* $\pi^{\varepsilon}(G)$ with revisions of players' conjectures $(\phi_i^t)_{(i,t)\in N\times T}$ according to a protocol for a game G is a tuple

$$\pi^{\varepsilon}(G) = \langle \Pr, (\Pi_i^t)_{i\in N}, (B_i^t)_{i\in N}, (\phi_i^t)_{(i,t)\in N\times T}\rangle$$

with the following structures: the players have a common prior μ on Ω, the protocol Pr among N, $\Pr(t) = (s(t), r(t))$, is fair and it satisfies the conditions that $r(t) = s(t+1)$ for every t and that the communications proceed in rounds. The revised information structure Π_i^t at time t is the mapping of Ω into 2^{Ω} for player i. If $i = s(t)$ is a sender at t, the message sent by i to $j = r(t)$ is M_i^t. An n-tuple $(\phi_i^t)_{i\in N}$ is a revision process of individual conjectures. These structures are inductively defined as follows:

– Set $\Pi_i^0(\omega) = \Pi_i(\omega)$.
– Assume that Π_i^t is defined. It yields the distribution

$$\mathbf{q}_i^t(a, \omega) = \mu([\mathbf{a} = a] | \Pi_i^t(\omega)).$$

Whence

[6] C.f.: Parikh and Krasucki [14].

[7] There exists a time m such that for all t, $\Pr(t) = \Pr(t+m)$. The *period* of the protocol is the minimal number of all m such that for every t, $\Pr(t+m) = \Pr(t)$.

- R_i^t denotes the set of all the state ω at which i is *rational* according to his conjecture $\mathbf{q}_i^t(\cdot\,;\omega)$; that is, each i's actual action a_i maximises the expectation of his pay off function g_i being actually played at ω when the other players actions are distributed according to his conjecture $\mathbf{q}_i^t(\cdot\,;\omega)$ at time t.[8]
- The message $M_i^t : \Omega \to 2^{\Omega}$ sent by the sender i at time t is defined as a robust information:

$$M_i^t(\omega) = \bigcap_{a_{-i}\in A_{-i}} \left\{\xi \in \Omega \ \mid \left|\mathbf{q}_i^t(a_{-i},\xi) - \mathbf{q}_i^t(a_{-i},\omega)\right| < \varepsilon \right\}.$$

Then:

- The revised partition Π_i^{t+1} at time $t+1$ is defined as follows:
  - $\Pi_i^{t+1}(\omega) = \Pi_i^t(\omega) \cap M_{s(t)}^t(\omega)$ if $i = r(t)$;
  - $\Pi_i^{t+1}(\omega) = \Pi_i^t(\omega)$ otherwise,
- The revision process $(\phi_i^t)_{(i,t)\in N\times T}$ of conjectures is inductively defined as follows:
  - Let $\omega_0 \in \Omega$, and set $\phi_{s(0)}^0(a_{-s(0)}) := \mathbf{q}_{s(0)}^0(a_{-s(0)},\omega_0)$
  - Take $\omega_1 \in M_{s(0)}^0(\omega_0) \cap B_{r(0)}([g_{s(0)}] \cap R_{s(0)}^0; p)$,[9] and set $\phi_{s(1)}^1(a_{-s(1)}) := \mathbf{q}_{s(1)}^1(a_{-s(1)},\omega_1)$
  - Take

$$\omega_{t+1} \in M_{s(t)}^t(\omega_t) \cap B_{r(t)}([g_{s(t)}] \cap R_{s(t)}^t; p),$$

and set

$$\phi_{s(t+1)}^{t+1}(a_{-s(t+1)}) := \mathbf{q}_i^{t+1}(a_{-s(t+1)},\omega_{t+1}).$$

The specification is that a sender $s(t)$ at time t informs the receiver $r(t)$ his/her individual conjecture about the other players' actions with a probability greater than his/her belief. The receiver revises her/his information structure under the information. She/he predicts the other players action at the state where the player p-believes that the sender $s(t)$ is rational, and she/he informs her/his the predictions to the other player $r(t+1)$.

We denote by ∞ a sufficient large τ such that $\mathbf{q}_i^{\tau}(\cdot\,;\omega_\tau) = \mathbf{q}_i^{\tau+1}(\cdot\,;\omega_{\tau+1}) = \mathbf{q}_i^{\tau+2}(\cdot\,;\omega_{\tau+2}) = \cdots$. Hence we can write $\mathbf{q}_i^{\tau}$ by $\mathbf{q}_i^{\infty}$ and ϕ_i^{τ} by ϕ_i^{∞}.

Remark 2. This communication model is a variation of the model introduced by Matsuhisa [6].

---

[8] Formally, letting $g_i = \mathbf{g}_i(\omega)$, $a_i = \mathbf{a}_i(\omega)$, the expectation at time t, $\mathbf{Exp}^t$, is defined by $\mathbf{Exp}^t(g_i(a_i,\mathbf{a}_{-i});\omega) := \sum_{a_{-i}\in A_{-i}} g_i(a_i,a_{-i})\ \mathbf{q}_i^t(a_{-i},\omega)$. An player i is said to be rational according to his conjecture $\mathbf{q}_i^t(\cdot\,,\omega)$ at ω if for all b_i in A_i, $\mathbf{Exp}^t(g_i(a_i,\mathbf{a}_{-i});\omega) \geq \mathbf{Exp}^t(g_i(b_i,\mathbf{a}_{-i});\omega)$.

[9] We denote $[g_i] := [\mathbf{g}_i = g_i]$.

3 The Result

We can now state the main theorem :

Theorem 1. *Suppose that the players in a strategic form game G have the knowledge structure with μ a common prior. In the ε-robust Bayesian belief communication process $\pi^{\varepsilon}(G)$ according to a protocol* Pr *among all players in the game, the n-tuple of their conjectures $(\phi_i^t)_{(i,t)\in N\times T}$ converges to a mixed strategy Nash equilibrium of the game in finitely many steps.*

The proof is based on the below proposition:

Proposition 1. *Notation and assumptions are the same in Theorem 1. For any players $i, j \in N$, their conjectures $\mathbf{q}_i^{\infty}$ and $\mathbf{q}_j^{\infty}$ on $A \times \Omega$ must coincide; that is, $\mathbf{q}_i^{\infty}(a;\omega) = \mathbf{q}_j^{\infty}(a;\omega)$ for every $a \in A$ and $\omega \in \Omega$.*

Proof. On noting that Pr is fair, it suffices to verify that $\mathbf{q}_i^{\infty}(a;\omega) = \mathbf{q}_j^{\infty}(a;\omega)$ for $(i,j) = (s(\infty), r(\infty))$. Since $\Pi_i(\omega) \subseteq [a_i]$ for all $\omega \in [a_i]$, we can observe that $\mathbf{q}_i^{\infty}(a_{-i};\omega) = \mathbf{q}_i^{\infty}(a;\omega)$, and we let define the partitions of Ω, $\{W_i^{\infty}(\omega) \mid \omega \in \Omega\}$ and $\{Q_j^{\infty}(\omega) \mid \omega \in \Omega\}$, as follows:

$$W_i^{\infty}(\omega) := \bigcap_{a_{-i}\in A_{-i}} [\mathbf{q}_i^{\infty}(a_{-i},*) = \mathbf{q}_i^{\infty}(a_{-i},\omega)] = \bigcap_{a\in A} [\mathbf{q}_i^{\infty}(a,*) = \mathbf{q}_i^{\infty}(a,\omega)],$$
$$Q_j^{\infty}(\omega) := \Pi_j^{\infty}(\omega) \cap W_i^{\infty}(\omega).$$

It follows that

$$Q_j^{\infty}(\xi) \subseteq W_i^{\infty}(\omega) \quad \text{for all } \xi \in W_i^{\infty}(\omega),$$

and hence $W_i^{\infty}(\omega)$ can be decomposed into a disjoint union of components $Q_j^{\infty}(\xi)$ for $\xi \in W_i^{\infty}(\omega)$;

$$W_i^{\infty}(\omega) = \bigcup_{k=1,2,\dots,m} Q_j^{\infty}(\xi_k) \quad \text{for} \quad \xi_k \in W_i^{\infty}(\omega).$$

It can be observed that

$$\mu([\mathbf{a} = a] | \, W_i^{\infty}(\omega)) = \sum_{k=1}^{m} \lambda_k \mu([\mathbf{a} = a] | \, Q_j^{\infty}(\xi_k)) \tag{1}$$

for some $\lambda_k > 0$ with $\sum_{k=1}^{m} \lambda_k = 1$.[10]

On noting that $W_j^{\infty}(\omega)$ is decomposed into a disjoint union of components $\Pi_j^{\infty}(\xi)$ for $\xi \in W_j^{\infty}(\omega)$, it can be observed that

$$\mathbf{q}_j^{\infty}(a;\omega) = \mu([\mathbf{a} = a] | \, W_j^{\infty}(\omega)) = \mu([\mathbf{a} = a] | \, \Pi_j^{\infty}(\xi_k)) \tag{2}$$

for any $\xi_k \in W_i^{\infty}(\omega)$. Furthermore we can verify that for every $\omega \in \Omega$,

[10] This property is called the *convexity* for the conditional probability $\mu(X|*)$ in Parikh and Krasucki [14].

$$\mu([\mathbf{a}=a]|\, W_j^{\infty}(\omega)) = \mu([\mathbf{a}=a]|\, Q_j^{\infty}(\omega)). \tag{3}$$

In fact, we first note that $W_j^{\infty}(\omega)$ can also be decomposed into a disjoint union of components $Q_j^{\infty}(\xi)$ for $\xi \in W_j^{\infty}(\omega)$. We shall show that for every $\xi \in W_j^{\infty}(\omega)$, $\mu([\mathbf{a}=a]|\, W_j^{\infty}(\omega)) = \mu([\mathbf{a}=a]|\, Q_j^{\infty}(\xi))$. For: Suppose not, the disjoint union G of all the components $Q_j(\xi)$ such that $\mu([\mathbf{a}=a]|\, W_j^{\infty}(\omega)) = \mu([\mathbf{a}=a]|\, Q_j^{\infty}(\xi))$ is a proper subset of $W_j^{\infty}(\omega)$. It can be shown that for some $\omega_0 \in W_j^{\infty}(\omega) \setminus G$ such that $Q_j(\omega_0) = W_j^{\infty}(\omega) \setminus G$. On noting that $\mu([\mathbf{a}=a]|G) = \mu([\mathbf{a}=a]|\, W_j^{\infty}(\omega))$ it follows immediately that $\mu([\mathbf{a}=a]|\, Q_j^{\infty}(\omega_0)) = \mu([\mathbf{a}=a]|\, W_j^{\infty}(\omega))$, in contradiction. Now suppose that for every $\omega_0 \in W_j^{\infty}(\omega) \setminus G$, $Q_j(\omega_0) \neq W_j^{\infty}(\omega) \setminus G$. The we can take an infinite sequence of states $\{\omega_k \in W_j^{\infty}(\omega) \mid k = 0, 1, 2, 3, \dots\}$ with $\omega_{k+1} \in W_j^{\infty}(\omega) \setminus (G \cup Q_j^{\infty}(\omega_0) \cup Q_j^{\infty}(\omega_1) \cup Q_j^{\infty}(\omega_2) \cup \cdots \cup Q_j^{\infty}(\omega_k))$ in contradiction also, because Ω is finite.

In viewing (1), (2) and (3) it follows that

$$\mathbf{q}_i^{\infty}(a;\omega) = \sum_{k=1}^{m} \lambda_k \mathbf{q}_j^{\infty}(a;\xi_k) \tag{4}$$

for some $\xi_k \in W_i^{\infty}(\omega)$. Let ξ_ω be the state in $\{\xi_k\}_{k=1}^m$ attains the maximal value of all $\mathbf{q}_j^{\infty}(a;\xi_k)$ for $k = 1, 2, 3, \cdots, m$, and let $\zeta_\omega \in \{\xi_k\}_{k=1}^m$ be the state that attains the minimal value. By (4) we obtain that $\mathbf{q}_j^{\infty}(a;\zeta_\omega) \leq \mathbf{q}_i^{\infty}(a;\omega) \leq \mathbf{q}_j^{\infty}(a;\xi_\omega)$ for $(i,j) = (s(\infty), t(\infty))$.

On continuing this process according to the *fair* protocol Pr, it can be plainly verified: For each $\omega \in \Omega$ and for any $t \geq 1$,

$$\mathbf{q}_i^{\infty}(a;\zeta'_\omega) \leq \cdots \leq \mathbf{q}_j^{\infty}(a;\zeta_\omega) \leq \mathbf{q}_i^{\infty}(a;\omega) \leq \mathbf{q}_j^{\infty}(a;\xi_\omega) \leq \cdots \leq \mathbf{q}_i^{\infty}(a;\xi'_\omega)$$

for some $\zeta'_\omega, \cdots, \zeta_\omega, \xi_\omega, \cdots \xi'_\omega \in \Omega$, and thus $\mathbf{q}_i^{\infty}(a;\omega) = \mathbf{q}_j^{\infty}(a;\omega)$ because $\mathbf{q}_j^{\infty}(a;\zeta_\omega) \leq \mathbf{q}_j^{\infty}(a;\omega) \leq \mathbf{q}_j^{\infty}(a;\xi_\omega)$ and $\mathbf{q}_i^{\infty}(a;\zeta) = \mathbf{q}_j^{\infty}(a;\xi)$ for every $\zeta, \xi \in \Omega$. in completing the proof.

Proof of Theorem 1: We denote by $\Gamma(i)$ the set of all the players who directly receive the message from i on N; i.e., $\Gamma(i) = \{\, j \in N \mid (i,j) = \ \mathrm{Pr}(t)$ for some $t \in T\}$. Let F_i denote $[\phi_i^{\infty}] := \bigcap_{a_{-i} \in A_i} [\mathbf{q}_i^{\infty}(a_{-i};*) = \phi_i^{\infty}(a_{-i})]$. It is noted that $F_i \cap F_j \neq \emptyset$ for each $i \in N$, $j \in \Gamma(i)$.

We observe the first point that for each $i \in N$, $j \in \Gamma(i)$ and for every $a \in A$, $\mu([\mathbf{a}_{-j} = a_{-j}] \,|\, F_i \cap F_j) = \phi_j^{\infty}(a_{-j})$. Then summing over a_{-i}, we can observe that $\mu([\mathbf{a}_i = a_i] \,|\, F_i \cap F_j) = \phi_j^{\infty}(a_i)$ for any $a \in A$. In view of Proposition 1 it can be observed that $\phi_j^{\infty}(a_i) = \phi_k^{\infty}(a_i)$ for each $j, k, \neq i$; i.e., $\phi_j^{\infty}(a_i)$ is independent of the choices of every $j \in N$ other than i. We set the probability distribution σ_i on A_i by $\sigma_i(a_i) := \phi_j^{\infty}(a_i)$, and set the profile $\sigma = (\sigma_i)$.

We observe the second point that for every $a \in \prod_{i \in N} \mathrm{Supp}(\sigma_i)$, $\phi_i^{\infty}(a_{-i}) = \sigma_1(a_1) \cdots \sigma_{i-1}(a_{i-1}) \sigma_{i+1}(a_{i+1}) \cdots \sigma_n(a_n)$: In fact, viewing the definition of σ_i we shall show that $\phi_i^{\infty}(a_{-i}) = \prod_{k \in N \setminus \{i\}} \phi_i^{\infty}(a_k)$. To verify this it suffices to show that for every $k = 1, 2, \cdots, n$, $\phi_i^{\infty}(a_{-i}) = \phi_i^{\infty}(a_{-I_k}) \prod_{k \in I_k \setminus \{i\}} \phi_i^{\infty}(a_k)$: We prove it by induction on k. For $k = 1$ the result is immediate. Suppose it is true for $k \geq 1$. On noting the protocol is fair, we can take the sequence of sets of players $\{I_k\}_{1 \leq k \leq n}$ with the following properties:

(a) $I_1 = \{i\} \subset I_2 \subset \cdots \subset I_k \subset I_{k+1} \subset \cdots \subset I_m = N$:

(b) For every $k \in N$ there is a player $i_{k+1} \in \bigcup_{j \in I_k} \Gamma(j)$ with $I_{k+1} \setminus I_k = \{i_{k+1}\}$.

We let take $j \in I_k$ such that $i_{k+1} \in \Gamma(j)$. Set $H_{i_{k+1}} := [\mathbf{a}_{i_{k+1}} = a_{i_{k+1}}] \cap F_j \cap F_{i_{k+1}}$. It can be verified that $\mu([\mathbf{a}_{-j-i_{k+1}} = a_{-j-i_{k+1}}] \,|\, H_{i_{k+1}}) = \phi^{\infty}_{-j-i_{k+1}}(a_{-j})$. Dividing $\mu(F_j \cap F_{i_{k+1}})$ yields that

$$\mu([\mathbf{a}_{-j} = a_{-j}] \,|\, F_j \cap F_{i_{k+1}}) = \phi^{\infty}_{i_{k+1}}(a_{-j})\mu([\mathbf{a}_{i_{k+1}} = a_{i_{k+1}}] \,|\, F_j \cap F_{i_{k+1}}).$$

Thus $\phi^{\infty}_j(a_{-j}) = \phi^{\infty}_{i_{k+1}}(a_{-j-i_{k+1}})\phi^t_j(a_{i_{k+1}})$; then summing over a_{I_k} we obtain $\phi^{\infty}_j(a_{-I_k}) = \phi^{\infty}_{i_{k+1}}(a_{-I_k-i_{k+1}})\phi^{\infty}_j(a_{i_{k+1}})$. It immediately follows from Proposition 1 that $\phi^{\infty}_i(a_{-I_k}) = \phi^{\infty}_i(a_{-I_k-i_{k+1}})\phi^{\infty}_i(a_{i_{k+1}})$, as required.

Furthermore we can observe that all the other players i than j agree on the same conjecture $\sigma_j(a_j) = \phi^{\infty}_i(a_j)$ about j. Noting that i is rational at ω_∞ we conclude that each action a_i appearing with positive probability in σ_i maximises g_i against the product of the distributions σ_l with $l \neq i$. This implies that the profile $\sigma = (\sigma_i)_{i \in N}$ is a mixed strategy Nash equilibrium of G, in completing the proof. □

4 Concluding Remarks

We have observed that in a communication process associated with p-belief system, their predictions induces a mixed strategy Nash equilibrium of the game in the long run. It is well to end some remarks on related literatures.

The **S5**-knowledge model is an operator model equivalent to the Kripke semantics for the modal logic **S5** (= KT45), which is the binary relation on a state-space satisfying reflectivity, transitivity and symmetry. The **S4**-knowledge model is equivalent to the Kripke semantics for the modal logic **S4** (= KT4), which semantics is the binary relation on a state-space satisfying the two relations: reflectivity, transitivity. The p-belief system is an extension of the **S5**-knowledge model,[11] but it is independent of the **S4**-knowledge model: they are both independent and weak models of the **S5**-knowledge model.

Matsuhisa [6] and [8] introduced the communication model for **S4**-knowledge model, and he established that the communication model leads to a Nash equilibrium. Furthermore Matsuhisa [7] showed a similar result for ε-mixed strategy Nash equilibrium of a strategic form game in the **S4**-knowledge model, which gives an epistemic aspect in Theorem of E. Kalai and E. Lehrer [5]. These papers treats communication models by sending the exact information on the players conjectures about the other players' actions through messages.

This article highlights a communication among the players in a game by sending an approximate information about the ations through ε-robust massages, and shows that the convergence to an exact Nash equilibrium is guaranteed even in such communication after long run. The main theorem in this article is valid in the ε-robust communication for the **S5**-knowledge model. There is an agenda

[11] See Remark 1.

to further research; first, to extend our main theorem to **S4**-knowledge model, which gives another generalisation of the theorem for the **S5**-knowledge model, because it coincides with the theorems in Matsuhisa [6] and [8] when $\varepsilon = 0$, and secondly, to unify all the communication models in the preceding papers ([6], [8], [11], [9], [10]) including the result presented in this article.

References

1. Aumann, R.J., Brandenburger, A.: Epistemic conditions for mixed strategy Nash equilibrium. Econometrica 63, 1161–1180 (1995)
2. Bacharach, M.: Some extensions of a claim of Aumann in an axiomatic model of knowledge. Journal of Economic Theory 37, 167–190 (1985)
3. Binmore, K.: Fun and Games. D. C. Heath and Company, Lexington, Massachusetts USA, pp. xxx+642 (1992)
4. Jordan, J.S.: Bayesian learning in normal form games. Games and Economic Behaviour 3, 60–81 (1991)
5. Kalai, E., Lehrer, E.: Rational learning to mixed strategy Nash equilibrium. Econometrica 61, 1019–1045 (1993)
6. Matsuhisa, T.: Communication leading to mixed strategy Nash equilibrium I. In: Maruyama, T. (ed.) Mathematical Economics, Suri-Kaiseki-Kenkyusyo Kokyuroku, vol. 1165, pp. 245–256 (2000)
7. Matsuhisa, T.: Communication leading to epsilon-mixed strategy Nash equilibrium, Working paper (2001). The extended abstract was presented in the XIV Italian Meeting of Game Theory and Applications (IMGTA XIV), July 11-14 (2001)
8. Matsuhisa, T.: Communication leading to a Nash equilibrium without acyclic condition (S4-knowledge case), M. In: Bubak, M., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2004. LNCS, vol. 3039, pp. 884–891. Springer, Heidelberg (2004)
9. Matsuhisa, T.: Bayesian communication under rough sets information. In: Butz, C.J., et al. (eds.) 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-ITA 2006 Workshop Proceedings, pp. 378–381 (2006)
10. Matsuhisa, T.: Communication leading to Nash equilibrium through robust messages - S5 knowledge model case. In: Dress, A.W.M., Xu, Y., Zhu, B. (eds.) COCOA. LNCS, vol. 4616, pp. 136–145. Springer, Heidelberg (2007)
11. Matsuhisa, T., Strokan, P.: Bayesian belief communication leading to a Nash equilibrium in belief. In: Deng, X., Ye, Y. (eds.) WINE 2005. LNCS, vol. 3828, pp. 299–306. Springer, Heidelberg (2005)
12. Monderer, D., Samet, D.: Approximating common knowledge with common beliefs. Games and Economic Behaviours 1, 170–190 (1989)
13. Nash, J.F.: Equilibrium points in n-person games. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 36, pp. 48–49 (1950)
14. Parikh, R., Krasucki, P.: Communication, consensus, and knowledge. Journal of Economic Theory 52, 178–189 (1990)

Manager-Agent Framework of a Medical Device Communication for u-Healthcare Services with USN

Yung Bok Kim[1] and Jae-Jo Lee[2]

[1] Sejong University
KunJa-Dong, Kwang-Jin-Ku, Seoul, Korea 143-747
yungbkim@sejong.ac.kr

[2] Korea Electrotechnology Research Institute (KERI)
Gyeonggi-Do, Korea 426-170
jjlee@keri.re.kr

Abstract. A medical device communication framework for a ubiquitous healthcare service based on ISO11073 with USN (ubiquitous sensor network) was studied in view of a ubiquitous computing and networking environment. We introduce the USN for e-healthcare service in smart environments. Beyond e-healthcare service, as an application of u (ubiquitous)-healthcare service for USN including PLC technology, we implemented a real-time health-monitoring service with a cost-effective Web server. We introduce the u-healthcare service based on the framework ISO11073/IEEE1073 (i.e. X73) with USN including PLC technology for a sink node or gateway without power limitations.

1 Introduction

In the sensor network, especially in wireless sensor networks (WSN), the research has mainly focused on routing, aggregation, and energy efficient data management algorithms inside one sensor network. The deployment, application development, and standardization aspects are not usually addressed [1]. We focused on the study of an application service, i.e. a u-Healthcare service, among various services based on sensor networks composed of wired sensors or wireless sensors in the USN. We defined the USN as a sensor network composed of ubiquitously prevalent wired-sensors as well as ubiquitously prevalent wireless- sensors, which are connected in the network. Our main interest is to find a killer application based on manager-agent application using PLC (Power Line Communication) for a wired sink node or gateway, among various communication technologies related to sensor networks. Beyond the e-healthcare service, the ubiquitous healthcare (u-healthcare) service will be more evolved in the emerging ubiquitous computing and networking society.

Some background in researches related to smart environments and e-healthcare is necessary because wired-sensor networks are essential in the smart environment. 'Smart environments for all' was introduced in a special thematic session by

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 664–673, 2008.

Nussbaum [2], and he stressed that smart environments would have significant potential to increase the quality as well as the efficiency of healthcare. Development of smart home technologies dedicated to people with disabilities provides a challenge in determining accurate requirements and needs in dynamic situations; and Feki et al. [3] introduced the integration of context awareness and multimodal functionalities in the smart environment. Information such as the availability of resources, user profiles, location, input controls and services can be used to improve the interaction between users and their environment.

In the emerging USN, sensor devices could be greatly miniaturized, harvest energy from the environment and communicate with other networks and devices integrated in our homes or our cities as smart environments. Recently, the new application of health monitoring and medical care based on USN is gaining popularity among researchers and offers genuine promise for future practical uses. We will introduce a manager-agent framework of medical device communication for u-healthcare services with USN in smart environments.

The following sections are organized as follows. We will discuss the background of the medical device communication framework and the USN for the ubiquitous healthcare service in the evolving computing and networking environments. We will present the X73 framework for healthcare service. Beyond the e-healthcare service, as a primitive application for ubiquitous healthcare services, we introduce our application of the real-time health monitoring service using health-monitoring sensors and health information web server in the mobile Internet before the proliferation of ubiquitous computing and the networking environment. For quality of service (QoS), we introduce an evaluation scheme for u-healthcare services in smart environments for point-of-care with PLC technology in a sink node or gateway. Finally we conclude our study with considerations for further research.

2 Backgrounds

The need for health monitoring in the e-healthcare service is gradually increasing and its application is becoming feasible with sensor network technology. By adopting tiny wireless sensor network devices in the ubiquitous networking environment with some sort of specific health monitoring system, regular patients and elderly people can be observed in smart environments that provide an e-healthcare service. For this purpose, wearable vital sign sensors can be attached to the body, allowing continuous communication transferring the sensed physical status of the wearer. Health monitoring with ubiquitous sensor networks (USN) in the ubiquitous networking environment has some outstanding features compared to traditional medical healthcare systems.

Health monitoring based on USN in the ubiquitous networking environment provides a totally different healthcare system scenario. Wireless sensor nodes are initially small and generally use batteries rather than power cables, thereby eliminating power limitation problems. The general features of the sensor network such as tiny sensor nodes, network construction and self-configuration allow it to potentially be used in medical care monitoring applications, even with power limitation problem.

With the USN, we need to consider the advantages, cost-effectiveness, and security issues of both (i.e. wired and wireless) sensor nodes/networks in the ubiquitous healthcare service, instead of insisting on the use of wireless sensor/networks. We should consider the wired sensor/network based on PLC technologies, especially for the following point-of-care medical service with USN.

The standard for point-of-care connectivity [4] establishes a set of specifications that allow seamless multi-vendor interoperability and communication between point-of-care devices, data concentrators, and clinical information systems. A document on point-of-care connectivity has been developed by the CLSI (Clinical and Laboratory Standards Institute) Subcommittee on point-of-care connectivity. The core of the standard is a group of three specifications developed by the Connectivity Industry Consortium (CIC). The specifications describe the attributes of the access point; communication protocols between the device and the access point; and communications between the data manager and clinical information systems. The collaborative effort among providers and manufacturers has produced a set of specifications acceptable to both.

The most influential standards in the vendor community derive from bodies which have achieved international influence and authority beyond the formal International/European standards bodies, e.g. the U.S. based HL7 for messaging, DICOM for imaging, IEEE for medical device communications [5]. We introduce these briefly as follows.

HL7, Health Level Seven, a reference to the 7th layer of the OSI model, was founded in 1987 by several vendors of software for the health care industry [6], and funded by American manufacturers of medical equipment and accredited by the American National Standards Institute (ANSI). It is a standard for the exchange of medical messages. Their goal was to develop consensual message formats to facilitate better interoperability of Hospital Information Systems (HIS). It develops its own syntax, in the seven levels of the protocol stack, for representing the information in a simple structure composed of segments and field labels (each one identified by its data type). Like DICOM, it exchanges the results of observations related to vital signs and biomedical signals, but it is not applicable to the interconnection of devices.

DICOM (Digital Imaging Communication) is a standards organization creating, and maintaining standards for the communication of biomedical diagnostic and therapeutic information in disciplines using digital images and associated data. It was formed by the American College of Radiologists (ACR) and the National Electrical Manufacturers Association (NEMA). It occupies a privileged position in medical imaging since it is very widespread among the healthcare community and manufacturers. It includes some directives for the exchange of biomedical signals, particularly ECG, but it is not applicable to the interconnection of monitoring devices.

ISO11073/IEEE1073 (known as X73), Standard for Medical Device Communications: a family of documents that defines the entire seven layer communication requirements for the 'Medical Information Bus' (MIB). This is a robust, reliable communication service designed for intensive care units, operating rooms, and emergency room bedside devices. The goal of X73 is to improve the interoperability and plug-and-play capacities of the different medical devices and medical information systems. Peripheral Area Network Interface (PAN-IF) connects an application hosting device, such as a personal computer, cell phone, or monitoring hub, to a PAN device, which is either a sensor or an actuator. The PAN-IF upper layers are implemented using the ISO/IEEE 11073-20601 Optimized Exchange Protocol, which leverages work from the ISO/IEEE 11073 Medical Device Communications working group. [7]

3 Framework for u-Healthcare

The following diagram [8] based on a manager-agent framework of medical device communication shows the upper layer communication stack in X73, or layered set of protocol and service components. Manager/agent (client/server) framework for communication systems is defined in the standard. Application processes in both the manager and the agent provide the (user) functionality of the device. For example, the agent has at least one application process that is able to process signals from an invasive blood pressure transducer; the manager application process is responsible for collecting and archiving vital signs information that is provided by the agent. The agent provides vital signs information in form of objects that are defined in the Domain Information Model. The Medical Data Information Base (MDIB) contains the set of object instances that is provided by the agent.

ACSE (Association Control Service Element) is the ISO/OSI standard for association control. CMDISE (Common Medical Device Information Service Element) is an object management service in principle, a lightweight version of the ISO/OSI CMISE (Common Management Information Service Element). The Remote Operation Service Element, ROSE, provides basic services used by the CMDISE (invoke an Operation, return the result of an operation, return an error, reject an operation). To comply with the definition of optimized encoding rules, a modified version of the ROSE is needed to work with the CMDISE.

ACSE means Association Control Service Element, MDIB means Medical Data Information Base, and MDSE means Medical Device Service Element, in Fig.1. The Session Layer and Presentation Layer produce a minimized overhead only.

The ISO11073/IEEE1073 (known as X73) set of standards for Point of Care Medical Device Communication based on a manager-agent framework of medicla device communication is the best-positioned international standard to provide interoperability with different sensors [6]. The X73 standard, currently in the development phase, is a single set of standards for complete connectivity among medical devices that contribute plug-and-play, transparency, and ease of use and configuration. X73 extends to the seven levels of the Open System Interface (OSI) protocol stack and absorbs EN 13734 (VITAL) for the upper layers, EN 13735 (INTERMED) for the intermediate layers, and the older 1073 standards (1073.3 and 1073.4) for the lower layers. Thus, it gives a full solution from the cable itself and connector to abstract representation of information and services. X73 distinguishes

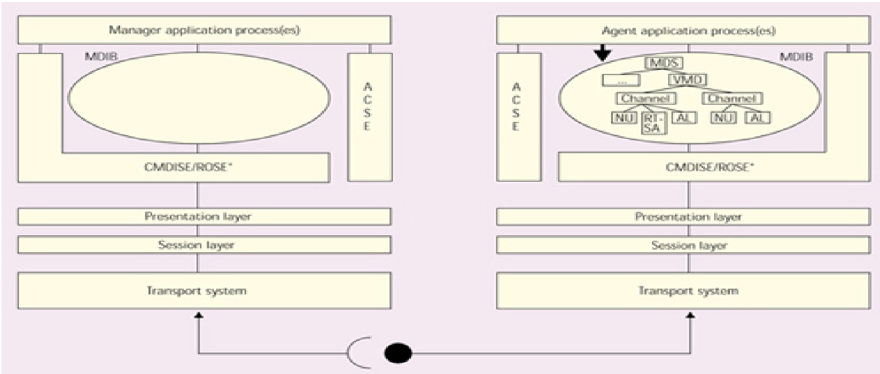


Fig. 1. Logical interface between two MIB-connected systems (Manager/Agent). (from the CEN 13734 standard for Vital Signs Information Representation).

four main standards groups: transport (e.g., wireless or cabled), services of general applications (e.g., for events or polling), device data (e.g., object-oriented model and representation terminology), and network communication standards (e.g., gateway between the representation of data and IEEE1073, DICOM, or HL7 messages). Toledo et al. [6] found a problem with the lack of X73 standardization of wireless connectivity. They plan to fix it when required by using the lower layers of the wireless technology (Bluetooth, Zigbee, etc), then merging them as possible with the upper layers, which actually follow the standard, waiting for the X73 to get updated to these technologies.

In the new and exciting field of Ubiquitous Sensor Networks (USN), e-healthcare services for medical care applications appear as a promising research area with a wide range of possibilities. Beyond e-healthcare services, we will introduce ubiquitous healthcare service and introduce the health monitoring sensor network based on mobile Internet [9] as a primitive application for the ubiquitous healthcare service.

The following use cases [10] reflect diversity, and there is throughout an assumption of use of existing or supplemented 11073 (X73) standard provisions.

UseCase#1: Mobile Wellness Monitoring of a Single Cardiac Patient;
UseCase#2: Wellness Weight Monitoring of a Single Patient Home;
UseCase#3: Chronic Respiratory Patient Management;
UseCase#4: Elderly patient follow-up;
UseCase#5: Cardiac Parameter Monitoring of a Single Patient Home;
UseCase#6: Secure Registration of Tele-monitoring Device;
UseCase#7: Body Area Network Monitoring;
UseCase#8: Critical cardiac alarms.

The UseCase#4 was considered for the elderly patient who lives alone or spends most of the day alone, and this case may be appropriate for the application of PLC based on a wired sensor/network. The monitoring service controls his vital signs once a day and generates a summary of his/her level of activity. Vital signs are recorded using a pressure cuff (blood pressure and heart rate) and a digital thermometer. A wearable 3D accelerometer records patient movements during all the day and generates parameters that summarize patient activity (% time sitting, % time standing, % time laying, % time sleeping) All this information is sent to a gateway (embedded PC located at the home).

For u-healthcare with the USN, we should consider both the wired sensor/network and the wireless sensor/network. Considering the WSN, we should refer to the aforementioned UseCase#1, #3 and #7 based on X73.

We studied the real-time health-monitoring applications, frequently acquiring personal information as well as writing analyzed information as a response to requests from clients or health-monitoring agents in the mobile Internet environment.

The service process for the u-healthcare system using a portable device is as follows: Measurement of biological signal/ analysis and display of the signal/ data transfer and recording/ diagnosis of transferred data. Looking at each phase of these four-step process, the measurement phase requires a composite function, network, cost effectiveness, lightness, unconfined and continuous measurements in order to acquire a great deal of various health indices. The monitoring process requires module information from a PDA (Portable Digital Assistant), laptop PC, cellular phone or other devices, and then the measured data shall be analyzed and transferred to the applicable medical organization. When the data is transferred through a wireless network environment, the transferred data shall have an interface with hospitals and/or other medical organization in order to prove the reliability of the analyzed data.

Further research, needs to consider the previous UseCase#7 based on X73. For wireless communication to transmit personal health information, the design includes the 802.15.4 standard for radio communication, used in the Zigbee specification, which offers promising reductions in power consumption. Sensors such as ECG, heartbeat, light, accelerometer and magnetometer are integrated into the Sensor board. All these sensors can be operated individually or in collaboration with other sensors. If we want to use external health sensor or transmit collected health information to local computer systems connected to the Internet, we can use the Interface board. It supports Serial/Parallel communication to communicate with external sensors or computers connected to the Internet; we study the PLC adaptor for the sensor node or gateway.

4 Experience for QoS

For the real-time health-monitoring network using mobile Internet in the USN environment, the dominating factor and the standard deviation of that random variable should be bounded within the deterministic response time. To be deterministic for real-time application, the estimation time should be bounded within deterministic time, and the interchange of data between the sensor nodes in watch phone and the server should be automatic except the information requested by the user; therefore the Web server should be efficient and if possible have high performance for the dedicated application. Further, the exchanged data and analyzed information should be as simple as possible with a simplified and efficient format. If possible, the bandwidth requirement for wireless or mobile Internet should be immune to network traffic conditions; also that would be ideal in case of degradation caused by other rich multimedia contents sharing the same network and server. The spent time in the Web server may be considered immune to the network and server condition. This system is based on wired or mobile Internet, and the monitored health data from sensors in a watch phone can be registered at any time using mobile Internet in the USN environment.

For the consistency of health-monitoring information and for a convenient user interface, we need a unified health-monitoring Web server for wired and mobile Internet. We need to consider the health-information center accessibility to the doctor or nurse as well as to the disabled and elderly within the concept of the X73 framework. We used a single Web server as a health information Web server for cost-effectiveness and the simplicity of management. This method offers effectiveness and efficiency for the real-time health-monitoring network and utilization of resources, e.g. the bandwidth for communication and the disk storage size for health information Web server for the patient. We also consider a reliability scheme.

We assume that the wrist phone with health-monitoring sensors wrote the health information regularly, which should be chosen carefully in further research with the X73 framework, to the health information Web server through the mobile Internet, and processes the data and analyzes for the request of the results in real-time way. Depending upon the frequency of writing the health information, the workload of the health Web server may change, and the interval of regular writing may be considered the arrival rate in the queuing performance analysis model for health-monitoring. The transmission packet unit for billing by mobile communication service provider is about 0.5 cents (U.S.); the packet size is 512 Bytes, which is the minimum packet size for billing in Korea. Therefore, if possible, the health-monitoring data for the wrist phone to the server through mobile Internet should be below 512 Bytes, and this size is also bounded much below 1.5Kbytes that is one emulated WML deck for performance evaluation of the health-monitoring Web server discussed later. As a reference, one SMS (Short Message Service) data, i.e. 80 Bytes, costs around 2.5 cents (U.S.) in Korea. Therefore, clearly one WML packet is less expensive than one SMS message in terms of the cost for wireless communication as a scheme in the USN environment.

The health Web server should have the capability to show the appropriate health contents, i.e. the HTML contents for the wired Internet as well as the mobile contents for many different kinds of mobile devices, e.g. WML, mHTML, etc. For the unified service, there are several constraints compared to the contents for the wired Internet. First of all, we should consider the various kinds of mobile devices as well as browsers for the mobile Internet, and each of those devices may have different capabilities in terms of images. We considered only text-based health-monitoring information from the wrist phone to the health-monitoring Web server and vice versa to be immune to any type of Internet traffic load as well as to minimize the mobile communication cost for cost-effective health-monitoring services based on X73 framework with PLC technology.

Important issues such as data quality of sensor nodes, real-time aggregation and estimation of random variables for comprehensive accessibility in relation to the u-Healthcare service should be considered. We are looking into implementing health-monitoring sensors in the wrist phone for the disabled and elderly as a primitive example with single-hop sensor nodes; that single-hop sensor node even with PLC sink node/gateway is a starting point to gain insights about important issues in terms of data analysis with random variables.

The wrist phone as a sensor node is a future product not yet implemented; thus we assume that the sensors for measuring the pulses, the strength of the pulses, and blood pressure, etc. would give the raw data about health of the disabled and elderly.

We studied real-time health-monitoring applications, frequently getting personal information. We further wrote analyzed information as a response to the request by clients or health-monitoring agents in the USN environment. We will discuss various aspects for the QoS of a ubiquitous application service.

For the real-time health-monitoring network using the mobile Internet in the USN environment, the dominating factor and the standard deviation of that random variable should be bounded within the deterministic response time. To be deterministic for real-time application, the estimation time should be bounded within deterministic time, and the interchange of data between the watch phone and the server should be automatic except for information requested by the user. Therefore the Web server should be efficient and have high performance for the dedicated application if possible, and the exchanged data and analyzed information should be as simple as possible with a simplified and efficient format. If possible, the bandwidth requirement for the wireless or mobile Internet in the USN environment should be immune to network traffic conditions. That would be ideal in terms of the degradation caused by the other rich multimedia contents sharing the same network and server.

This system is based on the wired or mobile Internet, and the monitored health data from sensors in a watch phone can be registered at any time using the mobile Internet with the domain name of test Web server for wired/mobile Internet in the USN environment. Actually this site has been used as a real-time information network server, and we considered it as a health-monitoring server because it can be used as any Web server for testing.

We studied the important QoS performance metric, delay, from the user's perspectives for the disabled, considering the application of pervasive (or ubiquitous) computing technologies in healthcare. The preparation time for the disabled to get a service (e.g. medical device, mobile device, etc.) is U (ubiquitous environment time metric); the time spent by the disabled with medical devices to make an appropriate action for the service is D (device time metric); the aggregate time to the medical healthcare server after the medical device for medical service is S (service time metric for communication and Web server); the time depending upon medical contents is C (contents time metric). If iteration of each variables is required, then subscript i, j, k, l, m can be used as U_i, D_j, S_k, C_l, P_m. Depending upon ubiquitous environment, device, service and patient, the iteration i, j, k, l, m may be different.

Among the above random variables, i.e. the performance metrics, (*U, D, S, C, P*) for the disabled or patient, the dominating factor, i.e. the random variable, may be different depending upon u-healthcare environment, device, service, contents and patient. We can represent the statistical values (i.e. mean, deviation, etc.) of random variables as performance metrics for QoS (quality of service) of real-time medical service, and the quality of u-healthcare service can be compared with the QoS performance metrics as comprehensive accessibility. Each performance metric can be used to enhance the QoS as follows. The ubiquitous environment time metric *U* will be shortened depending upon the proliferation of the smart environments including the PLC for u-healthcare. Device time metric *D,* i.e. the handling speed of the medical device, is one of the most important performance factors in any medical services for the disabled or patient. Service time metric *S* can be shortened according to the communication facility and Web server with the DB in the e-hospital. Contents time metric *C* can be shortened considering accessibility in contents design for the u-healthcare service. Patient time metric *P* may be different from person to person depending upon the disability of the person.

We can order the dominating factors in the overall performance at the user's perspective. The user interface design for the disabled with a medical device in a ubiquitous healthcare environment with PLC technology (i.e. *U* becomes smaller as the ubiquity increases) is important to decrease the device time metric *D* that is heavily related to the UI (user interface) convenience of medical devices for the disabled and handicapped users. With greater proliferation of the ubiquitous network environment including the PLC without power limitation, the time spent with ubiquitous devices with sensor nodes will become more critical. We should consider this QoS concept in the implementation of the X73 framework for the u-healthcare service with USN including PLC technology.

5 Conclusions

A manager-agent framework of medical communication for ubiquitous healthcare service based on X73 with USN was introduced. We focused on manager-agent framework for a u-Healthcare service based on wired sink nodes/gateways including PLC technology beyond WSN, and we showed comprehensive accessibility for QoS in the USN environments.

References

1. Aberer, K., Hauswirth, M., Salehi, A.: A Middleware For Fast And Flexible Sensor Network Deployment. In: VLDB 2006, Seoul, Korea, ACM, New York (2006)
2. Nussbaum, G.: Smart Environments for All Introduction to the Special Thematic Session. In: Miesenberger, K., Klaus, J., Zagler, W., Burger, D. (eds.) ICCHP 2004. LNCS, vol. 3118, pp. 889–890. Springer, Heidelberg (2004)
3. Feki, M.A., Renouard, S., Abdulrazak, B., Chollet, G., Mokhtari, M.: Coupling Context Awareness and Multimodality in Smart Homes Concept. In: Miesenberger, K., Klaus, J., Zagler, W., Burger, D. (eds.) ICCHP 2004. LNCS, vol. 3118, pp. 906–913. Springer, Heidelberg (2004)

4. POCT1-A2, vol. 26(28), Replaces POCT1-A, vol. 21(24) Point-of-Care Connectivity; Approved, Standard—2 edn. Clinical and Laboratory Standards Institute (CLSI) (2006)
5. IEEE1073 framework Draft V3.1. Report from the CEN/ISSS eHealth Standardization Focus Group. Current and future standardization issues in the e-Health domain: Achieving interoperability. eHealth Standardization Focus Group. July 23 (2004)
6. Toledo, P., Galarraga, M., Martinez, I., Serrano, L., Fernandez, J., Pozo,: Towards e-Health Device Interoperability: The Spanish Experience in the Telemedicine Research Network. In: IEEE 28th EMBS Conference, New York City, USA, August 30-September 3 (2006)
7. Acrroll, R., Cnossen, R., Schnell, M., Simons, D.: Continua: An Interoperable Personal Healthcare Ecosystem. IEEE Pervasive Computing, 90–94 (October-December, 2007)
8. ISO/IEEE P11073-20101/d06f, IEEE 1073.2.1.1TM/d06f, Date: 2002 September, Draft Standard for Health informatics – Point-of-care medical device communication - Application profiles – Base standard
9. Kim, Y.B., Kim, J.Y.: Real-Time Health-Monitoring Network for Disabled and Elderly People. In: Miesenberger, K., Klaus, J., Zagler, W., Burger, D. (eds.) ICCHP 2004. LNCS, vol. 3118, pp. 377–384. Springer, Heidelberg (2004)
10. CEN 2006 WG Working document: Use case examples, CEN TC 251/WGIV N 06 xxx. Health informatics: Point-of-care medical device communication -Use cases for mobile, ambulant and home care, July 18 (2006)

An Intelligent Multi Agent Design in Healthcare Management System

Dharmendra Sharma and Fariba Shadabi

School of Information Sciences and Engineering
University of Canberra
ACT, 2601, Australia

Abstract. Over the past few years, multi-agent approach has been successfully used in the development of large complex systems. Therefore, multi-agent approach can be considered as an effective approach for the development of complex e-health systems. The purpose of this study is to explore the use of intelligent multi-agent approach for developing e-health systems for the prediction of kidney transplant outcomes and the management of chronic diseases such as diabetes. The proposed kidney transplant outcome prediction is based on the use of a novel classification approach which is a combination of initial data preparations, preliminary classification by ensembles of neural networks, generation of new training data based on criteria of highly accuracy and model agreement, and decision trees.

Keywords: Multi-agent, e-health, kidney transplant, data mining.

1 Introduction

The economic and social benefits of accurately predicting medical outcomes are very high. As a result, the problem of improving predictive models has attracted many researchers. Over the past few years there has been great interest in the use of data mining tools across the healthcare spectrum.

Clinical data are known to be complex due to genetic and biological diversity of individuals, disease marker variability amongst individuals, the variability in the combination of drugs given to each individual and missing data, to name a few. Until now, most clinical data analysis has largely relied on the use of standard traditional strategies such as logistic regression [1] for knowledge discovery or predicting medical outcomes. Statistical techniques have been used successfully in a number of medical domains [2]. However, they do not always have the capacity for solving problems of high complexity. Consequently, researchers turned their attention to the search for alternative data modeling and prediction techniques that would result in better performance, simpler implementation and adoption in clinical practices.

Many researchers have recently turned their attention to Intelligent Data Mining (DM) and Machine Learning (ML) methods to mimic human functions (such as learning capacity and adaptation to changes) and to create computer programs for analysis of datasets [3, 4].

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 674–682, 2008.

Research shows that ML techniques can be applied in healthcare environments where an automated process must adapt to changing conditions, improve its performance based on previous data, extract knowledge from examples in a database, and deal with uncertain and incomplete medical knowledge [5].

There are many different types of clinical tasks to which ML tools can be applied. For example ML tools can assist detection of microcalcifications in mammography [6], analyze Sudden Infant Death Syndrome (SIDS) [7] and diagnose thyroid disorders [8].

One of the underlying purposes of this study is to explore the use of intelligent multi-agent approach for developing e-health systems for the prediction of kidney transplant outcomes. Currently, medical professionals around the world use blood types, the number of Human Leukocyte Antigen (HLA)[1] mismatches and the basic cross-match tests to determinate the best match between kidney donors and recipients. However, the current matching and testing scheme do not always provide a successful outcome. Literature search revealed the chance of a successful cadaver transplant surviving for one year after the transplant is about 85% and 50% at five years[2].

The proposed model is based on the use of a novel classification approach which is a combination of initial data preparations, preliminary classification by ensembles of neural networks, generation of new training data based on criteria of highly accuracy and model agreement, and decision trees. The proposed model can be integrated to a multi-agent system. Furthermore, it can tackles the problem of complexity of clinical data and non-transparency of black box approach such as neural network models by developing an algorithm with reference to the bagging-based ensemble and the hybrid decision tree-neural networks ensemble. The case studies described in this study are from a kidney transplant database (a complex scenario) and two well-known collections of benchmark data known as the Pima Indian Diabetes (a semi-complex scenario) and Wisconsin Cancer datasets (a non-complex scenario). The Pima Indian Diabetes and Wisconsin Cancer datasets are stored in the UCI repository [9, 10]. These case studies provide examples of the challenges involved in real life data mining in clinical settings.

Furthermore, we explore the use of intelligent multi-agent approach for developing e-health systems for the management of chronic diseases (such as diabetes) in general and hopefully better reveal the potential of e-health. E-health can be used to remove the time and distance barriers from the delivery of health care services and consultations to patients.

2 Background

2.1 Artificial Intelligence in Medicine

From the early 1950s, scientists and healthcare professionals have been aware of the potential of computer systems in medicine. In 1959, Ledley and Lusted published a

[1] The HLA is the general name of a group of genes in human cells that enable our body to tell the difference between our own cells and foreign ones.

[2] Transplant Service of the University of Texas Medical Branch at: http://www.utmb.edu/renaltx/srate.htm (Last accessed: 25 July 07).

paper entitled, "*The reasoning foundations of medical diagnosis*" [11]. In this paper they point out the potential for computer programs to assist physicians with diagnosis tasks. They applied their idea by using mathematical methods of Boolean algebra and symbolic logic to solve medical problems. Over the years AI systems, along with performing tasks that require reasoning with medical knowledge have played a significant role in the process of scientific research. Most significantly, AI systems have the capacity to learn, which has lead to the discovery of new phenomena and the creation of medical knowledge.

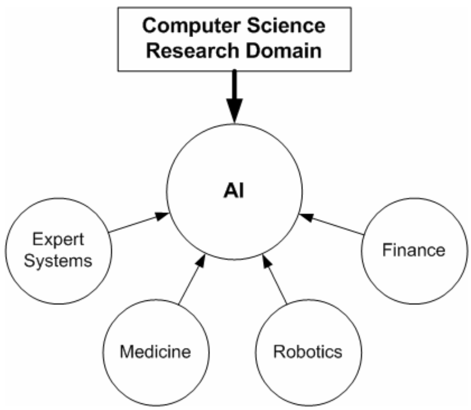


Fig. 1. Some of the application areas of artificial intelligence

Recently, AI systems have been involved in clinical decision support systems, mining medical databases, effective diagnostic assistance and treatment methodologies and intelligent systems for hospital, home and community health care. Diagnostic assistance is when AI systems help in developing a possible diagnosis for a patient based on their patient data. Image recognition and interpretation of medical films such as X-rays, CT and MRI scans, are areas were AI systems can indicate potential abnormal images. They can also be used in pharmacological sectors. For example, a learning system can be given examples of one or more drugs that weakly exhibit a particular activity, and based upon a description of the chemical structure of those compounds, the learning system can then suggest which of the chemical attributes are necessary for that pharmacological activity. Overall, AI can be considered as an attempt to build intelligent computer programs that behave like humans, furthermore it can also be described as a broad discipline with many applications, some of which are shown in Fig. 1.

2.2 Expert System Technology

Expert systems are designed to act as an intelligent assistant to a user or human expert in one problem domain [12]. The general structure of an expert system is shown in Fig. 2. An expert system usually has three components:

i. External data sources.
ii. A knowledge base (supplied by users), which stores the facts, related to the problem domain, their relations to one another, and perhaps some heuristic knowledge (less rigorous and more experiential). The knowledge within an expert system is generally represented in the form of a set of rules.
iii. A reasoning engine, which operates on the knowledge base and external data sources (performs the knowledge based reasoning) to draw conclusions, answers questions, and gives advice.

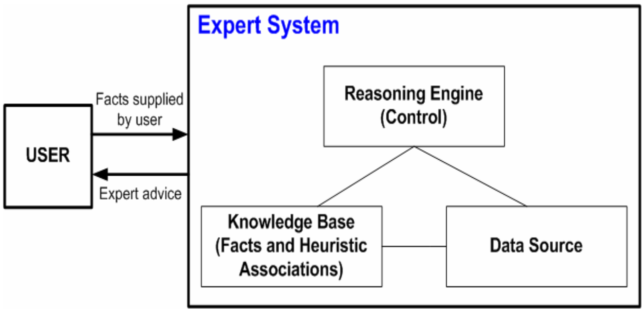


Fig. 2. Basic components of an expert system [12]

2.3 Multi Agents Technology

Agents are software entities that carry out some set of operations on behalf of a user or other programs. Agents usually have a set of beliefs, goals and rules to select plans and are capable of communicating messages concerning their beliefs and goals. Several agents can collectively and collaboratively form a Multi-Agent System (MAS) to perform complex tasks that cannot be done by individual agents. For researchers who are working in the field of Artificial Intelligence (AI), the term agent generally means a software entity that use AI techniques to learn and adapt to its environment in order to perform a set of actions and attain the goal of a user or other programs. Intelligent Multi-Agent System (IMAS) frameworks usually have capability to deal with complex systems and support agent mental state representation, reasoning, negotiation, planning and communications. Recently intelligent agents and multi-agent systems have attracted much interest in data mining community. AI-based data mining methods such as genetic programming have been applied to many multi-agent learning tasks [13]. Over the past few years GP has been successfully used for embedding intelligence in newly crated agents [14, 15].

There are numerous software agents in the market. Many of these software packages have proven to be very useful in variety of applications. Etzioni and Weld [16] provide a good reference for currently available intelligent software agents. One of the recent examples of software agent systems is Agent Academy (AA) [17]. The Agent Academy framework operates as an IMAS, which can train new agents or retrain its existing agents in a recursive mode. A user can issue a request to the Agent Factory (AF) for a new agent. In this system an untrained agent (UA) can have a minimal degree of intelligence that has been defined by the software designer. This agent then can be send to Agent Training (AT) module, where its world perception can be developed during a virtual interactive session with an Agent Master (AM). Perhaps the most important part of this framework is the part that store statistical data on prior agent behavior and experience. The Data Miner (DM) module then can use the stored information and mine the data, in order to generate useful knowledge about the application domain under the investigation. Knowledge extracted by the DM can be used by the AT module to train new agents or to enhance the behavior of existing agents.

3 Experimental Results

This study tries to handle the complexity of clinical data and rule extraction from individual black box classification models such neural networks by studying the data and designing a novel hybrid-learning model called RIDC-ANNE (see [18] for detail implementation of RIDC-ANNE).

This novel algorithm was successfully developed with reference to the data filtering, bagging-based ensemble and the hybrid decision tree-neural networks ensemble theories [19, 20]. The RIDC-ANNE approach assists the data preparation process by configuring an ensemble of bagged networks as a filter and identifying the regions in the data space that have been consistently misclassified or have high impact on the system performance. The RIDC-ANNE technique treats the ensembles as a black box or data filtering tool in order to provide a body of new training examples to feed standard rule induction or decision tree systems. This strategy considers the diversity and expertise of the component networks in the rule generation process. However, unlike other rule generation techniques (see for example: Shadabi, 2007), it does not focus on the identifying ensemble members that are relevant in explaining the prediction (output) associated with a particular case. Instead RIDCANNE explains the output of the ensemble based on a cluster of cases that consistently generate agreement across the classifiers with similar expertise. The use of RIDCANNE not only produces rule sets that are substantially easier to understand but it also reduces the computational cost, since it creates a cluster of new training examples that allow faster rule generation process.

The RIDC-ANNE approach uses the Bagging algorithm to train neural networks and extracts the input patterns (examples) that were included across the ANN series in the final results. In this strategy for each of the patterns or examples in the validation set, the votes were recorded first. Then the system searched for patterns or examples that caused at least 50% of the networks to agree on one category (fail or success). Then, it searched for patterns that caused 51% agreement among the networks. This procedure was repeated until the model reached the 100% agreement among the networks for one category (fail or success). It should be noted that as the model approaches the 100% agreement condition, fewer patterns were able to satisfy the condition and be included in the report. Finally, for each condition (i.e. network agreement), the prediction accuracy and the number of examples that satisfy the condition were reported by the system.

This investigation revealed that by using 19% of data points (84 cases), the validation set can reach a 70.2% accuracy rate (with 87% agreement among the networks).

It is important to note, these results are based on hold out evaluation scheme, See[19, 20]). Furthermore, a close study of the rule set generated in this scenario revealed that the donor state and age were considered two important factors in determining the success of kidney transplants.

Furthermore, the potential of the RIDC-ANNE strategy was also investigated in the application of two well-known medical datasets, Pima Indian Diabetes and Wisconsin Cancer and compared our results with the results obtained by other researchers whenever possible. For the Pima Indian Diabetes data, using the majority voting, the balanced test set reached around a 77% accuracy rate. This result was slightly better than previous comparison studies. However, by using only 93% (239 cases) of data points, it was possible to achieve an 80% accuracy rate with a 68% agreement among the networks. The accuracy rate also reached 87% with a 99% agreement among the networks, when only 152 examples (59% of data points) were used. Not surprisingly, by removing only 0.5% of the Wisconsin breast cancer (which is known to be a well recorded and non-complex dataset) data points, it was possible to generate 100% agreement among the networks and achieve an accuracy rate of 98.6%.

The following rule set was produced by applying the RIDC-ANNE approach (to kidney transplant data) based on the 84 examples whose class labels (outputs from the ensemble) were in agreement across 87% of classifiers:

a. *donstate <= 4 AND donage <= 43:* ***Success***
b. *donhosp <= 109 AND misa <= 1 AND refstate > 3 AND misb > 0 AND refhosp > 94:* ***Failure***
c. *misa <= 1 AND misb <= 1:* ***Success***
d*. donsex > 0 (i.e. Female):* ***Failure***
e. *misa > 1:* ***Failure***
f. *age > 27:* ***Success***

As demonstrated above, the rule set produced only six rules. These rules were valid for 97% of cases (82 cases) in the kidney transplant dataset. In the above rule set, the donor state and age were both considered to be important factors for the success of kidney transplants.

4 Combining the RIDC-ANNE and Multi Agent Technology

We do not encourage the use of MS Word, particularly as the layout of the papers (the position of figures and paragraphs) can change from printout to printout. Having said this, we do provide the template sv-lncs.dot to help MS Word users prepare their camera-ready manuscript and to enable us to use their source files for the online version. There are several reasons why IMAS is a suitable technical alternative for E-health systems (see for example: [21]). In this section we propose a methodology for IMAS development in the telemedicine system for kidney transplant problem, which relies on the ability of RIDC-ANNE or DM to generate knowledge models for agents. The roles of all agents and their responsibilities and services are detailed as follows (see also Fig. 3):

1. *Personal agent* is one type of interface agents, which provides the user with a graphical interface to the multi-agent system and initiates a search or shows the results of a query to the user. This agent also known as *Diagnosis agent* can be used to analyses the situation and makes an accurate judgment for the patient. It can also be designed in a way to implement the method of therapy agent.
2. *Monitoring agent* can be used to monitor the patients (after the surgery) in real time and send out the monitored data to data processing agent. It can also be designed in a way to send alert to the doctor in case of emergency (based on the collected data).
3. *Data processing agent* can provide statistic and integrate the monitored data.
4. *Therapy agent* has the capability to determine a proper therapy method.
5. *Consultation agent* can be used to provide consultation to the enquiry of patients and contacts with diagnosis or personal agent.
6. *Decision support agent* has the capability to provide decision support and cooperation for personal agent.
7. *Training agent* use knowledge extracted by the Data Miner in order to train new agents or enhance the behavior of agents already in running. The Data Miner is the core component of this IMAS prototype and can be used to provide the user with a number of DM algorithms. The DM algorithms are then validated before they can be store as a set of facts in knowledge based models of a rule engine. For Example, as described in previous section the generation of the rules can be done through RIDC-ANNE.
8. *Archival agent* integrates with the Medical database and has the capability to edit and store the patient record and therapy methods, and updates the database of the individual patients.

Figure 3 shows an agent-based model for designing an E-health system for the management of patients and the intersecting spheres of view of Agents of a dataset inside the Data Miner. The proposed Multi-agents system can potentially improve the speed of data mining task and can be used to automatically answer the user's questions in real time.

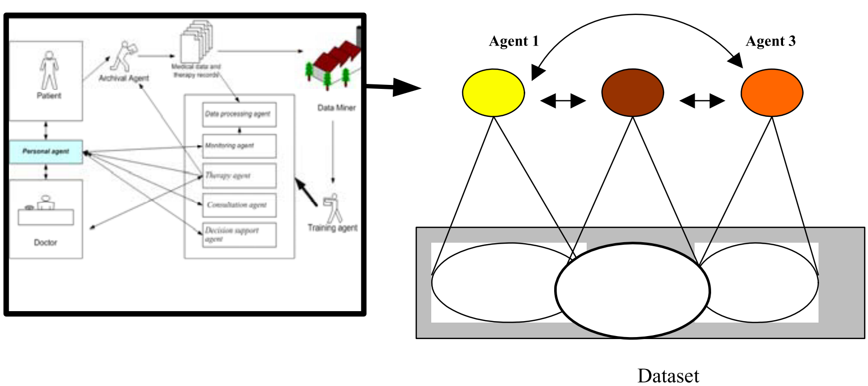


Fig. 3. Multi Agent Based architecture for designing an E-health system for the management of patients after Kidney transplantation

5 Conclusion

Recently, researchers in data mining and knowledge-based systems have turned to intelligent multi-agent approach. Research shows that integrating AI-based data mining modules into computerized patient records and clinical findings can provide a great chance for improving the quality of care. This study demonstrated the potential of E-health and the use of intelligent multi-agent approach for developing E-health systems for the prediction of kidney transplant outcomes.

In summary, the results showed that the application of the RIDC-ANNE model in predicting outcomes for a complex dataset such as kidney transplantation is promising. The method presented in this study proved useful in explaining the kidney transplant data and also provides basis for improved methods for matching the donor-recipient pairs with higher probability of successful transplant. The model can be easily integrated with the proposed multi-agent technology based model. Future work will investigate this approach in more details.

References

1. Hosmer, D.W., Lemeshow, S.: Applied logistic regression. John Wiley, New York (1989)
2. Bagley, S., White, H., Golomb, B.: Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. The Journal of Clinical Epidemiology 54(10), 979–985 (2001)
3. Schwartz, M.H., et al.: Using neural networks to identify patients unlikely to achieve a reduction in body pain after total hip replacement surgery. Medical Care 35, 1020–1030 (1997)
4. Pesonen, E.: Is neural network better than statistical methods in diagnosis of acute appendicitis? In: Pappas, C., Maglaveras, N., Scherrer, J.R. (eds.) Medical Informatics, pp. 377–381. IOS Press, Amsterdam (1997)
5. Li, L., et al.: Selecting pre-screening items for early intervention trials of dementia-a case study. Statistical Medicine 23(2), 271–283 (2004)
6. Woods, K.S., et al.: Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. Pattern Recognition and Artificial Intelligence 7(6), 1417–1436 (1993)
7. Wilks, P.A.D., English, M.J.: Accurate segmentation of respiration waveforms from infants enabling identification and classification of irregular breathing patterns. Medical Eng. and Physics 16(1), 19–23 (1994)
8. File, P.E., Dugard, P.I., Houston, A.S.: Evaluation of the use of induction in the develeopment of a medical expert system. Comp. and Biomedical Research 27(5), 383–395 (1994)
9. UCI Repository, UCI repository of machine learning databases (2005)(Access December 10, 2005), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
10. Mangasarian, L., Wolberg, H.W.: Cancer diagnosis via linear programming. Society for Industrial and Applied Mathematics (SIAM) 23(5), 1–18 (1990)
11. Ledley, R.S., Lusted, L.B.: Reasoning foundations of medical diagnosis. Science 130, 9–21 (1959)
12. Giarratano, J.C., Riley, G.D.: Expert Systems: Principles and Programming, 3rd edn., p. 624. Course Technology Publishing Company (1998)

13. Luke, S.A., Spector, L.: Evolving Teamwork and Coordination with Genetic Programming. Genetic Programming 1996. MIT Press, Cambridge (1996)
14. Iba, H.: Evolutionary Learning of Communicating Agents. Information Sciences 108, 1–4 (1998)
15. Iba, H.A., Kimura, M.: Robot, Programs Generated by Genetic Programming. Advanced Institute of Science and Technology, IS-RR-96-0001I, in Genetic Programming (1996)
16. Etzioni, O., Weld, D.S.: Intelligent Agents on the Internet - Fact, Fiction, and Forecast. IEEE Expert - Intelligent Internet Services (4), 44–49 (1995)
17. Mitkas, P.A.: Agent Academy: A Data Mining Framework for training Intelligent Agents. In: European Projects Presentation Day of the 13th Eur. Conf. on Machine Learning & 6th Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2002), Helsinki, Finland (2002)
18. Shadabi, F., Sharma, D., Cox, R.: Learning from Ensembles: Using Artificial Neural Network Ensemble for Medical Outcomes Prediction. In: Proceedings of the 3rd International Conference on Innovations in Information Technology (IIT 06)- Provided by IEEE Explore, Dubai, UAE (2006)
19. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
20. Shadabi, F.: Medical Outcome Prediction: A Hybrid Artificial Neural Networks Approach, in School of Information Sciences and Engineering, University of Canberra: Australia (2007)
21. Tian, J., Tianfield, H.: A Multi-agent Approach to the Design of an E-medicine System. In: Schillo, M., Klusch, M., Müller, J., Tianfield, H. (eds.) MATES 2003. LNCS (LNAI), vol. 2831, pp. 85–94. Springer, Heidelberg (2003)

A Middleware Platform Based on Multi-Agents for u-Healthcare Services with Sensor Networks

Yung Bok Kim[1], Chul-Su Kim[2], and Jun Wook Lee[2]

[1] Sejong University
KunJa-Dong, Kwang-Jin-Ku, Seoul, Korea 143-747
yungbkim@sejong.ac.kr
[2] RFID/USN Middleware Team, ETRI
Daejeon, Korea 305-700
{chulsu1, junux}@etri.re.kr

Abstract. We studied the middleware platform, i.e. COSMOS (Common System for Middleware of Sensor Network), based on multi-agents for u-healthcare services with various types of sensor networks such as the Zigbee wireless sensor network (WSN), the CDMA cellular network, RFID and IP-USN based on the 6LowPAN. Our project is focused on sensor network abstraction for various ubiquitous sensor networks (USN). We defined standard interfaces between USN middleware based on multi-agents and USN networks as well as application services with multi-agents. We implemented several USN services such as u-Healthcare services to consider important issues about middleware platform based on multi-agent applications.

1 Introduction

The fields of RFID and sensor networks are being studied actively as evolving technologies and services for new growth engine of Korean industry. Mobile RFID technologies, which use mobile phones equipped with RFID readers, were deployed as a new mobile service in Korea, after standardizing specifications by the Korean Mobile RFID Forum. A user with mobile device can get information associated with the specific RFID tag attached on a product. The various sensor networking technologies using wired RFID readers, mobile RFID readers, and sensor networks could be focused on restricted applications or on specific locations for environment monitoring services, etc. The main disadvantage with these sensing devices is lack of a common, scalable network infrastructure for collecting, filtering, extracting of the useful information as well as for enabling distributed queries from sensing devices.

In the sensor network, research has mainly focused on routing, aggregation, and energy efficient data management algorithms inside one sensor network. The deployment, application development, and standardization aspects have usually not been addressed [1]. In the last few years, however, various middleware solutions for

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 683–692, 2008.

sensor networks have emerged. These differ in terms of their models for querying and data aggregation, and their assumptions about the topology and other characteristics of the network. Naturally, the assumptions made for each particular middleware limit its potential applicability [2]. We need to study the USN middleware for various applications based on multi-agents in a unified and standardized way.

The middleware for ubiquitous networks, i.e. USN middleware, was considered to allow various applications with the available sensor information and computing resources in standardized way. For USN middleware based on multi-agents, we abstracted lower layers related to various sensor networks as well as RFID readers; various application services can be developed above the USN networks. USN middleware shall support context-aware processing based on qualified data-acquisition for context-aware work-flow based on multi-agents. Service discovery is a key function to implement context-aware applications for mobile users in ubiquitous computing environments; an efficient semantic service discovery scheme called UbiSearch [3] for a large-scale ubiquitous computing environment was introduced. We studied and present USN Directory Service for unified service for appropriate service discovery based on multi-agents.
We developed the USN middleware platform in the COSMOS project, a (Common System for Middleware of Sensor Network) middleware platform developed by ETRI in Korea. We developed also four kinds of application services operating on the platform based on multi-agents. We briefly introduce one application service among them, i.e. the u-Healthcare service.

In the following sections, we will discuss the background of the status of USN middleware, the USN middleware platform based on multi-agents, examples of service and important issues, u-Healthcare and an analysis model for QoS, and finally we will conclude our study.

2 Background of Middleware

Consider WSN middleware as a software infrastructure that glues together the network hardware, operating systems, network stacks, and applications [4]. A complete middleware solution should contain a runtime environment that supports and coordinates multiple applications based on multi-agents, and standardized system services such as data aggregation, control and management policies. There are many types of middleware based on various technologies for specific requirements. The basic goal of middleware is to provide various application services with abstractions based on multi-agents for lower layer transparency. The applications need not be concerned about how they can get required data physically. The concept of USN middleware is very similar to other normal middleware. But the difference is that USN middleware covers not particular devices but also various kinds of information acquisitive networks such as various sensor networks. In addition, it provides context-aware functionalities, intelligent event processing, real-time sensed-data processing. Furthermore, we also put the functionality of sensed-data mining into the USN middleware platform.

There have been many researches on sensor network middleware all over the world. MiLAN (Middleware Linking Applications and Networks) [5] suggests methods for user QoS as well as for the extension of the lifetime of sensor networks. MiLAN couples tightly the application's request with network operations. Cougar [6] considered the sensor networks as a big DB. It creates a set of queries, i.e. query plan, from an application query for in-network processing by edge or intermediate sensor nodes. It optimizes the resource usages of the whole sensor network, the reduction of network traffic as well as the lifetime extension of sensor network. There are more middleware platforms for the sensor network, and each middleware platform has its targeted purpose and characteristics, e.g. SINA (Sensor Information Networking Architecture and Applications) [7], DSWare (Data Service Middleware) [8], etc. Sensor networks can be used for various applications even including environment monitoring such as SensorWeb [9] and IrisNet [10]; these projects dealt with wide-area architectures to use distributed sensing devices.

The aforementioned middleware platforms based on multi-agents for sensor networks were developed to meet specific goals such as in-network processing, linking applications and networks, clustering, etc.; however, they were not considered enough to handle heterogeneous sensor networks. We tried to develop an intelligent and autonomous service platform in our USN middleware platform project: COSMOS. Our USN middleware shall cover various heterogeneous sensor networks, RFID reader including mobile RFID reader in mobile phone, CDMA mobile devices as a sensor/sink/gateway node, and IP-USN networks. Additionally, the COSMOS middleware platform supports intelligent processing capabilities such as sensed-data mining, context-awareness, intelligent management and unified metadata directory services on the basis of multi-agent system.

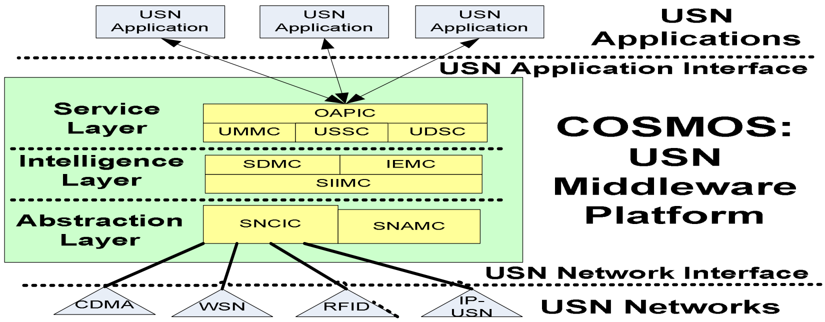
The standardization for interoperability is important because of the heterogeneous middleware platform, we are trying to contribute for standardization of middleware platform. At the latest ITU-T SG16 meeting in Geneva 2007, the concept of USN middleware was introduced as a new work area for USN standardization, and the discussion with revised contributions will be continued in the following meetings.

3 USN Middleware Platform Based on Multi-agents

Fig. 1 shows the architecture of COSMOS: a USN middleware platform based on multi-agents. This platform has been under development since 2006. The following Fig. 1 shows COSMOS middleware platform with acronyms are as follows:

USN middleware comprises of Service Layer, Intelligence Layer and Abstraction Layer. Service Layer comprises of OAPIC (Open API Component), UMMC (USN Middleware Management Component), USSC (USN Middleware Security Service Component) and UDSC (USN Directory Service Component).

This layer provides application interface for various services. Among the components in this service layer, we focus on the components for a unified and ubiquitous



OAPIC: Open API Component, UMMC: USN Middleware Management Component

USSC: USN Middleware Security Service Component, UDSC: USN Directory Service Component

SDMC: Sensor Data Mining Component, IEMC: Intelligent Event Management Component

SIIMC: Sensor Information Integration Management Component

SNCIC: Sensor Network Common Interface Component

SNAMC: Sensor Network Autonomous Management Component.

Fig. 1. COSMOS: a USN Middleware Platform

web service. UDSC contains various metadata to find appropriate resources such as components or sensor networks. If some services are required to monitor an area, then the USN Directory Service can help to look them up. USN Directory Service offers all relevant information such as location, wireless protocol, sensor type, number of sensor nodes, sensor network lifetime, etc. Scalability and modularity are very important factors in distributed environments. For a unified and ubiquitous service in heterogeneous middleware environments, we need to make this UDSC into a unified-UDSC. This will be discussed further as an issue for consideration in the following section.

The Sensor Network Intelligent Layer is composed of a SDMC (Sensor Data Mining Component), an IEMC (Intelligent Event Management Component) and a SIIMC (Sensor Information Integration Management Component). This layer provides intelligence for sensed-data processing, event processing and context information processing. It requests data from the wireless information infrastructure, i.e. USN networks, through Sensor Network Abstraction Layer.

The SDMC creates query plans to request data. This component is in charge of simultaneous scheduling for requests as well as processing for responses. This component provides mining functionality for sensed-data. The IEMC models an application specific event and notifies the requesting service of the result. It may use the SIIMC to obtain proper sensed information or use a sensor network monitoring component. The SIIMC models service-specific context information and defines the rules to capture the specific context. This component may use SDMC to obtain context information. This component enables USN middleware to provide intelligent context-aware processing.

The Sensor Network Abstraction Layer is comprised of the SNCIC (Sensor Network Common Interface Component), the SNAMC (Sensor Network Autonomous Management Component) and adaptor components. This layer provides wireless information infrastructure abstraction and sensor network monitoring functionality. The wireless information infrastructure includes RFID readers, various kinds of sensor networks, CDMA mobile devices as sensor/sink/gateway node and an IP-USN router/node, etc. This layer provides an abstract interface to the upper layer to use any kind of USN network. The adaptor is responsible for transforming data format and command format between USN middleware and actual wireless infrastructure protocols. The abstraction layer is realized for various USN networks by various adaptors.

4 Services and Issues

We developed four different services to implement USN middleware functionalities and to realistically estimate the performance. The four services are water and sewage service, u-SilverCare service, disaster surveillance service and ocean environment monitoring service. We developed these services respectively, and we will integrate the four services to provide real-time integrated information services. The wireless sensor network is deployed with sufficient sensor nodes which can sense temperature, smoke and gas. Sensed-data is transmitted to USN middleware. The USN middleware based on multi-agents receives real-time sensed-data from the sensor network and then processes them according to the service requests. The USN middleware can provide sensed-data to the application service. The USN middleware can provide some context-aware information defined by service requirements. To take context-aware functionality of USN middleware, each service is defined with certain rules to execute in that context.

When recognizing a fire, in the case context rule defined, the USN middleware runs automated work-flows which are defined by the service. The USN middleware makes an emergency call to the Emergency Hospital located nearby and alerts a fire occurrence to the Fire Defense Headquarter(s). The Disaster Surveillance Service is very critical to save many people and minimize the damage by fire. The above scenario makes it possible to automatically handle a fire without human intervention. The USN middleware is characterized by the coverage of a wireless infrastructure, context-awareness, and a monitoring and standardized interface for accessing the wireless infrastructure. We considered the characteristics of COSMOS middleware platform and will discuss some important issues for a unified and ubiquitous web service.

Unlikely other sensor network middleware, which deals only with a sensor network, USN middleware covers not only sensor networks but also RFID readers, CDMA mobile devices and IP-USN nodes. Support of RFID readers is very important in various service areas based on RFID tags and readers. For example, an RFID tag is attached to the refrigeration container for cold goods, and several sensor nodes are deployed for the refrigeration container. The RFID tag may be associated with the following information: container number, driver's name, type of products, appropriate environmental parameters, etc. If a container passes below a specific RFID reader, then the above information is sent to the appropriate application for service. At the same time, an application can obtain sensed-data via a sensor network deployed in the product container. By combining these two kinds of information, the application keeps tracking the status of the products with the associated information. Various service models may be introduced.

USN middleware support this service with two different ways.

 Loosely-coupled way: we designed the Sensor Network Abstraction Layer with two parts. First is the Sensor Network Common Interface Component (SNCIC), which deals with sensor network. The second is RFID Reader Interface Component (RIC), which deals with only RFID reader and mobile RFID reader. Therefore, there are two paths of data flow; one is for the RFID data flow through the RIC and the other is for sensed-data flow through the SNCIC.
 Tightly-coupled way: we designed Sensor Network Abstraction Layer with just one part. Sensor data and RFID data flow through one path, the SNCIC. We integrate sensed-data abstraction and RFID abstraction into the SNCIC. The RFID tag value is treated like sensed-data similar to normal sensed-value from sensor network. We expect to handle the smart RFID tag, which has both communication and sensing capabilities.

We abstracted the sensor network by defining the standard interface between the USN middleware and the sensor network. We decided to use the communication transport protocol with a TCP/IP. The abstraction was accomplished as following steps.

 First, we standardized the sensing type as temperature, remaining battery lifetime, pressure, gas gauge, bpm (beat per minutes), etc.
 Second, we defined common message formats used between the USN middleware and the adaptors. We categorize all possible messages into 3 groups. First group is data request/response. This group contains messages associated with sensor data request/response. The second group is command/notification. This group includes asynchronous sensor data request/notification, activate/deactivate of actuator and error report. The last group is for communication control between the USN middleware and adaptors. This definition of common message format will be refined after further development.

Context-aware functionality is accomplished as following steps.

 First, we defined a common model. Service can be defined in association with context-aware rules. It generates appropriate requests for sensed-data to get relevant information from information providers.
 Second, we defined a certain work-flow function. If target context is aware, then USN middleware can trigger some work flow, and we call that context-aware work-flow.

Context-aware processing may use the current status of sensor networks.

As an important issue for a unified and ubiquitous service, let's consider the real environments prevailing with various middleware platforms. With various middleware platforms, we need to consider a unified scheme for ubiquitous sensor web. We studied the unified USN directory service instead of each UDSC in a middleware platform as in the Fig. 2. The USN directory service should be out from the middleware and should have a unified and standardized interface and metadata format. We have tried to develop the unified UDSC using web technology and DB.

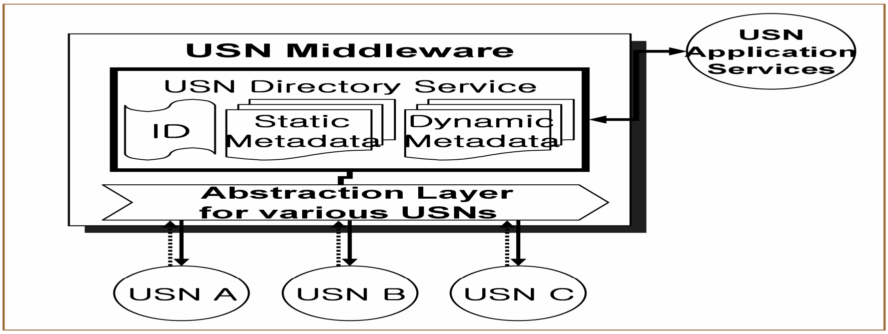


Fig. 2. USN Directory Service for Unified Service

5 u-Healthcare and Analysis

Among various application services mentioned in the previous section, we will focus on the u-healthcare service among them. Fig. 3 shows the u-SilverCare service as a specific example of the u-Healthcare service using the USN middleware platform.

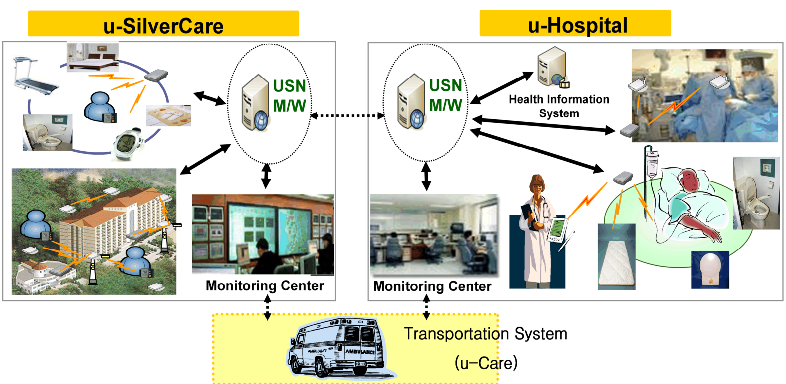


Fig. 3. u-SilverCare Service as a USN Service

Let us further discuss the u-Healthcare service considering important issues such as data quality of sensor nodes, real-time aggregation and estimation of random variables for comprehensive accessibility. Fig. 4 shows health-monitoring sensors in the wrist phone on the hand of the disabled and elderly persons as a primitive example with single-hop sensor nodes; that single-hop sensor node is a starting point to get insight concerning important issues in terms of data analysis with random variables. The wrist phone as a sensor node is a future product being implemented. We assume that the sensors for measuring the pulses, the strength of the pulses, and blood pressure, etc. would give the raw data about health of the disabled and elderly. We studied the real-time health-monitoring application, frequently getting personal information as well as writing analyzed information as a response to requests from clients or health-monitoring agents in the mobile Internet environment. We will discuss various aspects of the QoS of a ubiquitous application service.

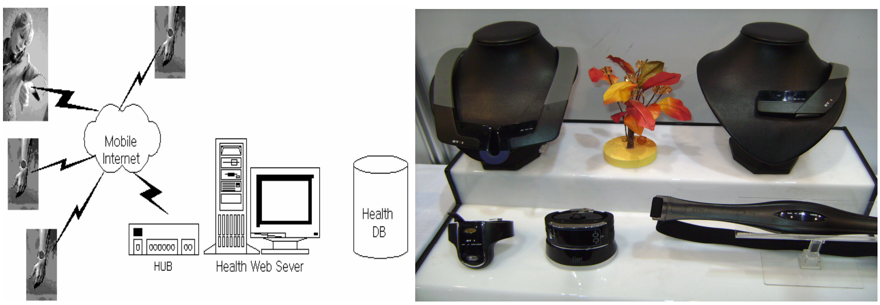


Fig. 4. u-Healthcare Service with Sensor Nodes (Wrist Phone, etc.)

For the real-time health-monitoring network using mobile Internet, as shown in the Fig. 4, the dominating factor and the standard deviation of that random variable should be bounded within a deterministic response time. To be deterministic for real-time application, the estimation time should be bounded within deterministic time, the interchange of data between the watch phone and the server should be automatic except the requested information by the user; therefore the Web server based on multi-agents should be efficient and have high performance for the dedicated application if possible, and the exchanged data and analyzed information should be as simple as possible with a simplified and efficient format. If possible, the bandwidth requirement for wireless or mobile Internet should be immune to network traffic conditions; that will also be ideal in terms of the degradation caused by the other rich multimedia contents sharing the same network and server.

The time spent in the Web server may be considered to be immune to the network and server condition, with short packets below 1.5Kbytes that may be the upper bound of packet size. In general, the frequency of health monitoring is much more than the frequency of analyzed or diagnosed information from the health Web server. This system based on multi-agents is based on wired or mobile Internet, and the monitored health data from the sensors in a watch phone can be registered at any time using mobile Internet with the domain name of test Web server for wired/mobile Internet. This site has actually been used as a real-time information network server and we considered it as the health-monitoring server because it can be used as any Web server for testing.

From the user's perspectives for the disabled, considering the application of pervasive (or ubiquitous) computing technologies in healthcare, we studied the important QoS performance metric, delay. The preparation time for the disabled to get a service (e.g. medical device, mobile device, etc.) is *U* (ubiquitous environment time metric); the time spent by the disabled with medical device to do appropriate action for service is *D* (device time metric); the aggregate time to the medical healthcare server after the medical device for medical service is *S* (service time metric for communication and Web server); the time depending upon medical contents is *C* (contents time metric). Among the above random variables, i.e. the performance metrics, (*U, D, S, C, P*) for the disabled, the most dominating factor, i.e. the random variable, may be different depending upon the U-healthcare environment, device, service, contents, and patient.

We can represent the statistical values (i.e. mean, deviation, etc.) of the random variables as performance metrics for QoS (quality of service) of real-time medical service, and the quality of the u-Healthcare service can be compared with the QoS performance metrics of comprehensive accessibility.

Each performance metric can be used to enhance the QoS as follows. Ubiquitous environment time metric *U* will be shortened depending upon the proliferation of smart environments for u-Healthcare. Device time metric *D,* i.e. handling speed of medical device, is one of important performance factors in any medical service for the disabled. Service time metric *S* can be shortened according to the communication facility and Web server with the DB in e-hospital. Contents time metric *C* can be shortened considering accessibility in contents design for the u-Healthcare service. Patient time metric *P* may be different from person to person depending upon the disability of the disabled.

We can order the dominating factors in the overall performance at the user's perspective. The user interface design for the disabled with a medical device in a ubiquitous healthcare environment (i.e. *U* becomes smaller as the ubiquity increases) is important to decrease the device time metric *D* that is heavily related to the UI (user interface) convenience of medical device for the disabled and handicapped users. With the performance metrics (*U, D, S, C, P*), we can estimate other metrics for QoS as follows: for example, total service time for a patient, *T*. We can use mean values of the random variables for the simplicity of metrics as follows: $\overline{U}$, $\overline{D}$, $\overline{S}$, $\overline{C}$, $\overline{P}$ and $\overline{T} = \overline{U} + \overline{D} + \overline{S} + \overline{C} + \overline{P}$.

In the case of web contents, the web-contents accessibility is more important than the elapsed time to access Web contents, if we consider the unified and ubiquitous web service. We found also that there was a correlation between web-contents accessibility and performance of comprehensive accessibility. We implemented the unified and ubiquitous web portal that both complies with the guidelines of web-contents accessibility and has good performance for soft-real-time unified and ubiquitous web service. With more proliferation of ubiquitous network environment, the time spent with the ubiquitous device with sensor nodes will become more critical in the multi-agents system.

6 Conclusions

We studied a USN middleware platform based on multi-agents, i.e. COSMOS, for ubiquitous environments. USN middleware covers various USN networks such as sensor networks, RFID readers including the mobile RFID reader, and CDMA mobile devices and IP-USN nodes. We introduced a USN middleware platform defined with standardized interfaces between wireless infrastructure and application services using multi-agents. The proposed USN middleware platform provides context-aware data processing. We focused on a u-Healthcare service to consider important issues such as data quality and comprehensive accessibility in the USN environment. As a further research, we are developing a unified scheme of metadata directory service in the heterogeneous USN middleware environment on the basis of multi-agents system.

References

1. Aberer, K., Hauswirth, M., Salehi, A.: A Middleware For Fast And Flexible Sensor Network Deployment. In: VLDB 2006, Seoul, Korea, September 2006, ACM, New York (2006)
2. Henricksen, K., Robinson, R.: A Survey of Middleware for Sensor Networks: State-of-the-Art and Future Directions. In: MidSens 2006, Melbourne, Australia, November 2006, ACM, New York (2006)
3. Kang, S., Kim, D., Lee, Y., Hyun, S.J., Lee, D., Lee, B.: Semantic Service Discovery Network forLarge-Scale Ubiquitous Computing Environments. ETRI Journal 29(5), 545–558 (2007)
4. Hadim, S., Mohamed, N.: Middleware: Middleware Challenges and Approaches for Wireless Sensor Networks. IEEE Distributed Systems Online 7(3), 1–23 (2006)
5. Murphy, A., Heinzelman, W.: MiLAN: Middleware Linking Applications and Networks. Thesis, Rochester, NY (January 2003)
6. Yao, Y., Gehrke, J.: The Cougar Approach to In-Network Query Processing in Sensor Networks. Thesis, Cornell University (2002)
7. Shen, C., Srisathapornphat, C., Jaikaeo, C.: Sensor Information Networking Architecture and Applications. IEEE Personel Communication Magazine, University of Delaware (2001)
8. Li, S., Lin, Y., Son, S., Stankovic, J., Wei, Y.: Event Detection Services Using Data Service Middleware in Distributed Sensor Networks. Thesis, University of Virginia (2003)
9. Delin, K., Jackson, S.: The Sensor Web: A New Instrument Concept. California Institute of Technology, Pasadena (2001)
10. Deshpande, A., Nath, S., Gibbons, P., Seshan, S.: IRIS: Internet-scale Resource-Intensive Sensor Services. Intel Research, U.C. Berkeley, Carnegie Mellon University (2003)

Measuring Similarity of Observations Made by Artificial Cognitive Agents

Grzegorz Popek and Radosław Katarzyniak

Institute of Information Science and Engineering, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-350 Wrocław, Poland
{grzegorz.popek, radoslaw.katarzyniak}@pwr.wroc.pl

Abstract. A model for artificial cognitive agent has been considered in which the agent can carry out autonomous observations of an external world and stores them in the so called base profiles. For this model four general approaches to defining similarity of stored observations have been proposed and a model for their integration into one computational mechanism has been presented. For each strategy an example similarity function has been given, fulfilling some commonsense expectations defined for strategies of comparing the content of individual observations.

Keywords: cognitive agent, empirical experience, distance measure, similarity measure.

1 Introduction

The ability to evaluate similarity of observed states of an external environment belongs to the class of basic cognitive abilities of intelligent agents. Effective strategies for determining such similarities influence very strongly the level of behavioral rationality of agents. Similarity measures are needed in various situations and for various purposes e.g. when actions are planed or results of actions are verified [7] and previous experiences are grouped [1,2]. In case of applied agent systems it very often happens that adequate similarity measures need to imitate human approaches to comparing empirical experiences. This requirement is especially important when artificial agents are designed in order to substitute human users in dangerous, harmful and inaccessible environments.

In this work the issue of designing appropriate similarity measures for effective comparison of made observations has been defined for a particular class of artificial cognitive agents [3]. This class of agents had been previously used to solve some interesting problems of inconsistent knowledge processing [3, 4, 5] and symbol grounding [1,2], where the concept of similarity appeared to be very needed and very important. However, in none of these works additional study was carried out into criteria of choosing appropriate similarity function between individual empirical experiences of artificial agents. Therefore, this problem has been discussed and studied in detail in an unpublished work [6]. The most important results from that research are presented below.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 693–702, 2008.

The following definitions are formulated for the above mentioned class of artificial agents and their knowledge base [1,2,6]:

1. Agents are equipped with strictly private knowledge bases extracted from collections of previously made and internally stored observations of an external world.
2. Agents are located in a dynamic world built from a given set Ω of atom objects $o_1,o_2,...,o_N$. Each object $o_1,o_2,...,o_N$ can exhibit particular properties from a given set $\Delta=\{P_1,P_2,..,P_K\}$. For each property $P\in\Delta$, each object $o\in\Omega$ and each time point t, the object o exhibits or does not exhibit the property P.
3. Agents can make at least partial observations of their external world. Each observation is related to a particular time point t and is represented in the so called base profile $BP(t)$. The structure of a t-related base profile $PB(t)$ is defined as the following structure.

$$BP(t) = <\Omega,P^{+}_{1}(t),P^{-}_{1}(t),P^{+}_{2}(t),P^{-}_{2}(t),...P^{+}_{K}(t),P^{-}_{K}(t) > \quad (1)$$

where:

a) t denotes a time point to which the base profile $BP(t)$ is related,
b) $P^{+}_{i}(t)\subseteq\Omega$ and for each object $o\in\Omega$ the condition $o\in P^{+}_{i}(t)$ holds if and only if the agent perceived o as exhibiting the property P_i at the time point t.
c) $P^{-}_{i}(t)\subseteq\Omega$ and for each object $o\in\Omega$ the condition $o\in P^{-}_{i}(t)$ holds if and only if the agent perceived o as not exhibiting the property P_i at the time point t.

4. Each knowledge state $KS(t)$ of an agent is always related to a particular time point $t\in T$ and is defined by the collection of all base profiles:

$$KS(t) =\{BP(l): l\in T \text{ and } l\leq^{TM}t\}. \quad (2)$$

As it has already been mentioned this relatively simple model for agents and their experience originated knowledge base has proved to be sufficient for defining at least some interesting problems of inconsistent knowledge processing and symbol grounding. *The major problem discussed in this work is how to compare two individual base profiles (two observations) in the way that preserves particular expectations defined for the same process by human agents.*

2 General Criteria for Measuring Similarity of Observations

The detailed analysis of natural strategies to perceive similarities between states of the same environments has resulted in the following general conclusion [6]: In order to design an appropriate application-oriented similarity measure one need to choose from two different conceptual levels on which this similarity should be perceived and defined. The first level can be called a micro-level of Ω and it requires a reference to states of given properties from Δ in each individual objects from Ω. The second level can be called a macro-level of Ω and contrary to the previous approach, it should involve generalized statistics and distributions of particular properties occurrence over all objects from Ω. This important and very natural difference in perceiving similarities and differences of world's states is similar to an original definition for micro and macro structures of universe used in the theory of choice and consensus [5].

The above general conclusion has led to the following descriptions of strategies for similarity measurement and further resulted in strict definitions for sets of requirements that constrain classes of appropriate similarity measures.

Strategy for measuring micro-similarity without a constrained set of properties. In this strategy it has been assumed that for each pair of base profiles $PB(t_1)$ and $PB(t_2)$ their similarity should result from and only from all individual differences between t_1-related and t_2-related distributions of all properties from Δ, verified for the same individual objects from Ω. These differences should take into account microstructures and microstates of individual objects in both time points t_1 and t_2, and then be generalized into one summarizing evaluation of overall difference of base profiles $PB(t_1)$ and $PB(t_2)$ (or similarity of these base profiles, what is equivalent).

Strategy for measuring micro-similarity with a constrained set of properties. This strategy is the above mentioned strategy for determining micro-similarity without a constrained set of properties, provided that only some of properties from Δ are applied.

Strategy for measuring macro-similarity without a constrained set of properties. In this strategy differences between states of individual objects in time points t_1 and t_2 are neglected and general tendencies in distributions of properties from Δ over all objects Ω are considered. In particular, for each property P the highest similarity of two observations is achieved if for each individual property P all objects in Ω exhibit it or do not exhibit it at all. In all other cases similarity decreases (which differentiates this strategy to micro similarity strategies).

Strategy for measuring macro-similarity with a constrained set of properties. This strategy is the above introduced strategy for determining macro similarity without a constrained set of properties, provided that only some of properties from Δ are applied.

3 Examples of Similarity Functions

In [6] the above approaches to measuring similarity between observations were discussed in detail and for each of them classes of appropriate similarity functions were proposed. Let the universe of all base profiles defined for a given set of objects Ω and a given set of properties Δ be denoted by $\mathfrak{P}$. The general similarity function is defined for all elements of Cartesian product $\mathfrak{P}\times\mathfrak{P}$.

3.1 Micro-Similarity Determination without a Constrained Set of Properties

Let two base profiles $BP(t_1)$ and $BP(t_2)$ be given:

$$\begin{aligned} BP(t_1) &= \left\langle \Omega, P^+_{1,1}, P^-_{1,1}, P^+_{1,2}, P^-_{1,2}, \ldots, P^+_{1,K}, P^-_{1,K} \right\rangle, \\ BP(t_2) &= \left\langle \Omega, P^+_{2,1}, P^-_{2,1}, P^+_{2,2}, P^-_{2,2}, \ldots, P^+_{2,K}, P^-_{2,K} \right\rangle \end{aligned} \tag{3}$$

An example similarity function that fulfills requirements accepted for this approach to observations' comparison can be defined as follows:

$$sim(BP(t_1),BP(t_2))=\frac{1}{2K}\sum_{i=1}^{K}\left(\frac{card\left(P_{1,i}^{+}\cap P_{2,i}^{+}\right)}{card\left(P_{1,i}^{+}\right)}+\frac{card\left(P_{1,i}^{+}\setminus\left(P_{2,i}^{+}\cup P_{2,i}^{-}\right)\right)}{2\cdot card\left(P_{1,i}^{+}\right)}+\right.$$
$$\left.+\frac{card\left(P_{1,i}^{-}\cap P_{2,i}^{-}\right)}{card\left(P_{1,i}^{-}\right)}+\frac{card\left(P_{1,i}^{-}\setminus\left(P_{2,i}^{+}\cup P_{2,i}^{-}\right)\right)}{2\cdot card\left(P_{1,i}^{-}\right)}\right), \quad (4)$$

where *card*(*A*) denotes cardinality of A.

3.2 Micro-Similarity Determination with a Constrained Set of Properties

Let two base profiles from 3.1 be given. Let a subset of property indexes $Ppt\subseteq\{1,2,...,K\}$ be given. In this case an example similarity function can be given as follows:

$$sim(BP(t_1),BP(t_2))=\frac{1}{2\cdot card(Ppt)}\sum_{i\in Ppt}\left(\frac{card\left(P_{1,i}^{+}\cap P_{2,i}^{+}\right)}{card\left(P_{1,i}^{+}\right)}+\right.$$
$$\left.+\frac{card\left(P_{1,i}^{+}\setminus\left(P_{2,i}^{+}\cup P_{2,i}^{-}\right)\right)}{2\cdot card\left(P_{1,i}^{+}\right)}+\frac{card\left(P_{1,i}^{-}\cap P_{2,i}^{-}\right)}{card\left(P_{1,i}^{-}\right)}+\frac{card\left(P_{1,i}^{-}\setminus\left(P_{2,i}^{+}\cup P_{2,i}^{-}\right)\right)}{2\cdot card\left(P_{1,i}^{-}\right)}\right). \quad (5)$$

This function is similar to the previous one and the additional value $2\cdot card(Ppt)$ is used to normalize its value.

3.3 Macro-Similarity Determination without a Constrained Set of Properties

In order to design an appropriate mechanism for determining macro similarity without a constrained set of properties one needs to answer two additional questions:

- How to describe in an effective way the actual tendency of distributions for a given property *P*?
- How to determine in an effective way the similarity of actual tendencies of distributions for given two time points t_1 and t_2?

Tendency of distribution of a certain property *P* is described in a numerical way as a number from an interval $[0,1]$. Value 0.5 reflects no tendency. As the value rises over 0.5, it reflects growing positive tendency (i.e. tendency for objects to exhibit corresponding property). At last, value 1 represents completely positive tendency. By analogy, values under 0.5 reflect growing negative tendency (i.e. tendency for objects not to exhibit corresponding property) and value 0 reflects completely negative tendency. In work [6] the following commonsense requirements were formulated to constrain the way the tendency correlated with a single property *P* should be evaluated:

Requirement 1
If all objects perceived at the moment t in terms of exhibiting property P exhibit this property, there is general positive tendency and it is reflected by a value of 1.

Requirement 2
If all objects perceived at the moment t in terms of exhibiting property P do not exhibit this property, there is general negative tendency and it is reflected by a value of 0.

Requirement 3
If no objects are perceived at the moment t in terms of exhibiting property P, there is no general tendency and it is reflected by a value of 0.5.

Requirement 4
Value of tendency function rises with rise of the number of objects perceived at the moment t as exhibiting property P.

Requirement 5
Value of tendency function drops with rise of the number of objects perceived at the moment t as not exhibiting property P.

Let $tend_i(t)$ denote the tendency of property P_i occurrence in a base profile $BP(t)$. This function can be given by the following formula:

$$tend_i(t) = \begin{cases} \dfrac{card\left(P_i^+(t)\right)}{card\left(P_i^+(t) \cup P_i^-(t)\right)} & when \quad card\left(P_i^+(t) \cup P_i^-(t)\right) > 0 \\ \dfrac{1}{2} & otherwise \end{cases} . \tag{6}$$

It can be proved that this function fulfills above listed requirements.

Another set of commonsense requirements was proposed in [6] to constrain a way in which the similarity with an use of tendency measures should be evaluated:

Requirement 1
If there are general tendencies (positive or negative) in an exhibition of a certain property and they are the same in both moments of time, value of similarity rises.

Requirement 2
If there are general tendencies (positive or negative) correlated respectively with every single property and they are the same in both moments of time, value of similarity is maximal and equal to 1.

Requirement 3
If there are general tendencies (positive or negative) in an exhibition of a certain property and they are opposite in corresponding moments of time, value of similarity drops.

Requirement 4
If there are general tendencies (positive or negative) correlated respectively with every single property and they are opposite in corresponding moments of time, value of similarity is minimal and equal to 0.

Requirement 5
If there is no tendency (positive or negative) in an exhibition of a certain property (i.e. no object are perceived in terms of exhibiting the property OR some objects are perceived as exhibiting the property, but about the same number of objects are perceived as not exhibiting the property), value of similarity drops.

Requirement 6
If there is no general tendency correlated with at least one property, the similarity is equal to 0, because there are no rational premises at the macro-level to deduct any correlation. In this approach it is stated, that two balanced states (i.e. with no tendencies) are completely different.

In consequence a related function for measuring macro similarity without a constrained set of properties can be given as follows:

$$sim(BP(t_1), BP(t_2)) = \frac{1}{K}\sum_{i=1}^{K}\left[tend_i(t_1) + tend_i(t_2) - 1\right]^2 . \quad (7)$$

3.4 Macro-Similarity Determination with a Constrained Set of Properties

An appropriate function for determination of macro-similarity with a constrained set of properties is defined in the similar way, provided that it is applied to a particular subset of the given set of properties. Let $Ppt \subseteq \{1,2,...,K\}$ be a subset of properties indexes. Corresponding function is defined as follows:

$$sim(BP(t_1), BP(t_2)) = \frac{1}{card(Ppt)} \cdot \sum_{i \in Ppt}\left[tend_i(t_1) + tend_i(t_2) - 1\right]^2 . \quad (8)$$

4 Integration of Strategies

One of the major practical problems that has to be solved in implementations of artificial cognitive agents, was to develop an integrated computational mechanism in which the above four approaches to measuring similarities could be effectively combined to fulfill specific requirements tailored to designed applications [6]. In order to achieve this desirable flexibility of integration n additional decomposition of particular similarity computation was proposed.

Namely, it has been proposed that each computation of similarity measure was divided into two separate steps:

Step1. *Computation of vector of partial similarity values at the level of single properties*. In this step, for each property P_i from the set of properties Δ, corresponding partial similarity value $simL_i$ is evaluated. It can be treated as a local similarity. All these values computed for i=1,2,…,K form a vector of partial similarity values.

Step2. *Computation of total similarity value based on partial similarity values*. In this step, value sim^* of similarity of two profiles is evaluated based on the already extracted vector of partial similarity values ($simL_1, simL_2, \ldots, simL_K$). Figure 1 shows how the process of computing sim^* goes on starting from computations of particular $simL_i$.

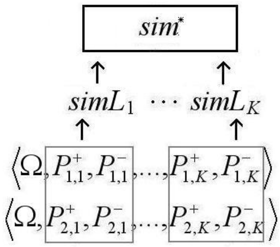


Fig. 1. Schema of decomposed process of base profiles' similarity evaluation

It means that each proposed similarity function $sim(BP(t_1),BP(t_2))$ can be realized as a superposition of two functions $simU$ and $simL$, where:

$$\begin{aligned} sim &: \wp\times\wp\to[0,1], \\ simU &: [0,1]^K\to[0,1], \\ simL &: \wp\times\wp\to[0,1]^K, \\ sim &= simU \circ simL. \end{aligned} \tag{9}$$

With an use of these notions, above mentioned steps take following form:

$$\begin{aligned} &\text{Step 1} \quad (simL_1,\ldots,simL_K) = simL(BP(t_1),BP(t_2)) \\ &\text{Step 2} \quad sim^* = simU((simL_1,\ldots,simL_K)). \end{aligned} \tag{10}$$

Let $Ppt \subseteq \{1,2,\ldots,K\}$ denote a given subset of properties' indexes and let a function $imp(i)$ be given (this function can be treated as a characteristic function of set of indexes):

$$imp(i) = \begin{cases} 1 & when \quad i \in Ppt \\ 0 & when \quad i \notin Ppt \end{cases}. \tag{11}$$

For all four strategies 3.1-3.4 function $simU$ is given by the formula:

$$sim((simL_1,\ldots,simL_K)) = \frac{1}{card(Ppt)} \sum_{i=1}^{K} imp(i) simL_i \,. \tag{12}$$

Function $simL$ is a vector function. Its coordinates are evaluated independently based on the same schema. For every $i \in \{1,2,\ldots,K\}$, corresponding coordinate $simL_i$ is given by a following parameterized formula

$$simL_i = \alpha_1\left(\frac{card\left(P_{1,i}^{+} \cap P_{2,i}^{+}\right)}{card\left(P_{1,i}^{+}\right)} + \alpha_2\frac{card\left(P_{1,i}^{+} \setminus \left(P_{2,i}^{+} \cup P_{2,i}^{-}\right)\right)}{card\left(P_{1,i}^{+}\right)} + \frac{card\left(P_{1,i}^{-} \cap P_{2,i}^{-}\right)}{card\left(P_{1,i}^{-}\right)} + \right.$$
$$\left. + \alpha_2\frac{card\left(P_{1,i}^{-} \setminus \left(P_{2,i}^{+} \cup P_{2,i}^{-}\right)\right)}{card\left(P_{1,i}^{-}\right)}\right) + \alpha_3\left(tend_i(t_1) + tend_i(t_2) - 1\right)^2 + \tag{13}$$
$$+ \alpha_4\left|tend_i(t_1) - tend_i(t_2)\right|.$$

Parameters are used in an above formula to allow adaptation to proposed strategies. Parameters α_1, α_2, α_3, α_4 should be interpreted as follows

- α_1 – corresponds to a degree in which the micro-similarity approach is taken into account in a process of similarity evaluation. Growth of this parameter reflects growth of an attention paid to this approach.
- α_2 – represents an interpretation of uncertainty. Low value of α_2 refers to a pessimistic attitude (e.g. when α_1=0 and information is missing – i.e. some properties have not been observed – the worst case scenario is assumed). Analogically high value of this parameter reflects an optimistic attitude toward unobserved facts. It is wider discussed in [6], where complete sub-dimension of uncertainty is proposed and analyzed.
- α_3 – corresponds to a degree in which the macro-similarity approach is taken into account in a process of similarity evaluation. Growth of this parameter reflects growth of an attention paid to this approach.
- α_4 – this parameter is not directly correlated with any of approaches mentioned in this paper and it is correlated with an modified, alternative version of macro-similarity approach. The tendency as described earlier is understood as an existence of a positive or a negative tendency. But it can also be understood in more probabilistic approach as a degree in which certain properties are distributed. In this modified approach, values close to 0.5 do not describe the lack of tendency but a certain level of property distribution. It is not described in detail in this paper, as the emphasis is put on the distinction between macro- and micro-approach, but can be used in further tests and application, so it is included in an integrated formula.

In addition, Values of parameters α_1, α_2, α_3, α_4, have to fulfill following claims

$$\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in [0,1], \quad \alpha_1 + \alpha_3 + \alpha_4 = 1. \tag{14}$$

Parametrization correlated with proposed approaches

Micro-Similarity Determination without a Constrained Set of Properties is modeled with a following set of parameters:

$$\alpha_1=1,\ \alpha_2=0.5,\ \alpha_3=0,\ \alpha_4=0, \quad Ppt=\{1,2,...,K\} \tag{15}$$

Micro-Similarity Determination with a Constrained Set of Properties is modeled with a following set of parameters:

$$\alpha_1=1,\ \alpha_2=0.5,\ \alpha_3=0,\ \alpha_4=0, \quad Ppt \text{ – defined in the approach} \tag{16}$$

Macro-Similarity Determination without a Constrained Set of Properties is modeled with a following set of parameters:

$$\alpha_1=0,\ \alpha_2=0,\ \alpha_3=1,\ \alpha_4=0, \quad Ppt=\{1,2,...,K\} \tag{17}$$

Macro-Similarity Determination with a Constrained Set of Properties is modeled with a following set of parameters:

$$\alpha_1=0,\ \alpha_2=0,\ \alpha_3=1,\ \alpha_4=0, \quad Ppt \text{ – defined in an approach} \tag{18}$$

It is easy to notice, that proposed parametrical model allows other sets of parameters' values. Therefore, combined approaches can be modeled and tested in final implementations. It is important task to choose values well suited for the modeled approach. Parameters can be either set by an expert or optimized through an experiment.

5 Final Remarks

Four approaches to measuring similarity of made observations have been proposed. These strategies have been defined for a particular model of empirical knowledge base. This knowledge base has been assumed to consist of individual observations that have ever been made by an artificial cognitive agent. These models for knowledge base and for the related artificial cognitive agent have already been used in other works to model some problems in inconsistent knowledge processing [3, 4, 5] and symbol grounding [1, 2]. In both of these applications an effective methods for similarity determination between two observations have been required, however the choice of applied similarity functions was rather arbitrary. The four strategies have been proposed and studied to resolve this situation.

This flexible mechanism for computing similarity measures will be further applied in implemented agent systems to provide them of desirable similarity evaluation strategies [6]. Further experiments are planned in which actual behavior of strategies will be tested.

References

1. Katarzyniak, R.: The language grounding problem and its relation to the internal structure of cognitive agents. Journal of Universal Computer Science 11(2), 357–374 (2005)
2. Katarzyniak, R.: Grounding a modal language of communication in artificial agents, EXIT Warsaw (2007)

3. Katarzyniak, R., Nguyen, T.N.: Reconciling inconsistent profiles of agent’s knowledge states in distributed multiagent systems using consensus methods. Systems Science 26(4), 93–119 (2000)
4. Nguyen, N.T.: Using distance functions to solve representation choice problem. Fundamenta Informaticae 48(4), 295–314 (2001)
5. Nguyen, N.T.: Consensus system for conflict solving in distributed systems. Information Sciences 147(1-4), 91–122 (2002)
6. Popek, G.: Strategies for determination of grounding sets for the language of modal formulas. MSc Thesis, Wrocław University of Technology, Faculty of Computer Science and Management, Wrocław (in Polish) (2007)
7. Popek, G.: An Action Planning Module Based on Vague Knowledge Extracted from Past Experiences. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 997–1003. Springer, Heidelberg (2006)

An Algorithm for Agent Knowledge Integration Using Conjunctive and Disjunctive Structures

Trong Hieu Tran and Ngoc Thanh Nguyen

Institute of Information Science and Engineering,
Wroclaw University of Technology, Poland
trong.hieu.tran@pwr.wroc.pl, thanh@pwr.wroc.pl

Abstract: In this paper an algorithm for integration of knowledge in multi-agent environments using logic disjunction and conjunction structures, is presented and analysed. The authors propose a distance function between the logic formulae and define a set of postulates (criteria) for integration. The worked out algorithm is analysed regarding the computation complexity and the proposed postulates.

Keywords: knowledge integration, consensus theory, multi-agent system.

1 Introduction

In multi-agent systems knowledge integration is very often needed. Knowledge integration is necessary when agents want to realize a common task, or when they want to merge their knowledge bases to create a new base for a new agent. Most often, if there is a higher level agent in the system who manages the activities of a group of lower level agents, then for making a total view of some subject this agent has to integrate the knowledge originating from the agents it takes care of. As an example, consider a multi-agent meta search engine [10] in which for a user query the managing agent gathers answers from so-called searching agents, next it integrates these answers and determines the final answer for the user.

In the Knowledge Management literature there are several known approaches for knowledge integration. However, because of the lack of space we will not analyze them deeply here. Each knowledge integration method consists of at least two elements: The first is the definition of the structure of knowledge and the second is a set of postulates (criteria) which set the purpose of integration process. One of the above-mentioned approaches is based on consensus theory [9] which is very useful for processing inconsistency of knowledge originating from different autonomous sources, such as agents or experts. Most often, a general criterion for a consensus-based knowledge integration method is based on the requirement that the result of integration should best represent the states of knowledge which are to be integrated. The method for knowledge integration presented in this paper is also consensus-based, and as the knowledge structure the authors use logic conjunctions and disjunctions. This kind of structures has been analyzed in work [9]. They are very useful for building agent knowledge bases. Although a lot of work has been done for

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 703–712, 2008.

this kind of structures, the original contribution of this work is based on defining a new distance function between the logic conjunctions (or disjunctions), and on working out a new and effective algorithm for integration.

The organization of the remaining part of this paper is presented as follows. In Section 2 the basic notions are introduced. Section 3 gives a brief view of conflicts in multi-agent systems. Section 4 contains the definition of the distance function between logic conjunctions (or disjunctions). In Section 5 the consensus-based integration problem is defined. Section 6 presents postulates for knowledge integration and their analysis. In Section 7 the integration algorithm is presented and analyzed. At last, some conclusions are included in Section 8.

2 Basic Notions

In this work we assume that in order to represent the knowledge an agent uses a finite set L of symbols for representing the truth of facts and events in the real world the agent acts. A symbol form L is called also an *atom*. A *literal* is defined as an expression a or $\neg a$ where a is an atom. A literal without symbol "¬" is called a *positive literal*, otherwise it is called a *negative literal*.

A conjunctive formula of literals (*conjunction* in short) is a logic formula which is represented by the following expression:

$$t_1 \wedge t_2 \wedge \ldots \wedge t_k$$

where t_i is a literal (positive or negative) for $i = 1, 2, \ldots, k$.

Similarly, a disjunctive formula of literals (*disjunction* in short) is a logic formula which represented by the following expression:

$$t_1 \vee t_2 \vee \ldots \vee t_k$$

where t_i is a literal (positive or negative) for $i = 1, 2, \ldots, k$.

In a conjunction or disjunction we assume that there is no repetition or inconsistency among literals. It means that each atom occurs at most once in a formula.

By $Conj(L)$ and $Clause(L)$ we denote the set of all conjunctions and the set of all disjunctions with symbols from set L, respectively.

Because of the limitation of set L and referring to the above restrictions, sets $Conj(L)$ and $Clause(L)$ are also finite.

Notice that a logic formula x can be divided in two sets of symbols: the first consists of the symbols creating positive literals, and the other consists of the remaining symbols. Thus a formula x can be represented by a pair (x^+, x^-) in which x^+ is the set of symbols creating positive literals appearing in x and x^- is the set of symbols occurring in negative literals.

Example 1

For $a, b, c, d \in L$,

If $x = a \wedge \neg b \wedge \neg c \wedge d$, we have $x = (x^+, x^-)$ in which $x^+ = \{a, d\}$ and $x^- = \{b, c\}$.

If $y = a \vee b \vee c$, we have $y = (y^+, y^-)$ in which $y^+ = \{a, b, c\}$ and $y^- = \emptyset$.

For convenience, in the next sections we denote *Form*(*L*) as one of sets *Conj*(*L*) and *Clause*(*L*).

3 Conflict in Multi–agents Systems

In a multi-agent system the simplest conflict takes place when two agents have different opinions on the same subject [8]. According to Pawlak [11], a conflict is specified by following elements: a set of agents, a set of issues, and a set of opinions of agents about these issues. The agents refer to the issues in a concrete context. Each agent can present his opinion regarding an issue using symbols (+) as "yes", (–) as "no", and (0) – if it is neutral. Thus we can state that a conflict should take place if there are at least two agents whose opinions on some issue differ from each other.

In general, a conflict includes the following main components [8]:

Conflict body: specifies the direct participants of the conflict.
Conflict subject: specifies the issues to which the conflict refers.
Conflict content: specifies the opinions of the participants on the conflict subject.

Referring to Pawlak's approach, the body of conflict is the set of agents who participate in conflict; the conflict subject refers to a set of dissenting issues; and the content is a collection of tuples representing the opinions of the participants.

Example 2
We have a conflict in which conflict body is a set of agents $\{a_1, a_2, a_3, a_4, a_5\}$ who dissent about a set of issues $\{i_1, i_2, i_3, i_4\}$ as conflict subject and the conflict content is given by the following information table:

U	i_1	i_2	i_3	i_4
a_1	–	+	0	0
a_2	+	–	0	–
a_3	0	+	0	+
a_4	+	0	–	0
a_5	0	+	+	–

Within a conflict we can determine several conflict profiles. A conflict profile is a multi-set (i.e. with repetitions) of opinions which are generated by the agents regarding an issue. In the above example, we have four conflict profiles which are $P_1 = \{–, +, 0, +, 0\}$, $P_2 = \{+, –, +, 0, +\}$, $P_3 = \{0, 0, 0, –, +\}$, and $P_4 = \{0, –, +, 0, –\}$. Information system theory [11, 12] has been proved to be very suitable for managing this kind of conflicts.

In [8] the author expanded the definition of conflict to treat with distributed systems, in which the values of attributes representing conflict content are not only elementary values but sets of them. Moreover, conflict content is partitioned into three groups, the first includes opinions of type "*The fact should take place*", the second includes opinions of type "*The fact could not take place*", and the last includes opinions of type "*I do not know if the fact takes place*". These groups represent three typed of knowledge: *positive*, *negative*, and *ignorance* [6].

There are several methods for resolving the conflict problems, these methods are presented in [4, 8, 9]. In this paper we define conflict in the similar way. We assume that a conflict profiles is a multi-set of logic formulae and we will use consensus methods to resolve the conflict.

4 Distance Functions Conjunctions and Disjunctions

Distance functions are generally understood as tools which allow measuring the difference between objects in the same space. Distance functions have been applied in many fields of sciences, for example Mathematics, Physics, and Biology. In the field of Knowledge Management distance functions have some significant applications. For instance, they are used to define the macrostructure of a universe in consensus problems [8] and to determine the median of elements and the distance between semilattices [7], to measure the inconsistency in knowledge bases [5], or the distance between equivalence relations [3].

In this paper, we generally define a distance function between logic formulas as follows:

Definition 1
By a distance function between formulae from set Form(L) we understand the following function:

$$d: Form(L) \times Form(L) \rightarrow R^{+}$$

where R^+ denotes the set of non–negative real numbers.

In [13] the authors defined some intuitive conditions for distance functions between conjunctions and disjunctions, such as *Translation*, *Reciprocity*, *Degrees of dissent*, and *Maximum value*. The relationships between these conditions have been analyzed and some distance functions have been proposed. It has also been stated in the mentioned work that distance functions for conjunctions are an important tool serving to find out the resolution for conflict situations in knowledge integration processes.

In this paper we define a general distance function for conjunctions and disjunctions which uses the sets of occurrences of literals.

Definition 2 *For $\alpha \in L$ and $x \in Form(L)$ let*

$$occur(\alpha, x) = \begin{cases} -1 & iff \quad \alpha \in x^{*-} \\ 0 & iff \quad \alpha \notin x^{*+} \cup x^{*-} \\ 1 & iff \quad \alpha \in x^{*+} \end{cases}$$

Example 3. For $\alpha, \beta, \gamma, \delta \in L$.
Let $x = \alpha \wedge \neg\beta \wedge \gamma$ we have

$$occur(\alpha, x) = occur(\gamma, x) = 1, \; occur(\beta, x) = -1, \text{ and } occur(\delta, x) = 0.$$

Let $y = \neg\alpha \vee \beta$ we have

$$occur(\alpha, y) = 1,\ occur(\beta, y) = -1\ \text{, and}\ occur(\gamma, y) = occur(\delta, y) = 0.$$

Function *occur* shows clearly the status (as positive or negative, or non-occurring) of a literal in a formula.

We define the distance function between logic formulae as follows:

Definition 3. *Let* $x, y \in Form(L)$

$$d(x, y) = \sqrt{\sum_{\alpha \in L}(occur(\alpha, x) - occur(\alpha, y))^2}$$

Obviously, the distance function d is a metric. Moreover, as it has been shown in [13], if *Form*(*L*) is *Conj*(*L*) then this distance function satisfies all postulates of *Translation*, *Reciprocity*, *Degrees of dissent*, and *Maximum value.*

5 Integration Based on Consensus Technique

In general, consensus theory deals with problems of data analysis in order to extract valuable information. Consensus methods enable to determine a version of data, which, at first, should best represent a set of previously given data, and, at second, should be a good compromise acceptable for parties that are in conflict because of their authorship of the original data.

The consensus problem for resolving inconsistency of a set of opinions should be formulated as follows: *Given n opinions* $O_1, O_2, \ldots, O_n$, *one should determine one which could best represent these ones.*

The chosen opinion is called a *consensus* of the given opinions.

The consensus has to satisfy following general conditions:

– It should not much differ from the given opinions, and
– The differences between the consensus and the given opinions should not much differ from each other.

Formally, assuming U as a countable universe, by U_0 we denote the set of all nonempty, finite subsets with repetitions of U. For given set $X \in U_0$ (called a conflict profile) if we have some choice rule, we can choose all those elements of U which satisfy this rule, then we should have a consensus choice function:

$$c : U_0 \to 2^U.$$

An element of set $c(X)$ is called a *consensus* of set X.

Consensus technique is as a powerful tool used in the alternatives ranking problem [1, 14] or the committee election problem [2].

In this paper, the consensus technique is used to determine the knowledge integration of agents, in which the knowledge of each agent is represented as a conjunction or disjunction. We assume that a conflict profile is a multi-set being a subset of *Conj*(*L*) or *Clause*(*L*). We define the following integration task:

For a given conflict profile of conjunctions (or disjunctions)

$$X = \{x_i = (x_i^+, x_i^-) \in Form(L): i = 1, 2,\ldots,n\}$$

it is needed to determine a conjunction (or disjunction) $x^* \in Form(L)$ which best represents the given conjunctions (or disjunctions).

This problem has been formulated in [9]. In this paper we propose a new algorithm for its solution.

6 Postulates for Knowledge Integration

In this section we introduce some postulates which are used to integrate knowledge based on consensus technique. In what follows we denote $\Pi(Z)$ as the set of all finite subsets with repetitions of set Z.

Definition 4. *By a consensus function for profiles of logic formulae we understand a function:*

$$C: \Pi(Form(L)) \rightarrow 2^{Form(L)}$$

which satisfies one or more of the following postulates:
P1. For each formula $(x^{*+}, x^{*-}) \in C(X)$ there should be:

a) $\bigcap_{x\in X} x^+ \subseteq x^{*+}$

and

b) $\bigcap_{x\in X} x^- \subseteq x^{*-}$.

P2. For each formula $(x^{*+}, x^{*-}) \in C(X)$ there should be:

a) $x^{*+} \subseteq \bigcup_{x\in X} x^+$

and

b) $x^{*-} \subseteq \bigcup_{x\in X} x^-$.

P3. For each symbol $z \in L$, if the form of appearance (that is as a positive or negative literal) of z occurs only once in X, this form of appearance of z should be occur in $x^* \in C(X)$.

P4. A consensus $x^* \in C(X)$ should minimize the sum of distances:

$$\sum_{x\in X} d\ (x^*, x) = \min_{x'\in Form(L)} \sum_{x\in X} d\ (x', x) \cdot$$

By C_{co} we denote the set of all consensus functions determined by Definition 4.

Some commentary for postulates is given as follows:

- Postulates $P1$a and P1b [9] mean that the common part of profiles should be included in the component of consensus for positive profiles and negative profiles respectively. The sense of this postulate follows from the Pareto criterion [14]: If all voters vote for the same candidate then he should be finally chosen. These conditions are very intuitive, because they fit for a very common principle stated that if all agents (or experts) consent some event should take place (or should not take place), then the consensus should take this fact into account.

- The main idea of Postulates *P*2a and *P*2b [9] is based on the requirement that the consensus should not exceed the profiles. This means that the positive and negative components of consensus should be included in the sums of positive profiles and negative profiles, respectively. These postulates come from the rule of the closed world: Objects which do not belong to the world, do not exist.
- Postulate *P*3 implies the superiority of knowledge, it states that if there is only one agent (or expert) has an opinion on an issue, this opinion should preserve in consensus.
- Postulates *P*4 [9] is a very popular criterion for consensus. This criterion is natural for satisfying the condition for "the best representation".

7 Analysis of Postulates and Integration Algorithm

In this section we present several properties of postulates and the relationships between consensus functions satisfying them.

The fact that a consensus function *C* satisfies a postulate *P* for a given profile *X* will be written as:

$$C(X) \mid\!- P.$$

The fact that a consensus function *C* satisfies a postulate *P* (that is the postulate is satisfied for all arguments of the function) will be written as:

$$C \mid\!- P.$$

At last, if a postulate *P* is satisfied for all functions from C_{co} then we will write:

$$C_{co} \mid\!- P.$$

We have the following properties of consensus functions:

Theorem 1
A consensus function which satisfies postulate P4 should also satisfy postulates P1 and P2, that is

$$(C \mid\!- P4) \Rightarrow ((C \mid\!- P1a) \wedge (C \mid\!- P1b) \wedge (C \mid\!- P2a) \wedge (C \mid\!- P2b))$$

for each $C \in C_{co}$.

Theorem 2
The following dependencies are true:

$$(\exists C \in C_{co})(\exists X \in \Pi(Conj(L))\text{:}$$
$$((C \mid\!- P1a) \wedge (C \mid\!- P1b) \wedge (C \mid\!- P2a) \wedge (C \mid\!- P2b) \wedge C(X) \mid\!- P4)$$

Proposition 1

Let $\alpha, \beta \in L$, $X \in \Pi(Form(L))$, $y=(y^+, y^-) \in Form(L)$, and $occur(\alpha, y) = occur(\beta, y) = 0$ if $|\sum_{x\in X} occur(\alpha, x)| > |\sum_{x\in X} occur(\beta, x)|$, we have

$$\min(\sum_{x\in X} d((y^+ \cup \alpha, y^-), x), \sum_{x\in X} d((y^+, y^- \cup \alpha), x)) \leq \min(\sum_{x\in X} d((y^+ \cup \beta, y^-), x), \sum_{x\in X} d((y^+, y^- \cup \beta), x))$$

Below we propose an algorithm for knowledge integration using logic formulae structures. The idea of this algorithm is based on the following steps: Firstly, we collect all literals occurred in all formulae, of course the result formula built from these literals satisfies postulates P1a, P1b, P2a, P2b and P4 (according to Proposition 1, because the number of occurrences of these literals is maximal). In the next step, we take the literal from the remaining set of literals occurred in all formulae such that the difference between the number of occurrences in the positive form and the negative form is the largest. We examine this literal if the condition of postulate P4 is satisfied. If so, we add this literal into the consensus. This step is repeated until not all the literals occurred in given formulae are examined.

The detail of algorithm is proposed as presented as follows.

Algorithm 1

Input: Profile $X \in \Pi(Form(L))$, where $X = \{(x_i^+, x_i^-): i = 1,2,\ldots,n\}$

Output: Consensus (x^{*+}, x^{*-}) satisfying postulates $P1$a, $P1$b, $P2$a, $P2$b, and $P4$.

Procedure:

BEGIN

1. Set $x^{*+} = \bigcap_{x\in X} x^+$;
2. Set $x^{*-} = \bigcap_{x\in X} x^-$;
3. Set $Z := (\{x_i^+: i = 1, 2,\ldots, n\} \setminus x^{*+}) \cup (\{x_i^-: i = 1, 2,\ldots, n\} \setminus x^{*-})$;
4. For each element $z \in Z$ calculate $f^+(z)$ as the number of occurrences of z^+ and $f^-(z)$ as the number of occurrences of $^-$ in sets belonging to profile X;
5. Calculate $S^* = \sum_{x\in X} d(x^*, x)$
6. While $Z \neq \emptyset$ do

   Begin

   6.1. Find $z \in Z$ such that $|f^+(z) - f^-(z)|$ is maximal;

   6.2. Set $Z := Z \setminus \{z\}$;

   6.3. Set $x'^+ := x^{*+} \cup \{z\}$, $x'^- = x^{*-}$;

   6.4. Calculate $S(x') = \sum_{x\in X} d(x', x)$,

   6.5. If $S(x') \leq S^*$ then set $S^* := S(x')$ and $x^{*+} := x'^+$;

   6.6. Set $x''^+ := x^{*+}$, $x''^- = x^{*-} \cup \{z\}$;

   6.7. Calculate $S(x'') = \sum_{x\in X} d(x'', x)$,

   6.8. If $S(x'') \leq S^*$ then set $S^* := S(x'')$, $x^{*-} := x''^-$ and $x^{*+} := x^{*+} \setminus \{z\}$;

   End

END.

It is easy to prove that the computation complexity of algorithm 1 is $O(m \cdot n)$ where $m = card(Z)$ and $n = card(X)$.

This algorithm satisfies postulates P1a, P1b, P2a, P2b, and P4. Indeed, the first and second steps ensure that the conditions of postulates P1a and P1b are fulfilled. Moreover, the conditions of postulates P2a and P2b are also satisfied because we have

$$\bigcap_{x\in X} x^{+} \subseteq \bigcup_{x\in X} x^{+} \text{ and } \bigcap_{x\in X} x^{-} \subseteq \bigcup_{x\in X} x^{-} .$$

The third step sets the scope of examined literals from all literals occurred in the given formulae. Thus it guarantees the result of the next steps to not exceed the scope. Owing to this the last result will always satisfy the conditions of postulates P2a and P2b.

The last steps enroll the remaining literals into the result formula step by step based on Proposition 1 and the condition of postulate 4.

As the result of these steps, our algorithm satisfies all given postulates.

Example 4

Let's consider a real world consisting of potential symptoms for a new disease. Let there be a set of doctors:

$$Agent = \{d_1, d_2, d_3, d_4\},$$

and a set of symbols:

$$\boldsymbol{L} = \{t_1, t_2, t_3, t_4\},$$

which represent the truth of the following facts:

- t_1: High temperature;
- t_2: Hack;
- t_3: Runny nose;
- t_4: High pressure;
- t_5: Languor.

The doctors are asked for the characteristic symptoms for the new disease. For this common subject they generate the following set of knowledge states:

Agent	*Knowledge state*
d_1	$\neg t_1 \wedge \neg t_2 \wedge t_4 \wedge t_5$
d_2	$\neg t_1 \wedge t_2 \wedge t_3 \wedge t_4 \wedge \neg t_5$
d_3	$\neg t_1 \wedge t_2 \wedge \neg t_3 \wedge t_4$
d_4	$\neg t_1 \wedge t_2 \wedge t_3 \wedge t_4$

Let $x^* = (x^{*+}, x^{*-})$ be the integration knowledge of these doctors. By applying Algorithm 1 we have $x^{*+} = \{t_2, t_3, t_4\}$ and $x^{*-} = \{t_1\}$. Thus, we have the common symptoms of this disease as hack, runny nose, high pressure and normal temperature. Symptom t_5 does not appear in the integration result.

8 Conclusions

In this paper an algorithm for integration of knowledge in multi-agent environments using logic disjunction and conjunction structures has been presented and analysed. The authors propose a set of four postulates (criteria) for integration. As it has been proved, the algorithm satisfies only three of these postulates. The future works should concern the analysis of this algorithm regarding postulate P_3 and working out an algorithm satisfying all the proposed postulates.

References

[1] Arrow, K.J.: Social Choice and Individual Values. Wiley, New York (1963)

[2] Bock, H.H., Day, W.H.E., McMorris, F.R.: Consensus rules for committee elections. Mathematical Social Sciences 35, 219–232 (1998)

[3] Day, W.H.E.: The complexity of computing metric distances between partitions. Mathematical Social Science 1, 269–287 (1981)

[4] Deja, R.: Using Rough Set Theory in Conflicts Analysis, Ph.D. Thesis (Advisor: A. Skowron), Institute of Computer Science, Polish Academy of Sciences, Warsaw (2000)

[5] Grant, J., Hunter, A.: Measuring inconsistency in knowledge bases. Journal of Intelligent Information Systems 27(2), 159–184 (2006)

[6] Katarzyniak, R., Nguyen, N.T.: Reconciling inconsistent profiles of agents' knowledge states in distributed multi-agent systems using consensus methods. System Science 26(4), 93–119 (2000)

[7] McMorris, F.R., Mulder, H.M., Powers, R.C.: The median function on median graphs and semilattices. Discrete Appl. Math. 101, 221–230 (2002)

[8] Nguyen, N.T.: Consensus System for Solving Conflicts in Distributed Systems. Journal of Information Sciences 147, 91–122 (2002)

[9] Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management. Springer, London (2008)

[10] Nguyen, N.T., Gandza, M., Paprzycki, M.: A Consensus-based Multi-agent Approach for Information Retrieval in Internet. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)

[11] Pawlak, Z.: An Inquiry into Anatomy of Conflicts. Journal of Information Sciences 109, 65–78 (1998)

[12] Skowron, A., Rauszer, C.: The Discernibility Matrices and Functions in Information Systems. In: Słowinski, E. (ed.) Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory, pp. 331–362. Kluwer Academic Publishers, Dordrecht (1992)

[13] Tran, T.H., Nguyen, N.T.: Distance functions for logic conjunctions in knowledge integration tasks. In: Proceedings of the 16th International Conference on Systems Science, vol. 2, pp. 454–462 (2007)

[14] Young, H.P., Levenglick, A.: A Consistent Extension of Condorcet's Election Principle. SIAM J. App. Math. 35, 285–300 (1978)

A Note on Root Choice for Parallel Processing of Tree Decompositions

Yueping Li[1] and Yunting Lu[2]

[1] Department of Computer Science, Sun Yat-sen University
Guangzhou 510275, P.R. China
leeyueping@gmail.com
[2] Shenzhen Institute of Information Technology
Shenzhen 518029, P.R. China
lu-yunting@163.com

Abstract. This paper deals with the root choice strategy for a tree decomposition when multiple agents(processors) are deployed. Tree decomposition is one of the most important decompositions in graph theory. It not only plays a role in theoretical investigations but also has widely practical applications [4]. The first step of solving problem using tree decomposition is to choose a root. And the root choice affects the time complexity when parallel processing is employed. We propose an algorithm to determine the root which makes the latest completion time minimum. In addition, remarks and future works are given.

Keywords: tree decomposition, treewidth, parallel processing, algorithm.

1 Introduction and Terminology

Tree decomposition is usually used to solve hard problems in graphs with bounded treewidth. It appears that many problems that are intractable (e.g., NP-hard) for general graphs become polynomial or linear-time-solvable, when restricted to graphs of bounded treewidth (see [4] for an overview). That is because that provided a tree decomposition, many hard problems can be divided into the small parts of the original graph. Due to this property, dynamic programming can be adopted to tackle the problem for the whole graph when the small parts have been solved.

Graphs of bounded treewidth appear in a large number of research fields. In real world, most communication networks have small treewidth. Yamagucki, Aoki, and Mamitsuka [14] have computed the treewidth of 9712 chemical compounds from the LIGAND database, and discovered that all but one had treewidth at most three; the one exception had treewidth four. Thorup [10] showed that the control flow graph of goto-free programs, written in one of a number of common imperative programming languages (like C, Pascal) have treewidth bounded by small constants.

The concept of treewidth was introduced by Robertson and Seymour [9]. In the 1980s, several related or equivalent notions were invented independently : partial k-trees (Arnborg and Proskurowski, e.g., [2][3]), clique trees (Lauritzen and

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 713–722, 2008.

Spiegelhalter [8]), recursive graph classes (Borie [5][6]) and k-terminal recursive graph classes (Wimer [11][12]). The notions treewidth and tree decomposition have been used most widely.

Let $G = (V, E)$ be a graph. A tree decomposition TD of G is a pair (T, X), where $T = (I, F)$ is a tree, and $X = \{X_i | i \in I\}$ is a family of subsets of V, one for each node of T, such that:

1. $\bigcup_{i \in I} X_i = V$
2. for every edge $v, w \in E$, there is an $i \in I$ with $v \in X_i$ and $w \in X_i$, and
3. for all $i, j, k \in I$, if j is on the path from i to k in T, then $X_i \bigcap X_k \subseteq X_j$.

The width of a tree decomposition $((I, F), \{X_i | i \in I\})$ is $\max_{i \in I} |X_i| - 1$. The treewidth of a graph G, denoted by $tw(G)$, is the minimum width over all possible tree decompositions of G.

We call the vertices in a tree decomposition "nodes" in order to avoid confusion with the vertices of a graph. An example of a graph G of treewidth two and a tree decomposition of the graph is given in Fig.1.

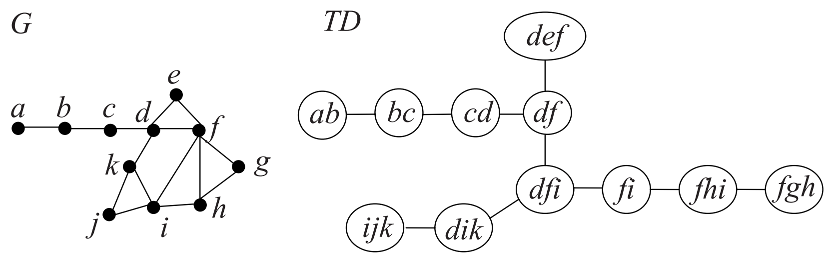


Fig. 1. A graph G of treewidth two and a tree decomposition of G

Dynamic programming algorithms for bounded treewidth graphs have been studied extensively. Most these algorithms have two stages: the first stage finds a tree decomposition of the graph and the second is a dynamic programming stage which solves the problem in a bottom-up traversal of the tree (of the tree decomposition).

In the bottom-up traversal, the computations of the nodes can be performed simultaneously. For example, the leaf nodes can be calculated synchronously. So multiple agents which processors are embedded can be employed. It is natural that parallel processing can reduce the time complexity dramatically. But it leads to a new problem that how to choose the root in order that the whole computation takes minimum time. It is clear that the definition of tree decomposition does not point out the root. In this paper, we focus on the root choice strategy of tree decompositions.

We consider the computation of the table for one node as a job. So the nodes of a tree decomposition are a set of jobs $j \in N = \{1, 2, \ldots, |V(T)|\}$. Assume that it takes unit processing time to compute each node and there are m parallel identical processors. As the classical scheduling problem, each processor can compute at most one job at a time and each job can be computed on any of the processors. It is clear that one node cannot be computed if one of its children has not been calculated. So we define the descendance relation $i \rightarrow j$ if the node i is one of the descendants of j. We investigate the problem is to find a root of the input tree decomposition in order to minimize the makespan, i.e. the completion time of the last job.

This paper is organized as follows: Section 2 presents the preliminary results. Section 3 and Section 4 propose and analyse our algorithm. Experimental results are given in Section 5.

2 Preliminary Results

Once the root is determined, our tree can be considered as an *intree* of which the nodes set is the same and the edges are directed according to the descendance relation. It is straightforward that there is no ancestor of the root. Assume that the root locates at *level* 1. The nodes which are the children of level 1 nodes are at level 2, and so on (see Fig. 1). Let l_{max} be the highest level of the whole tree.

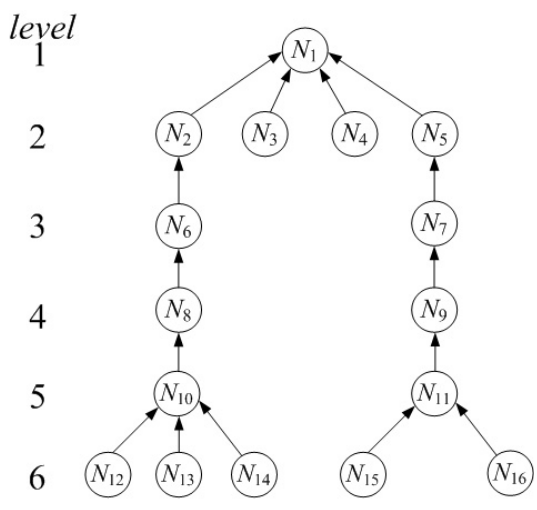


Fig. 2. A tree with $l_{max} = 6$

In the literature of scheduling, our problem is denoted by $P|p_j = 1$, *intree* $|C_{max}$ where $C_{max} = \max C_j | j \in N$ and C_j is the completion time of job j. Hu [7] proposed the classical "Level Algorithm" to solve this problem. The idea is simple: The processors choose the nodes which have high level to compute.

Consider the situation that we have three processors and the jobs of which the descendance relation constraints depicted in Fig. 2. A schedule solution is given as follows:

Table 1. A solution using Level Algorithm for Fig. 2.

	Time =1	2	3	4	5	6	7
Processor3	N_{14}	N_{12}	N_{10}	N_8			
Processor2	N_{15}	N_{13}	N_3	N_4	N_5		
Processor1	N_{16}	N_{11}	N_9	N_7	N_6	N_2	N_1

Hu's Level Algorithm

1) Let T be the tree rooted at r.

2) Compute the levels for all nodes by a traversal (e.g. DFS).

3) Set $t = 1$.

4) Let the queue Q store the nodes satisfy one of the following conditions:

-It has no child;

-Its children have been scheduled at time $[0, t-1]$.

5) Order the nodes of Q in the level decrease order.

6) Let S be the set of m jobs in the front of Q.
(If the number of nodes in Q is fewer than m, let S be the set of these nodes.)

7) Delete S from Q.

8) Schedule the jobs of S at time t.

9) Increase t by 1.

10) If there is one node not scheduled, Goto Step 4).

Furthermore, Hu [7] gave a formula to determine the value of C_{max}. Let $L(i)$ record the number of the nodes at level i. In the case that

$$\max_r \{ \frac{\sum_{j=1}^{r} L(l_{max} + 1 - j)}{r} \} \leq m \tag{1}$$

Hu proved that $C_{max} = l_{max}$.

In another case that

$$\max_r \{ \frac{\sum_{j=1}^{r} L(l_{max} + 1 - j)}{r} \} > m \tag{2}$$

choose a smallest integer $c \geq 1$ such that

$$\max_r \{ \frac{\sum_{j=1}^{r} L(l_{max} + 1 - j)}{r + c} \} \leq m < \max_r \{ \frac{\sum_{j=1}^{r} L(l_{max} + 1 - j)}{r} \} \tag{3}$$

Then we have $C_{max} = l_{max} + c$ in the latter case.

3 Description of the Algorithm

Let the tree to be scheduled be T with n nodes. Step 2) of Hu's Level Algorithm can be performed in $O(n)$ time. Once $L(i)$ is obtained for each $i \in [1, l_{max}]$, the value of C_{max} can be computed in $O(l_{max})$ time according to the Formulas (1), (2) and (3) in Section 2. Hence, the essential part is how to get the values of the array L, effectively.

It is natural to solve our problem by calculating C_{max} for each node of T. This can be done in $O(n^2)$ time. However, the time complexity is too expensive. We propose an algorithm which runs in $O(dn)$ time where d is the length of the longest path in tree T.

Some necessary definitions are given first. Let $e = (x, y)$ be an edge of $E(T)$ where $x, y \in V(T)$. And u, r are two nodes of the tree T. Let ST_r^u be the subtree rooted at u when the root of T is r. Two examples of ST_r^u and ST_r^v marked with dotted curves are shown in Fig. 3. Let $L_r^u(i)$ record the number of nodes at level i of the subtree ST_r^u. In the graph appears in the upper part in Fig. 3, we have that $L_r^u(1) = 1$, $L_r^u(2) = 2$ and $L_r^u(3) = 3$. And $L_r^v(1) = 1$, $L_r^v(2) = 2$, $L_r^v(3) = 3$, $L_r^v(4) = 5$ and $L_r^v(5) = 3$ in the lower graph. In the case that $u = r$, $L_r^r(i)$ stores the number of nodes at level i with respect to the tree T rooted at r (i.e. ST_r^r). And let $l_{max}(ST_r^u)$ be the highest level of the subtree ST_r^u. It is clear that $l_{max}(ST_r^u) = 3$ and $l_{max}(ST_r^v) = 5$ in Fig. 3.

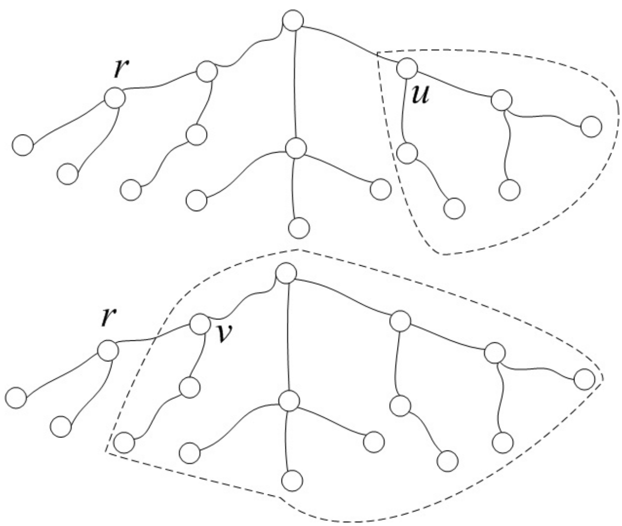


Fig. 3. ST_r^u and ST_r^v

What we want is an effective way to compute the values of L_r^r for each node r. First, we focus on how to determine L_r^r for one certain root r. Let r_0 be an arbitrary node. Along with the calculation of $L_{r_0}^{r_0}$, the values of $L_{r_0}^v$ for all $v \in V(T)$ are also obtained according to the following procedure:

Procedure for $L_{r_0}^v$

1) For each leaf node u, set $L_{r_0}^u(1) = 1$, $l_{max}(ST_{r_0}^u) = 1$ and label u "*visited*".
2) Let v be a non-*visited* node and the children of which have been *visited.*
3) Initial the values of $L_{r_0}^v$ and $l_{max}(ST_{r_0}^v)$ be 0.
3) For each v's child w do
4) Begin set $L_{r_0}^v(i+1) += L_{r_0}^w(i)$ for all $i \in [1, l_{max}(ST_{r_0}^w)]$
5) If $l_{max}(ST_{r_0}^w) > l_{max}(ST_{r_0}^v)$, then assign $l_{max}(ST_{r_0}^w)$ to $l_{max}(ST_{r_0}^v)$.
6) End
7) Increase $l_{max}(ST_{r_0}^v)$ by 1.
8) If there is a node is not *visited*, Goto Step 2).

Suppose we have obtained the values of L_r^r. Let r' be a node which is adjacent with r. We investigate that how to calculate $L_{r'}^{r'}$ effectively. Assume that $L_u^v(i) = 0$ where $i > l_{max}(ST_u^v)$ or $i < 1$. We have the following property:

$$\begin{cases} L_{r'}^{r'}(i) = L_r^r(i-1) - L_r^{r'}(i-2) + L_r^{r'}(i) & \text{if } i \geq 2, \\ L_{r'}^{r'}(i) = 1 & \text{if } i = 1. \end{cases} \tag{4}$$

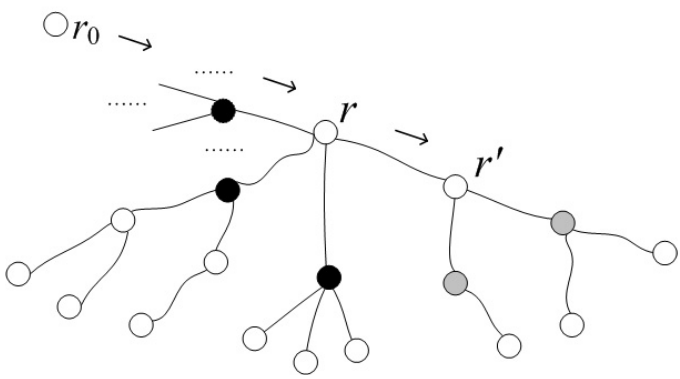


Fig. 4. ST_r^r and $ST_{r'}^{r'}$

We present the proof of Formula (4) as follows: It is trivial that $L_{r'}^{r'}(1) = 1$. We need to show $L_{r'}^{r'}(i) = L_r^r(i-1) - L_r^{r'}(i-2) + L_r^{r'}(i)$ for $i \geq 2$. It is straightforward that the nodes of $ST_{r'}^{r'}$ are composed by the node r', $ST_{r'}^r$ and $ST_r^{r'}$. The nodes of $ST_r^{r'}$ lie in the same level with respect to $ST_{r'}^{r'}$ due to the same subtree root. For the nodes of $ST_{r'}^r$, their levels are increased by one when considered as nodes in $ST_{r'}^{r'}$. Let i be an integer such that $2 \leq i \leq l_{max}(ST_{r'}^r)$. We have $L_{r'}^r(i) = L_r^r(i) - L_{r'}^r(i-1)$. Then, $L_{r'}^{r'}(i) = L_{r'}^r(i-1) + L_r^{r'}(i) = L_r^r(i-1) - L_r^{r'}(i-2) + L_r^{r'}(i)$. It follows.

Next, we introduce the method to maintain the values of $L_r^{r'}$ in Formula (4). The node r_0 is the first node of which $L_{r_0}^{r_0}$ is calculated as mentioned above.

Consider r_0 as the root of T. And perform a DFS (Depth First Search) starting from r_0. During the DFS, it is clear that the first time we visit a node r' is through the edge from its parent, denoted by r, to r'. Since DFS starting from r_0, it implies that $L_r^{r'}(i) = L_{r_0}^{r'}(i)$ for $i \in [1, l_{max}(ST_r^{r'})]$ and $ST_r^{r'}$ is the same as $ST_{r_0}^{r'}$. Hence, we could use $L_{r_0}^{r'}$ instead of $L_r^{r'}$ in Formula (4). The reason is that the values of $L_{r_0}^{r'}$ are fixed with respect to r_0 while the recalculations of $L_r^{r'}$ are expensive.

4 The Algorithm for Root Choice

Input: ***A tree*** T ***and the number of multiple agents, denoted by*** m
Output: ***The node minimizes*** C_{max} ***when chosen as root of*** T

1) Let P be a longest path of T and d be the length of P.

2) Let r_0 be an arbitrary node of T and set $C = +\infty$.

3) Calculate $L_{r_0}^v$ for all $v \in V(T)$ according to *Procedure for* $L_{r_0}^v$ in Section 3.

4) Perform a DFS starting from the node r_0.

4.1) For each node r' which is visited for the first time Do

4.2) Begin calculate the values of $L_{r'}^{r'}$ according to Formula (4);

4.3) Compute C_{max} considered r' as the root according to Formulas $(1 \sim 3)$;

4.4) If $C > C_{max}$, then assign C_{max} to C and store r' in *Root*;

4.5) End //foreach

5) Return the node *Root*.

The correctness of our algorithm is guaranteed by the validity of Formula (4) which is given in Section 3.

Since Steps (4.1$\sim$ 4.5) run in $O(d)$ time, the complexity of Steps (4) and (4.1$\sim$ 4.5) is bounded by $O(dn)$. It is clear that *Procedure for* $L_{r_0}^v$ can be done in $O(n)$ time. In the worst case, d equals n. Usually, d is in $o(n)$. For example, $d \in O(\log n)$ for full binary trees. Thus, the complexity of our algorithm is $O(dn)$ which is superior to the one runs in $O(n^2)$ time.

5 An Illuminating Example

Suppose there are four agents and the tree T is given in Fig. 5. At first, we choose the node N_{11} to be r_0. Calculate the values of $L_{r_0}^v$ according to our algorithm. The results are given in Table 2. We only show the values of non-leaf nodes.

After calculating $L_{r_0}^v$ for each node v, one DFS is employed to obtain the values of $L_{r'}^{r'}$. And we show the computing procedure when the search moving from the node N_{11} to N_{15}. According to the computation of Formula (4) mentioned above, we use the formula $L_{r'}^{r'}(i) = L_r^r(i-1) - L_{r_0}^{r'}(i-2) + L_{r_0}^{r'}(i)$.

We have $L_{N_{15}}^{N_{15}}(2) = L_{N_{11}}^{N_{11}}(1) - L_{r_0}^{N_{15}}(0) + L_{r_0}^{N_{15}}(2) = 1 + 1 = 2$,
$L_{N_{15}}^{N_{15}}(3) = L_{N_{11}}^{N_{11}}(2) - L_{r_0}^{N_{15}}(1) + L_{r_0}^{N_{15}}(3) = 3 - 1 + 1 = 3$,
$L_{N_{15}}^{N_{15}}(4) = L_{N_{11}}^{N_{11}}(3) - L_{r_0}^{N_{15}}(2) + L_{r_0}^{N_{15}}(4) = 5 - 1 + 2 = 6$,
$L_{N_{15}}^{N_{15}}(5) = L_{N_{11}}^{N_{11}}(4) - L_{r_0}^{N_{15}}(3) + L_{r_0}^{N_{15}}(5) = 8 - 1 + 2 = 9$.

The node N_{11} is the desired root where $C_{max} = l_{max} + 1 = 6 + 1 = 7$. One solution is given in Table 3. And N_{15} is an alternative root where $C_{max} = l_{max} + 2 = 5 + 2 = 7$. When the DFS is finished, we can get the values of L_v^v for all nodes. We present the values in Table 4.

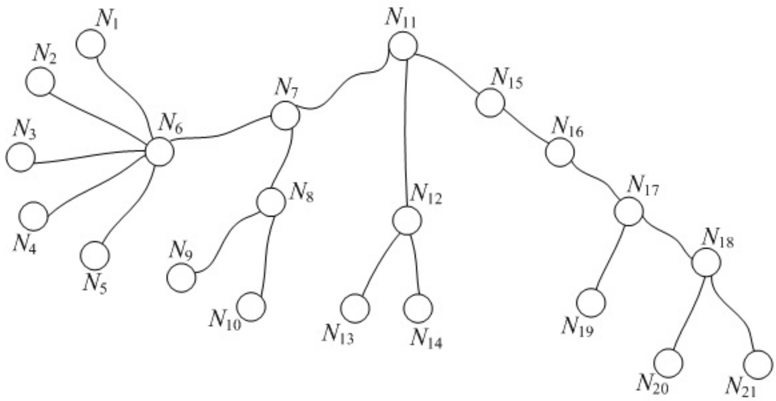


Fig. 5. A tree with 21 nodes

Table 2. The values of $L_{r_0}^v$

	Level								
	1	2	3	4	5	6	7	8	9
$L_{r_0}^{N_6}$	1	5							
$L_{r_0}^{N_8}$	1	2							
$L_{r_0}^{N_7}$	1	2	7						
$L_{r_0}^{N_{12}}$	1	2							
$L_{r_0}^{N_{18}}$	1	2							
$L_{r_0}^{N_{17}}$	1	2	2						
$L_{r_0}^{N_{16}}$	1	1	2	2					
$L_{r_0}^{N_{15}}$	1	1	1	2	2				
$L_{r_0}^{N_{11}}$	1	3	5	8	2	2			

Table 3. One solution when the root is N_{11}

	Time =1	2	3	4	5	6	7
Processor1	N_{21}	N_{18}	N_{17}	N_{14}	N_{15}	N_7	N_{11}
Processor2	N_{20}	N_9	N_3	N_{13}	N_{12}		
Processor3	N_{19}	N_5	N_2	N_{16}	N_6		
Processor4	N_{10}	N_4	N_1	N_8			

Table 4. The values of $L_{N_i}^{N_i}$

	Level								
	1	2	3	4	5	6	7	8	9
N_1	1	1	5	2	4	3	1	2	2
N_2	1	1	5	2	4	3	1	2	2
N_3	1	1	5	2	4	3	1	2	2
N_4	1	1	5	2	4	3	1	2	2
N_5	1	1	5	2	4	3	1	2	2
N_6	1	6	2	4	3	1	2	2	
N_7	1	3	9	3	1	2	2		
N_8	1	3	2	7	3	1	2	2	
N_9	1	1	2	2	7	3	1	2	2
N_{10}	1	1	2	2	7	3	1	2	2
N_{11}	1	3	⑤	⑧	②	②			
N_{12}	1	3	2	3	8	2	2		
N_{13}	1	1	2	2	3	8	2	2	2
N_{14}	1	1	2	2	3	8	2	2	2
N_{15}	1	2	3	⑥	⑨				
N_{16}	1	2	3	4	4	7			
N_{17}	1	3	3	1	2	4	7		
N_{18}	1	3	2	1	1	2	4	7	
N_{19}	1	1	2	3	1	2	4	7	
N_{20}	1	1	2	2	1	1	2	4	7
N_{21}	1	1	2	2	1	1	2	4	7

6 Conclusions and Future Works

We settle the root choice problem of tree decompositions when multiple agents are available, i.e. parallel processing can be applied. An $O(dn)$ algorithm is proposed where d is the length of a longest path and n is the number of nodes. Our algorithm is superior to the algorithm enumerates the root with all the nodes which runs in $O(n^2)$ time. Since the root choice affects the latest completion time (i.e. C_{max}) dramatically, our work helps to optimize the schedule of the agents. Since tree decomposition can be employed to solve many hard problems, parallel processing on tree decomposition will make the computation effectively. And our investigation answers the root choice problem which is essential.

Since we calculate the values of L_r^r in DFS order, the branch-and-cut technique can be applied. That is, we can determine all nodes in certain subtree cannot be the optimal root. If so, we skip to search the subtree. This happens probably when we visit the subtrees of which the height is low. For example, all the nodes of the subtree $ST_{N_{11}}^{N_{18}}$ have l_{max} greater than 7.

Furthermore, our algorithm is not optimal in time complexity. We find there are more properties of the values of C_{max} in trees. A more effective algorithm might be developed by means of these properties. This will be given in the subsequent paper.

References

[1] Arnborg, S.: Efficient algorithms for combinatorial problems on graphs with bounded decomposability - A survey. BIT 25, 2–23 (1985)
[2] Arnborg, S., Lagergren, J.: Easy problems for tree-decomposable graphs. Journal of Algorithms 12, 308–349 (1991)
[3] Arnborg, S., Proskurowski, A.: Linear time algorithms for NP-hard problems restricted to partial k-trees. Disc. Appl. Math. 23, 11–24 (1989)
[4] Bodlaender, H.L.: Treewidth: Characterizations, Applications, and Computations. In: Fomin, F.V. (ed.) WG 2006. LNCS, vol. 4271, pp. 1–14. Springer, Heidelberg (2006)
[5] Borie, R.B.: Recursively Constructed Graph Families. PhD thesis, School of Information and Computer Science, Georgia Institute of Technology (1988)
[6] Borie, R.B., Parker, R.G., Tovey, C.A.: Automatic generation of lineartime algorithms from predicate calculus descriptions of problems on recursively constructed graph families. Algorithmica 7, 555–581 (1992)
[7] Hu, T.C.: Parallel sequencing and assembly line problems. Operations Research 9, 841–848 (1961)
[8] Lauritzen, S.J., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. The Journal of the Royal Statistical Society Series B (Methodological) 50, 157–224 (1988)
[9] Robertson, N., Seymour, P.D.: Graph minors (II) Algorithmic aspects of tree-width. J. Algorithms 7, 309–322 (1986)
[10] Thorup, M.: Structured programs have small tree-width and good register allocation. Information and Computation 142, 159–181 (1998)
[11] Wimer, T.V.: Linear Algorithms on k-Terminal Graphs. PhD thesis, Dept. of Computer Science, Clemson University (1987)
[12] Wimer, T.V., Hedetniemi, S.T., Laskar, R.: A methodology for constructing linear graph algorithms. Congressus Numerantium 50, 43–60 (1985)
[13] Wu, T.H.: Fiber network service survivability. Artech House, Inc., Norwood (1992)
[14] Yamaguchi, A., Aoki, K.F., Mamitsuka, H.: Graph complexity of chemical compounds in biological pathways. Genome Informatics 14, 376–377 (2003)

An Agent's Activities Are Controlled by His Priorities

Huiliang Zhang, Shell Ying Huang, and Yuming Chang

School of Computer Engineering, Nanyang Technological University,
639798, Singapore
zhlpku@yahoo.com, assyhuang@ntu.edu.sg
http://www.ntu.edu.sg

Abstract. Activity scheduling mechanism plays a critical role in the correct behaviour of BDI agents. For example, a robotic agent to serve at home should carry out the right activities at the right times. However the scheduling of deliberation about new beliefs and the scheduling of intention execution have not been carefully studied in most BDI systems. Usually if there is any differentiation of urgency among different tasks, a constant utility/priority value is used by a task selection fnction. We argue that priorities should be allowed to change with time and a linear function of time may not be the best for all tasks. In this paper, we propose to enrich the BDI framework with an extension which consists of 2 processing components, a PCF (Priority Changing Function) Selector and a Priority Controller. With this extension priorities of desires/intentions may have different initial values and may be changed with time according to the chosen PCFs. We propose a method of constructing PCFs which model the change of priorities in human behaviors when dealing with several things at the same time. We also propose a method to realize the change of the priorities of existing desires/intentions due to the generation of new beliefs/desires/intentions if necessary. We show by simulation experiments that Ramp function and especially the Sigmoid function can control the activities of an agent better than constant priorities with respect to getting tasks of various importance and urgency done with smaller Mean Earliness and smaller Mean Tardiness.

1 Introduction

Bellman defines AI in [1] as the automation of activities that we associate with human thinking, activities such as decision making, problem solving, and learning. One such activity is to decide when is the appropriate time to think about a certain matter or to do something. For an intelligent agent, this means it should know when to deliberate and when to act in addition to being able to deliberate on how to achieve a goal and how to carry out a plan. There has been significant amount of work on solving the "How" problem but not the "When" problem. For example, the PRS (procedural reasoning system) obtains the ability of reasoning in complex ways about dynamic processes while keeping appropriate responsiveness and control [5]. AgentSpeak(L) [8] and LORA (logic of rational agents) [9] are two sets of operational semantics defined for BDI agents. The decision is made through logic reasoning. All these works are solutions to the "How" question.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 723–732, 2008.

The “When” question, that is, the scheduling of deliberation about new beliefs and the scheduling of intention execution are usually omitted in these BDI systems. For example, in AgentSpeak(L) [8], the selection function S_I selects an intention to execute from the intention set *I*. The detailed selection criteria are not specified. We believe the scheduling of intention is crucial in an agent’s ability to cope with the changing world. Some scheduling mechanisms appear in subsequent researches. In AgentSpeak(XL) [2], an extension version of AgentSpeak(L), a task scheduler is incorporated into the interpreter to decide how to select intentions. The set of intentions in the AgentSpeak(L) is converted into a corresponding TÆMS task structure. The 'enables' and 'hinders' relationships among the plans indicate which plan may be executed first. Another method is shown in the JAM agent architecture [4]. The intention selection is based on the utility value of the plan set by the user. The intention with higher utility will be executed first.

We consider the problem of deliberation scheduling and intention scheduling in an agent who will behave like an “average human”. We associate a single priority value with each desire or intention to facilitate the scheduling of deliberations and intention executions. For humans, their priorities change with time. The priorities may be affected by how close it is to the deadline of a task, or how appropriate it is to do a certain task at a certain time from human point of view. The deadlines of tasks may also change, either forward or backward. Therefore to represent the priority of a goal or an intention of an agent realistically, a function of time should be used instead of a constant value. Representing the priority of a desire/intention by a function of time instead of a constant also has another advantage. If the priority is a constant and there are other desires/intentions with higher priorities continuously being generated, this lower priority desire/intention will never have a chance of being considered.

Currently, the control of the priority changing with time has not been adequately researched even though some work has been done in the artificial life community. In [6], a priority control mechanism for behavioral animation is proposed. The priority is set at minimal value immediately after the agent displays certain behaviour like drinking. Then this priority increases with time. The increased priority will induce the agent to drink again. However, expecting the priorities of all desires and intentions to change in the same manner is not realistic. Different desires and intentions should be allowed to change their priorities in various suitable ways.

A parallel BDI agent framework was proposed in [10] to achieve better reactivity and rationality in intelligent agents. This framework equips a BDI agent with the natural abilities of doing several things at the same time and the ability of prioritizing the deliberations and intention executions according to the urgency of the tasks. The level of priority is used in the scheduling of the desires/intentions in the agent. However this mechanism has the problem that with priorities set at constant levels, some desire or intention may be starved indefinitely by desires or intentions with higher levels of priorities.

In this paper, we proposed a priority control extension to the parallel BDI agent framework [10] in order to support the capabilities of representing the changing importance of different desires and intentions. This is an extension of the paper in [11] where only independent intentions are considered. Pre-defined *Priority Changing Functions(PCFs)* are associated with the desires and intentions. A Priority Controller will compute the priority value of the desires and intentions to help the scheduling decisions to be made at various time moments. We propose a method of constructing PCFs which model the change of priorities in human behaviors when dealing with several things at the same time. The remainder of this paper is structured as follows. In Section 2, we present the BDI agent framework with the proposed extension of priority control. In Section 3, we discuss the mechanisms of priority controls proposed. The priority control functions are evaluated in Section 4. At the end of the paper, we summarize the contribution of the paper and future work.

2 BDI Agent Framework with the Priority Control Extension

The parallel BDI agent framework consists of three main components running in parallel: the belief manager, the intention generator and the intention executor. These three components represent the three steps in the deliberation process of an agent: *detect*, *think* and *act* respectively. The three components will retrieve and update data in the three data structures: *beliefs*, *desires*, and *intentions*. The *beliefs* contain the agent's view of its environment and of itself. The *desires* will maintain the goals of the agent. The *intentions* store the plans to be executed to achieve the agent's goals. Higher priority beliefs, desires and intentions preempt the processing of lower priority ones in the parallel running components.

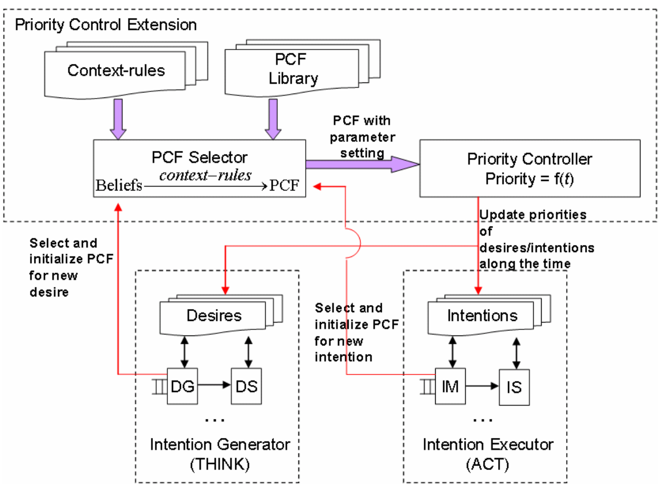


Fig. 1. Priority control extension to the parallel BDI framework (only parts of the original framework [10] that interact with the extension are shown)

To represent the dynamic change of priorities of a desire or intention, each desire or intention will be associated with a Priority Changing Function (PCF) which defines how the priority should change with time. The priority control extension to the agent architecture is shown in Figure 1. Two processing components are introduced into the BDI agent, a PCF (Priority Changing Function) Selector and a Priority Controller. When a desire/intention is generated, the intention generator will call on the PCF

Selector which will, based on some context-rules, (i) select a suitable PCF for the new desire/intention from the PCF Library and (ii) decide on the values of the parameters if any for the PCF. The signature of the function of the PCF Selector is as follows: $PCF\_Selector : desires \times PCFs \rightarrow PCFs$ and $PCF\_Selector : intentions \times PCFs \rightarrow PCFs$ where *desires* and *intentions* are the set of desires and the set of intentions of the agent respectively and *PCFs* is the set of priority changing functions in the PCF library. More discussions on the PCF Selector are given in Section 3.3.

The Priority Controller will be responsible for updating the priorities of the desires/intentions according to their PCFs as time passes. In order to have the priorities of desires and intentions assessed accurately but not computed unnecessarily, the Priority Controller will update the priorities of desires/intentions each time the BDI agent is to select a desire/intention to execute by

$$\text{Priority} = \text{f}(t) \tag{1}$$

where, f (t) is the PCF and t is the current time.

This extension allows a BDI agent to select suitable pre-defined PCFs for the desires/intentions and compute the priority values of desires/intentions at various points in time. Then the deliberation of the desires and the execution of the intentions can be scheduled by the Intention Generator or the Intention Executor based on their importance or urgency, represented by their priority values at the time. This enriches the BDI agent with the ability of realizing the scheduling of the desires/intentions in a more realistic way.

3 Priority Control

With the architectural support as described in the previous section, we present in this section the mechanism of priority control so that what an agent deliberates and acts on are decided by his priorities, just like the humans. The basic control of priorities comes from the Priority Changing Function (PCF) which defines how the priority of a desire or an intention should change with time. The PCF is defined by

$$\text{f}(t) = \text{Maximum priority} * \text{I}(t) \tag{2}$$

where *Maximum priority* is the highest value which a desire or an intention may have; I(t) is the function of influence factor with a range [0, 1]. *Maximum priority* is an intrinsic constant value of a desire or an intention where I(t) controls the changes in priority with time . This PCF with a suitable I(t) will be applied in the parallel BDI agent to provide some human-like behaviour. I(t) can be different for different desires and intentions.

Very often, human interests in a certain goal or intention go through a few phases. According to Maslow's theory of hierarchy of needs [7], a need will increase along with other needs. During the process a need is being satisfied, it will lose its dominance. So one phase is the period when the interests are getting stronger and stronger and we call it the reminding phase. Another phase is the period when the interests or appropriateness will become less and less and occasionally the goal or intention may even be forgotten so this is called the forgetting phase. There are also situations where someone has an unchanging interest to do something and this is called the unchanging phase. We will first describe how I(t) models the various phases of priority changing and then show how the various phases work together. Then we will present how the effect of a new belief/desire/intention on some existing desire/intention is modeled.

3.1 The Reminding Phase of a PCF

This is the phase where an agent has increasing inclination to deliberate on how to achieve a goal or to execute an intention plan. The priority of the goal or intention should increase gradually until it reaches its maximum value. We call this phase the 'reminding' phase. The manner by which the priority of a desire or an intention increases may be different from that of the others. We propose two different functions to model the way a priority may increase, the ramp and Sigmoid functions.

For the two functions, the value I(t) at $t = 0$ is y_0. The value of y_0 when multiplied by the maximum priority as shown in Equation (2) will return f(t) the initial priority of a desire or intention. And at $t = t_m$, the value of I(t) should be 1 or very close to 1. t_m is the time when f(t) is to reach its maximum priority value. The ramp function is a model where the priority increases at a constant rate. It is realized by:

$$\mathrm{I}(t) = y_0 + \frac{1-y_0}{t_m} * t \tag{3}$$

To simulate the reminding phase using the Sigmoid function, we extract part of the Sigmoid curve between [-α, α] in x-axis, $0<\alpha<0.5$. The original Sigmoid curve is shifted to right by $\ln\left(\frac{1}{\alpha}-1\right)$ in x-axis. Then the curve between [α, 1- α] in y-axis will be resized to [0, t_m] in x-axis and [y_0, 1] in y-axis. We get the equation:

$$\mathrm{I}(t) = \left[\left[1+\exp\left(\left(1-\frac{2t}{t_m}\right) * \ln\left(\frac{1}{\alpha}-1\right)\right)\right]^{-1} - \alpha\right] * \frac{(1-y_0)}{(1-2\alpha)} + y_0 \tag{4}$$

Sigmoid functions with different α, t_m and y_0 exhibit different shapes of the function. With a smaller α, the function simulates a slower rise at the beginning followed by a steeper increase of priority and then a slow rise near time t_m. By controlling the value of α, we can control the way the priority increases over time within the pattern of Sigmoid function.

The Sigmoid function has the property that the gradient of I(t) gradually decreases to zero when t approaches t_m. This simulates the increase in the agent's interest in a goal or an intention gradually stops as the interest reaches its maximum. For (4), (i) I(t) is increasing for $t \le t_m$; (ii) the rate of increase first increases for $0 \le t \le t_m - \sigma$ ($\mathrm{I}''(t) \ge 0$) and then decreases for $t_m - \sigma \le t \le t_m$ ($\mathrm{I}''(t) < 0$); (iii) the increase in function value and the rate of increase at t_m are 0, which means that the trend to increase the priority has stopped.

Table 1. Results with different α for the Sigmoid function

Inter-arrival mean = 25												
α	0.1 – 0.3			0.4 – 0.6			0.7 – 0.8			0.9		
	Pass (%)	Mean Earliness	Mean Tardiness	Pass (%)	Mean Earliness	Mean Tardiness	Pass (%)	Mean Earliness	Mean Tardiness	Pass (%)	Mean Earliness	Mean Tardiness
.100	25.6	-46.10	332.39	60.8	-43.57	34.90	83.4	-49.28	9.97	88.1	-49.84	4.56
.0500	28.2	-42.53	309.74	62.4	-42.04	33.35	81.7	-47.33	10.67	88.5	-51.30	4.99
.0250	35.7	-46.69	291.15	65.3	-45.85	37.64	81.6	-48.47	9.55	89.0	-50.77	3.16
.0218	31.7	-39.45	269.52	63.1	-44.27	26.94	82.8	-41.53	10.53	94.5	-46.96	1.13
.0187	29.6	-41.67	263.76	60.3	-40.34	42.54	78.7	-44.92	15.62	83.5	-48.49	5.75
.0125	33.4	-42.79	343.15	53.8	-38.57	44.40	72.0	-41.21	10.48	82.1	-45.00	4.36
0	21.4	-39.20	272.14	31.6	-69.08	46.57	42.8	-54.74	12.92	49.5	-59.01	3.61

To demonstrate the way to set an optimal value for α, a simulation experiment is done in which 100 events arrives according to an exponential distribution of mean = 25 time units. $t_m = T_C - T_E$, where T_C is the time by which the task should be completed, T_E is the execution time a certain task takes.. The T_E of each event is assigned a random execution time between 1 and 49 with mean = 25 time units. The T_C for each event is between $1.25T_E$ to $8T_E$ which means events are known in advance of their deadline by different amount of time, some are known much earlier than others. The events are divided into four groups according to their maximum priorities MP: 0.1-03, 0.4-0.6, 0.7-0.8 and 0.9. y_0 is 0.05MP. The experiment results are shown in Table 1. In the Table, Pass (%) is the percentage of events meeting deadlines, Mean Earliness is the Average time left for which jobs are finished before the deadline, and Mean Tardiness is the Average time by which jobs are finished after the deadline. Binary search is done for the value of α to achieve high Pass rate, small Mean Tardiness, especially for high priority events, and small Mean Earliness. As we see, with α = 0.0218, the agent has the best performance.

3.2 The Forgetting Phase and the Unchanging Phase of a PCF

The forgetting phase is the phase where an agent's interest in a goal or an intention plan is fading, or the appropriateness of doing a certain task is diminishing with time. For example, cleaning windows is better done in day time and it is getting late. In this case, the priority of the goal or the intention should be decreased gradually. So if the agent has other intentions with higher priorities like serving some guests, he should delay cleaning windows. If the priority decreases to a value below a threshold, which corresponds to that it is too late to clean windows, the intention will be removed (forgotten). The first significant study on memory was performed by Hermann Ebbinghaus and published in 1885 as *On Memory*. Ebbinghaus was the first to describe the shape of the forgetting curve [3]. This curve is the biological base on which we simulate the process of intention retention. The forgetting curve is described as:

$$R = e^{-t/S} \tag{5}$$

where R is the retention, which means the ability to retain things in memory; t is the elapsed time; S is the strength of memory, which means the duration of things in memory. The forgetting curves with different S exhibit different rates of decay of priority. The smaller S corresponds to a shorter retention interval. When S is a very big value, there will be little forgetting and priority will decrease very slowly.

The unchanging phase is the phase where an agent's interest in a goal or an intention plan is at a constant level. The priority of the goal or the intention is specified by a constant.

3.3 The Complete PCF

The complete PCF, or more specifically, the complete I(t) is formed by either any one of the single phased functions described earlier, or it is a concatenation of two or more single phased functions. For example, we can compose an I(t) by combining the functions of the reminding phase and the forgetting phase together. The result is a reminding-forgetting function:

$$\mathrm{I}(t) = \begin{cases} y_0 + \dfrac{1-y_0}{t_m} * t & \text{if } t <= \mathrm{t_m} \\ \exp\left(- \dfrac{t - \mathrm{t_m}}{\mathrm{S}} \right) & \text{otherwise} \end{cases} \tag{6}$$

An example of the situation modeled by this I(t) is that an agent intends to clean windows starting latest by t_m if he finds the time. After t_m it becomes less desirable to clean windows so the priority goes into the forgetting phase. We can in fact append a constant priority phase with value 0 for this intention for the evening and night time and another reminding phase for next morning. The same can be done to simulate Maslow's theory of hierarchy of needs [7], where a need will increase and after it is satisfied, it will lose its dominance. All these information can be considered by the PCF Selector in order to choose a suitable PCF. In the simplest case, a constant priority function can be chosen for a desire/intention that has no deadline and the sigmoid/ramp function for one that has. The default will be the sigmoid function. As shown by the experiments presented later in the paper, the sigmoid function helps to have tasks finished with smallest Mean Tardiness and Mean Earliness in general. When more information about a desire/intention is available, the PCF Selector will be able to choose the PCF that best suits the desire/intention.

3.4 Priority Change Caused by Other Beliefs/Desires/Intentions

It is noticed that the I(t) function as described earlier changes the priority of a desire or an intention in the absence of the effect of new beliefs, desires and intentions. However, a new belief/desire/intention may make an existing desire/intention more urgent or less urgent, therefore the priority of the affected desire or intention needs to be increased or decreased. For example, suppose the human master asked his robotic agent to wash his car while the agent is doing cleaning in the house and the robotic agent also has a few other things to do. The agent has the intention to wash the car but the priority is not as high as his other intentions. After a little while the master reminds the agent about washing his car. At this point the priority for washing the car should be increased. So the new belief that the car needs to be washed sooner should have the effect of increasing the priority of the intention of washing the car. Another scenario that will change the priority of an existing intention: the robotic agent has the intention to tidy up a room but his master tells him to iron a shirt in the next 10 minutes. The robotic agent generates the intention to iron the shirt and has to lower the priority of tidying up the room. In situations like these examples, the priority of an existing desire/intention at *t* and beyond is affected. *t* is the time when a new belief, a new desire or a new intention is generated and it is the moment when the priority of an existing desire/intention should be changed if it is affected.

To model the effect on the I(t) of the existing desire/intention, one simple way to do it is to change the T_C value and therefore the t_m value in I(t) is changed. This results in an increase or decrease of the I(t) value at t which brings the increase or decrease of the priority of the affected desire/intention.

4 Evaluation of PCFs

To examine how the PCFs perform in controlling the scheduling of activities, another set of experiments is done to compare the performance of agents with priorities changed by the Ramp, Sigmoid reminding functions, and with static priorities. The way the experiments are set up is the same as in Table 1. Three different event inter-arrival means are used in 3 groups of experiments. For each group of experiments, multiple runs are conducted to obtain the average of Pass rate, Mean Earliness and Mean Tardiness. The results are shown in Table 2.

The first group of experiments presents the situation where inter-arrival mean which is 20 is smaller than the average processing time of events T_E which is 25 time units. In such a situation, priority control with the static values produces very high Pass rate and small tardiness but also large earliness for the highest priority events. At the same time, the lowest priority events suffer from very low Pass rate and huge tardiness. The Ramp function has the lowest Pass rate for all events and the Sigmoid function returns better Tardiness and better earliness for almost all events but the Pass rate for all events except the lowest priority ones is lower than the static priority.

In the second group of experiments, the event inter-arrival mean is equal to the average processing time of events. Ramp function again produces the worst Pass rate but its tardiness and earliness are better than the static function for all events with the only exception in the tardiness of the highest priority events. Sigmoid function has better Pass rate than the static function for the lowest priority events and comparable Pass rate for other events. Its tardiness and earliness are all better than the static function for all events.

When the event inter-arrival mean is larger than the average processing time of events in the third group of experiments, Ramp function returns comparable Pass rate and better tardiness and earliness for all events than the static function except the slight increase in tardiness for the highest priority events. Sigmoid function generates better Pass rate for the lowest priority events and comparable Pass rate for other events. It also achieves smaller tardiness and earliness for almost all events.

The general conclusion from the experimental results is that static priority guarantees that high priority events(priority = 0.9) have high Pass rate and small Mean Tardiness at the expense of the lower priority ones(priority = 0.1 to 0.3). As the priority of events decreases, the Pass rate decreases and Mean Tardiness increases. This confirms that the lower priority events tend to be starved of processing time under static priority. The Ramp PCF and especially the Sigmoid PCF are able to reduce the tardiness and increase the Pass rate for low priority events, reduce the tardiness and earliness of most events with a small increase of tardiness for high priority events. This means more events are handled not too early and not too late and we can conclude that using these PCFs produces more desirable behaviour in BDI agents.

Table 2. Results of agents with different PCFs

PCF	0.1 – 0.3			0.4 – 0.6			0.7 – 0.8			0.9		
	Pass (%)	Mean Earliness	Mean Tardiness	Pass (%)	Mean Earliness	Mean Tardiness	Pass (%)	Mean Earliness	Mean Tardiness	Pass (%)	Mean Earliness	Mean Tardiness
Inter-arrival mean = 20												
Static	6.8	-57.342	1215.936	45.0	-80.769	257.975	84.5	-81.017	42.552	96.9	-86.467	5.009
Ramp	8.9	-31.288	494.061	24.7	-47.459	165.464	42.6	-46.182	78.571	58.1	-50.134	53.733
Sigmoid	14.0	-32.282	631.128	33.5	-31.702	93.431	56.8	-32.323	18.767	73.2	-34.469	7.467
Inter-arrival mean = 25												
Static	21.7	-66.847	637.102	64.2	-76.525	112.0	89.9	-89.622	25.009	98.1	-89.606	2.483
Ramp	23.2	-48.96	213.96	51.3	-52.48	60.75	71.2	-50.95	33.48	80.1	-59.16	14.71
Sigmoid	31.7	-39.45	269.52	63.1	-44.27	26.94	82.8	-41.53	10.53	94.5	-46.96	1.13
Inter-arrival mean = 30												
Static	48.8	-81.879	202.074	77.0	-82.931	58.982	92.6	-88.243	13.898	99.6	-90.023	0.330
Ramp	49.5	-61.059	93.621	79.4	-60.304	24.878	90.5	-63.874	8.912	95.9	-68.647	2.710
Sigmoid	55.4	-55.370	144.777	80.0	-52.034	18.946	92.7	-59.889	4.897	94.3	-60.029	1.949

5 Conclusion and Future Work

In this paper, we propose to enrich the BDI agent framework with the priority control extension consisting of 2 processing components, a PCF (Priority Changing Function) Selector and a Priority Controller. This provides the support at the architecture level for the control of priorities of desires and intentions that may have different initial values and will change with time according to the chosen PCF. This enables the automation of the scheduling of the agent's activities in a more human-like manner and removes the need for the scheduler to reason about deadlines when selecting activities. A method to compose priority changing functions from single phases of priority change is proposed. This is a flexible way to model the change of priorities in human activities. Different priority functions will introduce different behaviour in an agent's decision on when an activity should be done. We also propose a method to realize the change of the priorities of existing desires/intentions due to the generation of new beliefs/desires/intentions if necessary. Experimental results show that using these PCFs produces more desirable behaviour in BDI agents.

One shortcoming in this method at the moment is that it can only consider a one to many relationship, that is, the effect of one new belief/desire/intention on the existing desires/intentions. When intention A makes both intention B and intention C more urgent, changing the value of t_m in the PCFs of B and C separately and independently may not be appropriate. Another unresolved yet important issue is how to decide the maximum priority of a task. The parameter, if not set properly, may result in the agent switching between tasks in an unnatural way. However, we believe one way is to let the agent learn from its experience like humans learn time management. The agent may use a neural network and first go through supervised learning to obtain parameter values for a set of ‘anchor’ tasks. Subsequently, the agent may infer values for other tasks which are similar to some anchor tasks. These issues are the subjects of our future work.

References

[1] Bellman, R.E.: An Introduction to Artificial Intelligence: Can Computers Think? Boyd and Fraser Publishing Company (1978)

[2] Bordini, R.H., Bazzan, A.L.C., Jannone, R.d.O., Basso, D.M., Vicari, R.M., Lesser, V.R.: AgentSpeak(XL): efficient intention selection in BDI agents via decision-theoretic task scheduling. In: Proceedings of International Conference on Autonomous Agents, pp. 1294–1302 (2002)

[3] Ebbinghaus, H.: Memory: A Contribution to Experimental Psychology, New York: Teachers College, Columbia University

[4] Huber, M.J.: JAM: A BDI-theoretic mobile agent architecture. In: Proceedings of third International Conference on Autonomous Agents (Agents 1999), pp. 236–243 (1999)

[5] Ingrand, F.F., Georgeff, M.P., Rao, A.S.: An architecture for real-time reasoning and system control. IEEE Expert 7(6), 34–44 (1992)

[6] Lamarche, F., Donikian, S.: The orchestration of behaviours using resources and priority levels. In: Proceedings of the Eurographics Workshop, Manchester, U.K., September 2-3, 2001, pp. 171–182 (2001)

[7] Maslow, A.H.: Motivation and Personality. HarperCollins Publishers (1987)

[8] Rao, A.S.: AgentSpeak(L): BDI agents speak out in a logical computable language. In: Proceedings of the 7th European workshop on Modelling autonomous agents in a multi-agent world: agents breaking away, Einhoven, The Netherlands, pp. 42–55 (1996)

[9] Wooldridge, M.: Reasoning about rational agents. The MIT Press, Cambridge (2000)

[10] Zhang, H., Huang, S.Y.: A General Framework for Parallel BDI Agents. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT-2006), Hong Kong, China, December 18-22 (2006)

[11] Zhang, H., Huang, S.Y.: Dynamic Control of Intention Priorities of Human-like Agents. In: Proceedings of the 17th European Conference on Artificial Intelligence, Riva del Garda, Italy (August 28-September 1, 2006)

Mining Based Decision Support Multi-agent System for Personalized e-Healthcare Service*

Eunyoung Kang, Hee Yong Youn, and Ungmo Kim

School of Computer Engineering, Sungkyunkwan University,
440-776, Suwon, Korea
{eykang, youn, umkim}@ece.skku.ac.kr

Abstract. In this paper, we proposed a multi-agent based healthcare system (MAHS) which is the combination of a medical sensor module and wireless communication technology. This MAHS provides broad services for mobile telemedicine, patient monitoring, emergency management, doctor's diagnosis and prescription, patients and doctors and information exchange between hospital workers over a wide area. Futher more, MAHS is connected to a Body Area Network (BAN) and a doctor and hospital support staff. In this paper, we demonstrate how we can collect diagnosis patterns, classify them into normal, and emergency and be ready for an emergency by using the real-time biosignal data obtained from a patient's body. This proposed method deals with the enormous quantity of real-time sensing data and performs analysis and comparison between the data of patient's history and the real-time sensory data. In this paper, we separate Association rule exploration into two data groups: one is the existing enormous quantity of medical history data. The other group is real-time sensory data which is collected from sensors measuring body temperature, blood pressure, pulse. We suggest methods to analyze and model patterns of a patient's state for normal, and very emergency, and making decisions about a patient's present status by utilizing these two data groups.

Keywords: Healthcare System, Multi-Agent System, Ubiquitous Computing Environment, Mining, Association Rule.

1 Introduction

The development of mobile devices, such as PDA, laptop and notebook computers, and technologies, such as sensors, computerized chips, and wire/wireless networks, has made ubiquitous computing possible. Ubiquitous computing offers a user-oriented computing environment, where users can freely access computers anywhere, at any time. Healthcare is a field where ubiquitous computing is most widely used [1, 5].

e-Healthcare is developing from merely curing diseases which is the basis of the existing medical service to maintaining health and prevention of diseases [2, 3, 4]. To

* This work was supported in part by the Ubiquitous Autonomic Computing and Network Project, 21st Century Frontier R&D Program and by the university IT Research Center project (ITRC), funded by the Korean Ministry of Information and Communication.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 733–742, 2008.

manage health status and diagnosis, we need to have a method to manage a history of diagnosis data effectively, and extract useful knowledge based on that.

Association rule is the one of the data mining techniques used to extract knowledge from databases, and is being used by many researchers. There are many useful applications of this, such as the medical field, supervision of the ecosystem and environment, supervision of dangerous buildings, and tracking moving objects. The knowledge extracted from data by using Association rules are related to events or changing environments in the medical field, for example; " the body temperature and pulse are correlated", "when there is bleeding, the blood pressure goes down and pulse goes up". By analyzing the relations between data, we can predict a patient's present and future status [8, 9].

In this paper, we propose methods to collect diagnosis pattern, classify them into normal, and emergency and be ready for an emergency by using the real-time biosignal data obtained from a patient's body. In order to do this, we must deal with an enormous quantity of real-time sensing data and perform the analysis and comparison between the patient's historical data and the real-time sensory data. In addition, we must convey precise and fast diagnosis to doctors, so they can perform decision making. Therefore, in this paper, we separate Association rule exploration into two data group: one is the existing enormous quantity of medical historical data. The other group is real–time data from the body which is collected from sensors measuring body temperature, blood pressure, pulse. We suggest methods to analyze and model patterns of a patient's state for normal, and very emergency, and making decisions about a patient's present status by utilizing these two data groups.

We demonstrate a healthcare system that remotely captures real-time medical information with sensors to provide a patient with his doctor's diagnosis and prescription.

The remainder of the paper is organized as follows. Section 2 details the agent platform and related approaches. Section 3 provides the details of the proposed MAHS architecture and multi-agent system. Section 4 presents the proposed implementation of MAHS. Finally, Section 5 explains our conclusions.

2 Related Works

2.1 JADE and LEAP

The JADE (Java Agent DEvelopment Framework) is middleware developed by TILAB for the development of distributed multi-agent applications based on a peer-to-peer communication architecture. The advantage of the multi-agent system is that parallel processing with many agents enables the system to provide complicated services that couldn't be processed by a single agent [13]. The addition of new agents enables the system to be easily extended to handle new services. The system exchanges and shares information through co-operation among agents.

The LEAP (Lightweight Extensive Agent Platform) is probably the best known agent platform for small devices [14]. LEAP is the first attempt to implement a FIPA agent platform that runs seamlessly on both mobile and fixed devices over both wireless and wired networks. In essence, a module called LEAP, which is connected through wireless networks, enables optimization of all communication mechanisms for devices with limited resources.

2.2 Decision Support System and Mining

Lots of studies have introduced methods to make decision by processing a vast volume of data. Those methods are the ones to obtain information through statistical calculations, as of the volume and gain of information, and artificial intelligent elements. Such process requires a complex calculation and lots of time, and it's difficult to add to and maintain a new module. Bio-information conveyed from the body of a patient contains a great deal of data. Therefore, if there happens an emergency situation, you should determine his status real-time, quicker and more accurate results are required. To solve those challenges, we propose a method where a pattern has been accurately and quickly abstracted through mining to provide a flexible system for easily reflecting new information and module through an agent-based system.

2.3 Existing Approaches

There are researchers seeking connections between data mining techniques and decision making. Data mining techniques analyze the existing enormous quantities of data, and look for patterns by using changing data periodically. These patterns and knowledge can be relied upon because they are derived from a large quantity of data [6, 7]. Association rules are one of the representative mining techniques. There is an Apriori Algorithm and an FP-Growth method to search for Association rules. Methods have been suggested which analyze related information from a data stream collected from sensors during a limited time [7, 8, 10, 11, 12]. [10] suggests a method of data Stream Mining for Maximal Frequent Item sets to explore minimum frequent item sets. This method suggests generating a Summary of the Frequent Item set and taking only the data which has a greater frequency than the established support criteria. This method generates summary information from one scan and extracts the most frequent item set. However this method only analyzes Association rules about the one data stream collected from a sensor. [8] suggests a method to investigate event occurrence frequencies. This method calculates the event occurrence support, and investigates the occurrence frequencies, takes out event sequences which satisfy the maximum Support, and predicts the future events. However, these methods only consider periodic intervals of real-time sensing data, but not history data and real-time collecting data from sensors.

3 Proposed System Architecture

3.1 System Architecture for Healthcare Services

In this paper, we propose the multi-agent based Real-time Tele Healthcare System using a JAVA-based application and the JADE, a distributed agent platform. Fig. 1

shows the architecture for a remote diagnosis and prescription system. The system includes three different areas: Body Area Network (BAN) System, Surrogate System, and Hospital System.

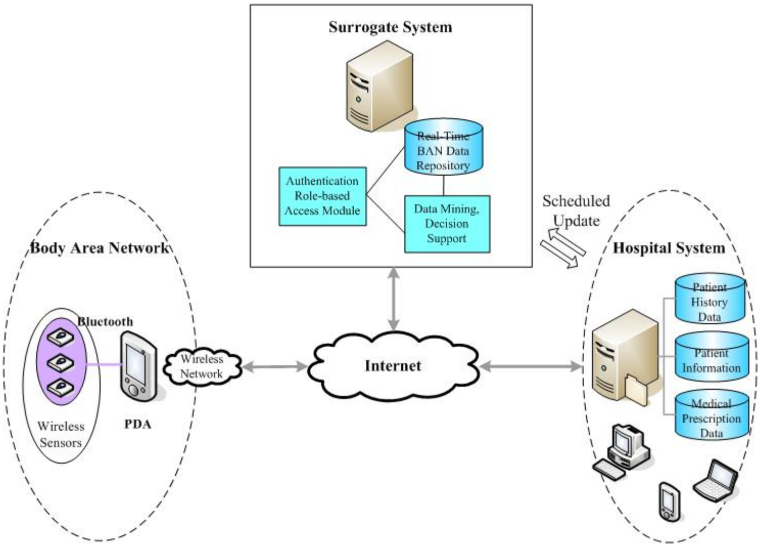


Fig. 1. Healthcare Services Architecture

The BAN system is the body area to which sensors are attached in order to capture biosignals, including blood pressure, body temperature, pulse and breathing. These data are transferred to the PDA, a patient's mobile device, through a wireless network. Sensors are connected with wireless interface boards to transfer the captured medical information. The connection is maintained until an unusual event occurs.

The second area is the surrogate system that serves as a hub between a patient and a hospital. An agent determines whether a patient is in a critical condition based on medical data transferred from the BAN System. If it is determined that there is an emergency, the data is transferred to the hospital system for enacting emergency measures, immediately after being stored in the surrogate system. If it is not an emergency, the data is merely stored in the surrogate system. For data stored in the surrogate system, necessary data is regularly saved to the central database of the hospital. These real-time data will be deleted after a certain period of time unless there is an emergency. This surrogate system connects the BAN system with the hospital sub-system. Data stored in the surrogate system is available to doctors and support staff in the hospital.

The third area is a hospital sub-system. If necessary, data is registered, retrieved, changed, updated and deleted by doctors, patients and hospital support staff.

3.2 The Multi-agent System

The multi-agent system is composed of six main components: patient monitoring agent, gate agent, supervisor agent, manager agent, decision support agent and doctor agent. This system has sensors responsible for acquiring information about surrounding devices and services. Information about a patient such as blood pressure (systolic/diastolic), body temperature, breathing and pulse, is provided to a supervisor agent. At this time, the Gate Agent verifies a patient's authentication of his request for services. The Manager Agent stores patient information in the hospital database and searches for the doctor in charge. In addition, this agent transmits messages of request for diagnosis and prescription from the doctor in charge. The Doctor Agent provides the Patient Monitoring Agent with the diagnosis and prescription for the patient.

3.2.1 Patient Monitoring Agent (PMA)

The Patient Monitoring Agent operates on a mobile device with the following functions: Firstly, it uses sensors to detect medical data from a patient, and peripheral data, including temperature and humidity. Secondly, it transfers the detected data along with details about the sender and device information to a surrogate system via the supervisor agent. Thirdly, it delivers a doctor's observations and diagnosis to his patient via the user interface.

3.2.2 Gate Agent (GA)

The Gate Agent verifies a patient's authentication of his request for services. Patients have different access rights to RBAC, in accordance with various privileges given by their roles.

3.2.3 Supervisor Agent (SA)

The Supervisor Agent operates between the mobile device and the hospital system, controlling the entire surrogate system. Firstly, it receives real-time medical data from a patient including the blood pressure (systolic/diastolic), body temperature, breathing and pulse. It saves the data into a repository, and then uses a specific pattern recognition module to analyze the data and compare it with normal conditions. If the value of data exceeds normal range (threshold), the agent sends an emergency alert message to a doctor or any other person with authority in the hospital via the manager agent, to take the appropriate emergency measures. If the value falls within the normal range (threshold), services will be discontinued when data is saved in a repository.

3.2.4 Manager Agent (MA)

The Manager Agent operates on a hospital sub-system. If the supervisor agent requests emergency measures, it searches for the doctor in charge, and related hospital support staff. This agent sends a message including the patient's historical data and which requests a diagnosis about a patient by the doctor in charge. This agent stores the diagnosis and opinion in the medical prescription database, including a Timestamp and result-id. In addition, this Medical History database stores biosignal information about a patient from the Supervisor Agent. It also manages necessary data for data retrieval, registration and update, and deletion.

3.2.5 Doctor Agent (DA)

The doctor's diagnosis of a patient is aided by messages from the Manager Agent. As well as this diagnosis, the Manager Agent sends an opinion to the patient. This diagnosis and the prescription data is stored in the Medical Prescription database. The stored diagnosis and prescription data are managed and maintained as historical records, which are used when required by patients.

3.2.6 Decision Support Agent (DSA)

In general, data about a patient's normal body temperature, pulse, and blood pressure is not very important even though it is collected frequently. By contrast, data about a patient's abnormal pulse, blood pressure, breathing and body temperature is very important even though it happens less frequently.

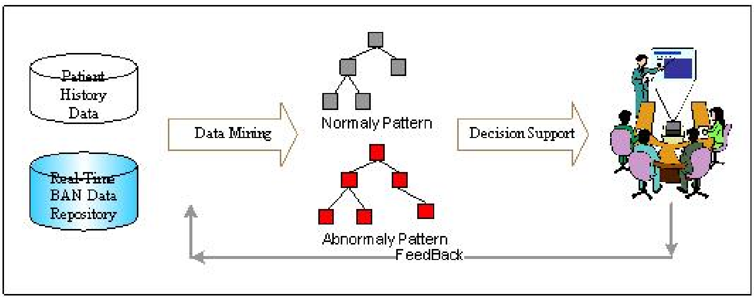


Fig. 2. Decision Support Scheme

Therefore, the Association rule exploration is very important for predicting the occurrence of less frequent, but important information. It is possible a patient with minor symptoms has a body temperature and blood pressure higher or lower than a normal every day range. However, we should not classify this patient condition as an emergency. Therefore, we should separate such data into two groups, historical data and real-time sensory data, then mine the data. By utilizing this, we can determine a patient's fast changing emergency status. Because we mine existing historical data taken from repository storage, we can analyze different time intervals. Fig. 2 shows the decision support scheme for a remote diagnosis and prescription.

In [7], this paper contributes to predicting events expected to follow by analyzing correlations among them collected through a sensor. In spite of such contribution, it tries to find correlations among multi-dimensional events where, among data collected for a given time, only data with more than the minimum support in their frequency of occurrence are chosen. Such support-based methods would lead to a problem of excluding data with a lower level in frequency of occurrence. However, such pieces of information as body temperature, pulse, and blood pressure that demonstrate a normal status of a patient are not so important even if they're frequently collected, while those showing his abnormal status are considered very important in spite of their lower level in frequency of occurrence. Therefore, mining that can predict the occurrence of events with a higher level of importance even if they show a lower level in frequency of occurrence, matters. We find only abnormal events by applying modified mining as shown in [7] to find a relational rule among data items.

Time / Data item	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
A	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
B	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
C	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}

Time / Data item	T_1	T_2	T_5	T_6	T_7	T_8
A	A_1	A_2	A_5	A_6	A_7	A_8
B	B_1	B_2	B_5	B_6	B_7	B_8
C	C_1	C_2	C_5	C_6	C_7	C_8

: normal range value
: abnormal value
minSup = 60%

Fig. 3. An example

Figure 3 shows an example of mining where there happens ten times of sensing on 3 data items collected by a sensor. Since the support ratio in the figure is 60%, [7] can be used to choose only the normal value of data item C as frequent item set (8/10=80%). But we apply our mining method to abnormal events with the value above or short of a normal range. That is, if the first support ratio is applied to T_1, T_2, T_5, T_6, T_7 and T_8, among their individual data item, abnormal, abnormal and normal for A, B and C, respectively, are picked up as frequent item set, which, in turn, frequent data pattern $\{(A_{ab}, B_{ab}, *),(A_{ab}, *, C_{nor})\}$ is selected.

4 Implementation

In this paper, we demonstrated the implementation of a multi-agent based Real-time Tele Healthcare System with the suggested architectures, such as the JADE, LEAP, J2SE, and PersonalJava. The Oracle database was used to store patient information, diagnosis and prescription and as a surrogate clinical repository.

We choose N (N=4, in this case) patients (students), and saved one month patient's sensor data to the medical historical database. It is not easy to recognize unusual diseases from a patient's sensory data. In order to achieve accurate recognition, we must consider all possible cases and use knowledge and medical diagnosis provided by a professional doctor. Our mining method uses the method from [7]. The standard of decision-making in our paper is based on the comparison between a normal pattern of patient history and the current sensory data for each category.

The following is pseudo code for decisions about patient disease states.

```
/* present disease state of a patient is being judged according to
   whether it is unusual or not
Input:  Patient historical data pattern(d1), present sensory data pattern(d2)
Output: state decision (special state(true), normalcy(false))
*/
for each data item i ∈ patient historical data
{
    If ( i ∈ normal range) then
       Add data item i to normal-queue;
    Else
       Add data item i to abnormal-queue
}
for each data item d ∈ abnormal-queue
    delta = | d2 – d1 |
    if ( delta ≥ threshold )
    then  return (true);   // emergency measure
    else  return (false);   // normal state
}
```

Fig. 4 shows the sequence diagram of remote medical services based on ACL message transfers among agents. Body temperature, breathing and pulse are measured every one second, the blood pressure every fifteen minutes. All collected medical data is transferred to a mobile device.

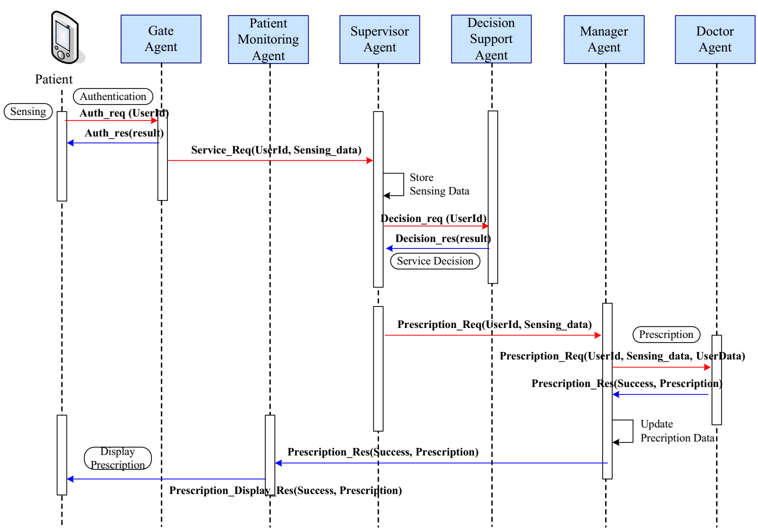


Fig. 4. Sequence Diagram of Multi-Agent based System

Fig. 5 shows the prototype of the implemented healthcare system. Fig. 5 (a) is the user interface for patient diagnosis and prescription via a PDA and Fig. 5 (b) is a Graph that could record both changes in patient temperature and in the surrounding temperature.

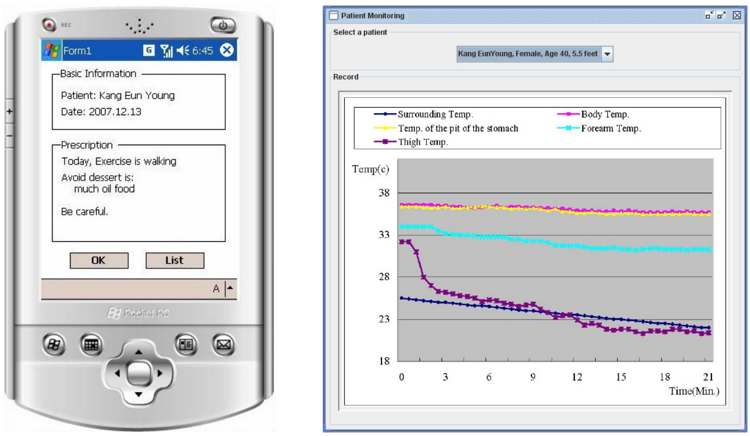


Fig. 5. (a) Patient's diagnosis and prescription (b) Result of Patient monitoring

5 Conclusions

This paper proposed a ubiquitous healthcare system by modeling real-time diagnosis and prescription services provided by a hospital system, based on collected medical and peripheral data. The proposed system provides an interconnection of patients and a hospital in a ubiquitous computing environment. In our multi-agent system, agents are involved in functions, such as using sensors to collect medical and peripheral data in real-time, storing the collected data in the surrogate system, determining whether a patient is in a critical condition, transferring to the hospital system data about the patient that has been determined to be critical, and finally delivering the doctor's diagnoses and prescriptions to the patient. However, determination of the patient's condition in medical terms based on the collected data requires further investigation. Introducing intelligent agent technologies to the context-aware health care system proposed in this paper is expected to create good value-added ubiquitous services.

References

1. Dokovsky, N., van Halteren, A., Widya, I.: BANip: enabling remote healthcare monitoring with Body Area Networks. In: International Workshop on scientific engineering of Distributed Java applications, pp. 27–28 (2003)
2. Camarinha-Matos, L., Afsarmanesh, H.: Virtual communities and elderly support. In: MIV 2001 in Advances in Automation, Multimedia and Video Systems, and Modern Computer Science, WSES, pp. 227–284 (2001)

3. Kim, S.: Ubiquitous Healthcare: The OnkoNet Mobile Agents Architecture. NODe, 265–277 (2003)
4. Signh, S., Ismail, B.I., Haron, F., Yong, C.H.: Architecture of Agent-Based Healthcare Intelligent Assistant on Grid Environment. In: Liew, K.-M., Shen, H., See, S., Cai, W. (eds.) PDCAT 2004. LNCS, vol. 3320, pp. 58–61. Springer, Heidelberg (2004)
5. Brones, T.: Supporting the Developers of Context-Aware Mobile Telemedicine Applications. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM-WS 2005. LNCS, vol. 3762, pp. 761–770. Springer, Heidelberg (2005)
6. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach. IEEE Transactions on Knowledge and Data Engineering 16(11) (2004)
7. Chen, G., Wu, X., Zhu, X.: Mining Sequential Patterns Across Data Streams. Univ. of Vermont Computer Science Technical Report(CS-05-04) (March 2005)
8. Kuramitsu, K.: Finding Periodic Outliers over a Monogenetic Event System. In: Proc. of UDM 2005, pp. 97–104 (April 2005)
9. Li, H., Lee, S., Shan, M.: Online Mining (Recently) Maximal Frequent Itemsets over Data Streams. In: Proc. ORIDE-SDMA 2005, pp. 11–18 (April 2005)
10. Han, H., Ryoo, H., Patrick, H.: An Infrastructure of Stream Data Mining, Fusion and Management for Monitored Patients. In: Proc. of 19th IEEE International Symposium on CBMS 2006, pp. 461–468 (June 2006)
11. Park, K.J., Ryou, H.: Anomaly Detection Scheme Using Data Mining in Mobile Environment. In: Kumar, V., Gavrilova, M.L., Tan, C.J.K., L'Ecuyer, P. (eds.) ICCSA 2003. LNCS, vol. 2668, pp. 21–30. Springer, Heidelberg (2003)
12. Lv, T.-Y., Su, T.-X., Wang, Z.-X.: An Auto-stopped Hierarchical Clustering Algorithm Integrating Outlier Detection Algorithm. In: Fan, W., Wu, Z., Yang, J. (eds.) WAIM 2005. LNCS, vol. 3739, pp. 464–474. Springer, Heidelberg (2005)
13. `http://www.fipa.org`
14. `http://leap.crm-paris.com`

A Multi-agent Mechanism in Machine Learning Approach to Anti-virus System

Minh Nhat Quang Truong[1] and Trong Nghia Hoang[2]

[1] Cantho In-service University, Vietnam
[2] Vietnam National University HCM City, Vietnam

Abstract: In this paper, we would like to introduce a multi-agent mechanism to protect target systems from computer virus infections. Using a machine learning approach, we first define a form of object knowledge to determine computer viruses and diagnosed objects. Second, we set up an association model of knowledge base and database. The database stores information of diagnosed objects. The knowledge base contains certain sets of deduction rules. Finally, we build two active agents to control virus infections. In an event-learning model, the first agent named Virus Auto-protect Agent is used to monitor all suspicious events. The second one, Virus Scanning Agent is used in an explanation-learning model to scan for viruses, to warn users of dangers and to restore the data from the previous state of safety. The experimentation results show that the anti-virus system can quickly recognize known and unknown computer virus infections.

Keywords: Anti-virus, Computer virus, Machine Learning, Multi-agent.

1 Introduction

Nowadays, computer viruses have become very widespread and they influence the data security of many computer systems. Therefore, it is necessary to improve the methods of identifying computer viruses. Our research aims to find a new approach for anti-virus technology. So we have set up a project called Machine Learning Approach to Anti-virus System (MAV). In this paper, we would like to introduce a multi-agent mechanism to protect target systems from both known and unknown computer virus infections.

First of all, we define a form of object knowledge to determine computer viruses and diagnosed objects. Second, we set up an Association Model of Knowledge Base and Database (AMKBD) to identify computer viruses. Like a medical reference book, the database stores information of diagnosed objects. The knowledge base contains certain deduction rules that describe threats to the target system. Finally, a multi-agent mechanism is used to control virus infections. Using an event-learning model, a Virus Auto-protect Agent (VAA) monitors any suspicious activity such as opening an executable file to write. With an explanation-learning model, a Virus Scanning Agent (VSA) learns to verify and explain the dangers stated above. If a virus is found, the system data will be restored from the previous state of safety.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 743–752, 2008.

2 Computer Virus and Anti-virus Technology

A computer virus is “any instruction, information, data or program that degrades the performance of a computer resource; disables, damages or destroys a computer resource; or attaches itself to another computer resource and executes when the host computer program is executed” [1].

Anti-virus technology researches the methods to recognize computer viruses and produce anti-virus programs (AV) to eliminate malicious codes from target systems [2]. Anti-virus technology has gone through three evolutions in methods of identifying computer viruses: signature-based approach, suspicious behavior-based approach and intention-based approach.

A signature-based anti-virus program can identify most known computer viruses from a dataset by using a database containing information of viruses. However, such anti-virus program can fail when lacking virus information. Therefore, many anti-virus programs cannot recognize new computer viruses [3].

Instead of using viral signatures, a suspicious behavior-based anti-virus program monitors the behaviors of all programs [4]. When a suspicious behavior is detected, for example, a program attempts to write to an executable file, the anti-virus program flags this event and prompts the user for action [5]. However, this approach cannot distinguish some similar actions although they have different purposes. For example, user’s programs sometimes create and delete temporary files, but viruses copy themselves and delete user’s files.

Improving on the previous ones, an intention-based anti-virus technology like Rudra (Sanrasoft, India) uses an ‘image’ of system data. When any serious are identified, such anti-virus programs try to restore the original data for the target system. This ‘return to the past’ technique can protect the target system from crashing for unknown reasons; possibly due to computer viruses. However, this approach needs a lot of system resources to manage the system information. In addition, it also requires a flexible mechanism to guard all suspicious events [6]. So, we use a multi-agent mechanism to solve this problem.

3 Machine Learning Approach to Anti-virus System

To improve the identifying methods of an anti-virus system, we create a conventional anti-virus program that utilizes a machine learning rule-based approach that involves these basic tasks: modeling a knowledge base, forming rule sets, discovering interesting rules from database using machine learning algorithms and applying rule-based diagnostics to recognize viral code from the target system [7].

There are many types of computer viruses, and each type has its own way of infecting the computer. In our approach, each computer virus class is defined with its own particular characteristics. A standard virus class has an object-oriented form:

Object: Virus class identification
Property: Attributes/behavior
Method: Treatment/direction outline

Analyzing computer viruses, we define five virus classes corresponding to widespread virus types (Table 1). Using an ID-virus library that contains updated virus characteristics (signatures, behaviors), MAV needs to be built with the set V_K of K vectors-virus samples $V_K = \{v_1, v_2, \ldots, v_k\}$ and to determine the existence of v_i in the dataset S. The conventional rule set has the form:

$$R: p_1 \wedge p_2 \wedge \ldots \wedge p_n \Rightarrow q$$

where p_i represents the virus characteristics and q is the result of the process [8].

Table 1. Five computer virus classes in MAV system

No.	Virus class	Description	Virus type
1	A-class	Application	Executable worm, Trojan horse
2	B-class	Boot record	Boot virus
3	C-class	ASCII text	Script worm, Text file virus
4	D-class	Document macro	Macro virus
5	E-class	Executable	Executable file virus (F-virus)

In this paper, we would like to present an event-explanation learning technique using AMKBD [9] to diagnose the virus file for E-class described as follows.

4 The Event-Explanation Learning Model in MAV System

The term "F-virus" denotes the kind of computer viruses that attach to executable files (called virus's subjects) of the operation system. If a file has been infected by an F-virus, the file's properties will change [10]. In our research, these subjects are PE (Portable Executable) files on Windows32 environment.

4.1 The Knowledge Base Organization

Analyzing the F-virus infection behaviors [11], we achieve two rule sets RS1 and RS2. The first set describes the events of instant infections, and the second explains the permanent changes of the subjects (Table 2). Using an IF-THEN rule editor, we build the knowledge base for MAV system.

4.2 The Database Organization

To supply the subject information for deduction rules, we build a database called *VerifyDB* for the system image as follows:

SUBJECT(**SUBJ_ID**, NAME, ADDRESS, TYPE, SIZE, ATTRIBUTE, DATETIME)
VERSION(**VER_ID**, RELEASE_DATE, VIRUS_COUNT)
HISTORY(**#VER_ID**, **#SUBJ_ID**, SCAN_DATE)

where SUBJ_ID and VER_ID are primary keys of relatives SUBJECT and VERSION. The foreign keys VER_ID and SUBJ_ID function as a primary key of HISTORY relative.

Table 2. Two sets of rules to recognize F-viruses

RS	Rule	Expression	
RS1	Rule 1	IF	*OpenFileToWrite* event happens
		THEN	File will be changed
	Rule 2	IF	*SetFileAttr* event happens
		THEN	File will be changed
	Rule 3	IF	*SetFileDate* event happens
		THEN	File will be changed
	Rule 4	IF	*GetFileAttr* event happens
		AND	NOT *FileChanged*
		THEN	File may be corrupted
	Rule 5	IF	*OpenFileToWrite* event happens
		AND	NOT *FileChanged*
		THEN	File may be corrupted
	Rule 6	IF	*GetFileDate* event happens
		AND	NOT *FileChanged*
		THEN	File may be corrupted
RS2	Rule 7	IF	*FileAttribute* was changed
		THEN	File was changed
	Rule 8	IF	*FileDateTime* was changed
		THEN	File was changed
	Rule 9	IF	*FileSize* was changed
		THEN	File was changed

4.3 The Inference Motor

The strategy of the discovery knowledge from the database is as follows:

- At first, MAV diagnoses the data set and updates all tested subject information into the 'reference booklet' *VerifyDB*.
- Permanently, MAV monitors whether the subject is accessed or not to control any virus behaviors.
- When a new subject attempts to access the target system, MAV will test it beforehand to prevent a virus from spreading.
- If an 'epidemic disease' appears, MAV does not need to retest all diseases, but to scan for only 'new diseases'.

Hence, we build an algorithm named SID (Searching-Inferring-Diagnosing) to recognize F-virus infections for target systems. For a data point:

- SID searches data point information in the database.
- If not found, SID tests the data point and updates the system image.
- Or else, SID needs to verify the data point for three test cases:

  Case 1: subject information of two present continuous testing times is different, with the same testing version program. SID warns about subject changes to confirm one of three events:
  - By virus: reporting, restoring the data point information from the database.
  - By user: updating the database, increasing the knowledge base.
  - Unknown: locking the subject, reporting, increasing the knowledge base.

Case 2: subject information of two present continuous testing times is the same, with different testing versions. SID only tests the subject for the last diseases. If it is found to be a virus, the subject is processed as a new one and its information is updated.
Case 3: subject information of two present continuous testing times is different, with different testing versions. SID only tests the subject for the last diseases. If it is found to be a virus, the subject is processed as in Case 2. Otherwise, it is processed as in Case 1.

5 The Multi-agent Mechanism in MAV System

From our study, RS1 rules are effective for a short time (just a few milliseconds of a system process depending on CPU speed), while RS2 rules are effective for a long time (a few seconds or even a month depending on the spread of viruses and the frequency of system usage). System messages draw certain events described in RS1 rules that are 'instant knowledge'. Such knowledge is easily covered by other system events. In contrast, the subject's change in RS2 rules belongs to 'permanent knowledge'. Our solution is to build two active agents VAA (Virus Auto-protect Agent) and VSA (Virus Scanning Agent) to control these system events.

5.1 The Virus Scanning Agent

Designed to work in a foreground mode, VSA exploits RS2 rules to acquire permanent knowledge of the system, to scan for viruses and to explain all damage (Figure 1). VSA has some tasks, such as:

1. Initializing working parameters, memory variables and user settings such as program factors/constants, true/false options…
2. Loading data points in need to test in an observation space.
3. Using NNSRM (Nearest Neighbor Rule-based Structural Risk Minimization) to classify observation space into two classes: CLASS_1 – possibly infected and CLASS_2 – absolutely pure data points, then removing CLASS_2 to reduce the data points to test [12].
4. Running the SID algorithm for CLASS_1 to scan for viruses.
5. Reporting scanning results, making interactive dialogues, acquiring user knowledge and eliminating viral codes for CLASS_1.
6. Updating and increasing the knowledge base and database for the system.

In these tasks, the 3rd and 4th consume a lot of system resources (running time, memory, disk space…). To distribute their tasks, we build two sub-agents for VSA using two program threads: CA (Classifying Agent) and SA (Scanning Agent). In this model, an FIFO queue is used as a shared memory zone for CA and SA. Their work is described as follows:

- The queue is declared as a dynamic array; therefore, it never overflows to ensure the CA finishes its work normally.
- Using NNSRM, the CA classifies the observation space, and then inserts each data point from CLASS_1 to the rear of the queue.

- While the CA is filling the queue, the SA tries to empty it by picking out each data point from the queue's front, then passing the point to SID procedure.
- If the queue is empty while the CA is still running, the SA waits until a data point appears in the queue.

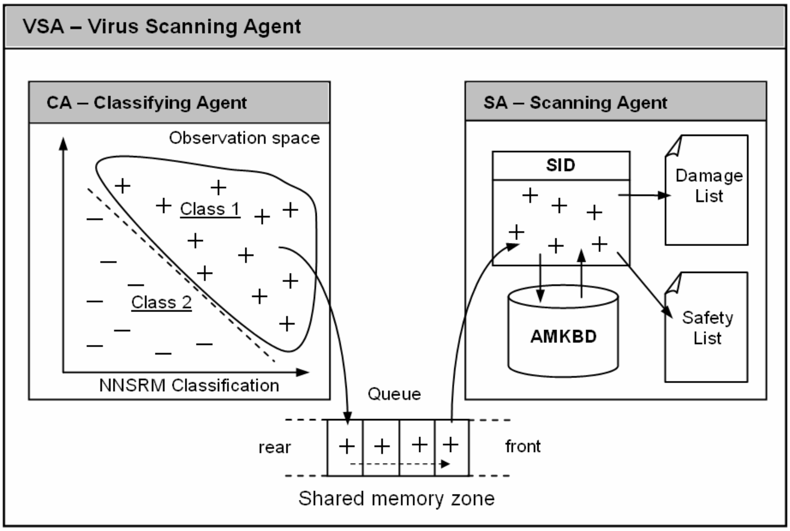


Fig. 1. The Virus Scanning Agent with two sub-agents

5.2 The Virus Auto-protect Agent

Different from VSA, VAA is built to run in a background mode for monitoring all system events. To acquire this instant knowledge, the VAA does the following:

1. Activates automatically when Windows starts.
2. Creates dynamic threads to distribute the VAA's tasks for mobile sub-agents.
3. Hooks file/folder/registry system API functions such as Create/OpenFile, Get/SetFileAttribute, Get/SetFileDate, Read/WriteFile, Close/DeleteFile, etc.
4. Monitors all running system processes such as creating a process, loading a dynamic link library, exploring storage devices, accessing a network, etc.
5. Downloads the database and knowledge base from the MAV remote server.

In our design, VAA's sub-agents are assigned to work autonomously at suitable positions in the target system. After being created, a sub-agent moves to system folders to guard Windows' events, while some others move to system gateways to supervise network accessions. Using SID, VAA flags and analyzes all system events by exploiting RS1 rules to alert users (Figure 2). This feature can also automatically eliminate dangerous events, without giving warnings on the graphic user interface.

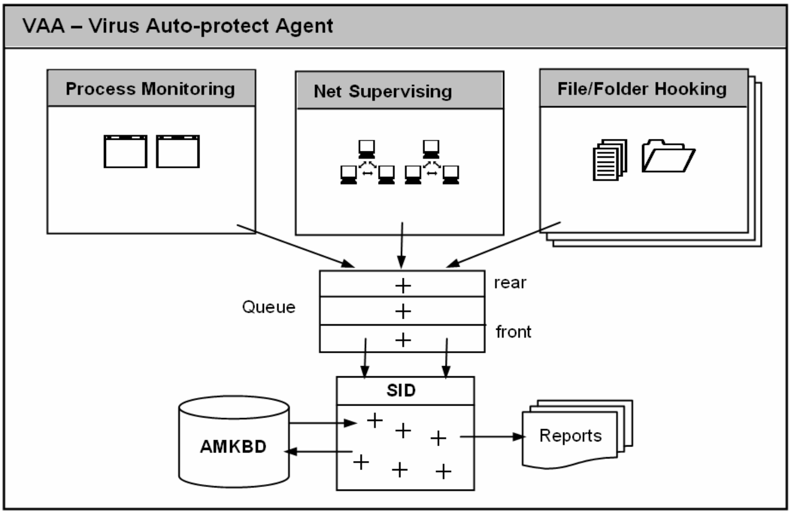


Fig. 2. The Virus Auto-protect Agent with mobile sub-agents

5.3 Co-operation of VSA and VAA

Although they use the same database, knowledge base and inference motor SID, the active agents VSA and VAA work independently and closely interact in the mode of process communication. For example, before writing to an infected PE file, VSA sends a message to require VAA not to guard this event. Similarly, while diagnosing a subject, VAA warns VSA of a local test. This interaction strategy helps the system to avoid conflicts.

6 Experiments

Using Borland Delphi 7.0, we build MAV with two program modules: VSA.EXE in an application mode and VAA.EXE in a service mode. Also, a dynamic link library named AMKBD.DLL is built to control virus infections through the *VerifyDB* database and SID algorithm. To evaluate the MAV results from other products, we experiment on program speed and effectiveness.

6.1 Testing MAV Speed

With a computer (Celeron 2.4 GHz CPU, 512 MB RAM) running on Windows XP for a dataset 8,000,000 KB of 56,100 files, we test program speed by:

- Running MAV without using SID (called PreMAV) many times and measuring the software runtime cost.
- Running MAV using SID many times and measuring the software runtime cost.
- Running other AVs many times and measuring the software runtime cost.

The AVs tested are Norton Anti-virus (NAV-Symantec, USA), VirusScan (McAfee, USA) and BKAV (BKIS, Vietnam). Table 3 shows the results of our study. Analyzing the charts of the runtime cost of testing programs (Figure 3), we see:

- Except MAV, the runtime cost charts of testing programs (including PreMAV) are nearly straight. It denotes that these anti-virus programs do not have learning ability because they spend a constant cost for a tested data set.
- The chart of MAV decreases strongly after learning the first time, then it is nearly straight. It means MAV is learning gradually stabilizes.

Using AMKBD multi-agent mechanism, MAV spends a runtime cost to search and learn activities. This effort requires 838-772 = 66 (seconds) with a rate of 8.55% on time cost of PreMAV. However, this cost is valuable because it helps MAV increase a rate of 1227.325 % runtime cost for the next four running times:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\text{Cost}_i\text{ (PreMAV)}}{\text{Cost}_i\text{ (MAV)}} = \frac{1}{4}\left(\frac{769}{74}+\frac{779}{61}+\frac{764}{59}+\frac{766}{59}\right) = 1227.325\ \%$$

Table 3. Testing anti-viruses runtime cost

Anti-virus	Time 1 (s)	Time 2 (s)	Time 3 (s)	Time 4 (s)	Time 5 (s)
BKAV	245	248	247	248	246
PreMAV	772	769	779	764	766
MAV	838	74	61	59	59
NAV	1365	1347	1315	1317	1320
SCAN	1506	1450	1462	1488	1475

6.2 Testing MAV Effectiveness

In this experiment, anti-virus programs tested are MAV, NAV, SCAN and Bit Defender (Softwin, Romania). Table 4 is the diagnostic result for a dataset of 1,000 viruses. Figure 4 expresses the test results. In their detection ability, MAV and Bit Defender have the same results at 957 and 959 viruses. Both these anti-viruses are better than NAV and Virus Scan at 907 and 906 viruses. With a few virus samples (890 viruses), MAV can still recognize as many viruses as other anti-virus programs, which have much bigger virus databases (of 72,020 and 253,993 viruses).

Table 4. Testing anti-viruses effectiveness

Anti-virus	Signatures	Version	Detection	Precision	Prediction	Omission
Norton Antivirus	72,020	9.05.15	907	889	18	93
Virus Scan	N/A	4.0.4682	906	877	29	94
Bit Defender v.8	253,993	7.05450	959	925	34	41
MAV	890	N/A	957	483	474	43

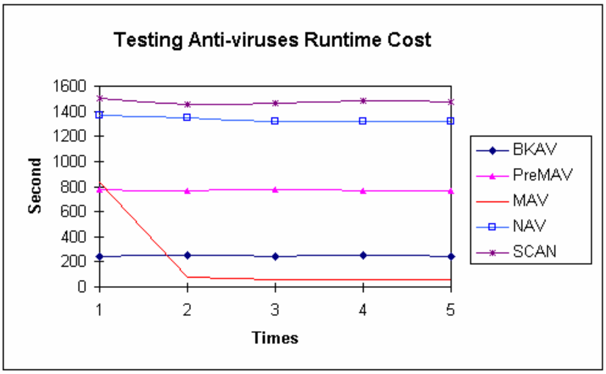


Fig. 3. The charts of testing Anti-viruses runtime cost

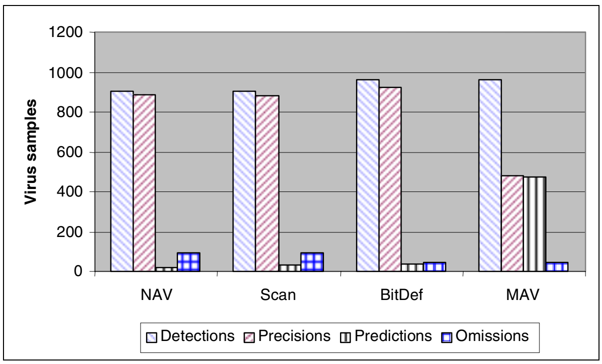


Fig. 4. Comparison of smart anti-viruses scanning

7 Conclusion and Future Research Work

Using the AMKBD multi-agent mechanism, the MAV not only detects computer virus infections but also improves program speed. However, when any software is updated, such a program may make inexperienced users worried. In addition, the program can have some difficulties with an infected system, where knowledge acquisition has not been performed. To solve these problems, we will combine different techniques of machine learning for the MAV system.

Acknowledgement

We would like to acknowledge Prof. Dr. Hoang Kiem (Vietnam National University, Ho Chi Minh City), who has helped us to finish this paper.

References

1. Bordera, M.: The Computer Virus War. Is The Legal System Fighting or Surrendering? Computer and Law, University of Buffalo School of Law (1997)
2. Bontchev, V.: Methodology of Computer Anti-Virus Research. Doctoral Thesis, University of Hamburg (1998)
3. Thuy, N.T., Quang, T.M.N.: Expert System Approach to Diagnosing and Destroying Unknown Computer Viruses. In: Proceedings of the IASTED International Conference Artificial Intelligence and Soft Computing, May 1998, pp. 371–374. IASTED/ACTA Press, Cancun-Mexico (1998)
4. Thuy, N.T., Quang, T.M.N.: A Global Solution to Anti-virus Systems. In: Proceedings of the 1st International Conference on Advanced Communication Technology, Muju-Korea, February 1999, pp. 374–377 (1999)
5. Szor, P.: The art of computer virus research and defense. ch. 11. Behavior Blocking. Addison Wesley Professional Press, Reading (2005)
6. Rabaiotti, J.: Counter Intrusion Software. PhD. Thesis, Computer Science, Cardiff University, pp. 38–43 (2007)
7. Kiem, H., Thuy, N.T., Quang, T.M.N.: A Machine Learning Approach to Anti-virus System. In: Proceedings of Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICS-IPSJ and IEICE-SIGAI on Active Mining, Hanoi-Vietnam, December 2004, pp. 61–65 (2004) (ISSN: 0913-5685)
8. Kiem, H., Thuy, N.T., Quang, T.M.N.: Using Null Data Processing to recognize variant computer viruses for Rule-based Anti-virus systems. In: Proceedings of the IEEE International Conference on Granular Computing. Atlanta-USA, May 2006, pp. 600–603 (2006)
9. Quang, T.M.N., Kiem, H.V., Thuy, N.T.: Association Model of Knowledge Base and Database in Machine Learning Anti-virus System. In: Proceedings of the WMSCI 2006 Conference, Florida-USA, pp. 277–282 (2006)
10. Arnold, W., Tesauro, G.: Automatically Generated Win32 Heuristic Virus Detection. In: Proceedings of the 2000 International Virus Bulletin Conference (2000)
11. Schultz, M.G., Eskin, E., Zadok, E., Stolfo, S.J.: Data Mining Methods for Detection of New Malicious Executables. In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA (May 2001)
12. Kiem, H., Thuy, N.T., Quang, T.M.N.: Machine Leaning Approach to Anti-virus Expert System with Nearest Neighbor Rule-based Structural Risk Minimization. In: Proceedings of the RIVF'05, the International Conference in Computer Science, Cantho, Vietnam, February 2005, pp. 295–298 (2005)

Data-Mining Model Based on Multi-agent for the Intelligent Distributed Framework*

Romeo Mark A. Mateo, Insook Yoon, and Jaewan Lee

School of Electronic and Information Engineering, Kunsan National University
68 Miryong-dong, Kunsan, Chonbuk 573-701, South Korea
{rmmateo, jwlee}@kunsan.ac.kr

Abstract. Recent researches in distributed system include intelligent resource finding, dynamic replication and adaptive load balancing schemes which focus on improving specific technique. In this paper, an intelligent distributed framework is presented to address the use of intelligent models for adaptive distributed object groups. Moreover, this paper proposes the agent-based data-mining model for implementing adaptive schemes using data mining algorithms and efficient interactions using multi-agent system. The *k*-means algorithm constructs group classes of object, multilayer perceptron classifies the client requests using the classes constructed from *k*-means and patterns generated from Apriori algorithm determine the next object needed to be replicated. For efficient interactions, the data mining is modeled in multi-agent system. Simulation result using the proposed model shows great improvements on serving clients by minimizing delay time and optimizes system performance by efficient load distribution.

1 Introduction

The current emerging technologies like ubiquitous, mobile and peer-to-peer systems is part of distributed system. Most concepts and techniques of distributed system are legacy to current technologies which consider the efficient communication of components and coordination of actions in the computer networks. The main goal of distributed system is to share effectively data or resources within the network. There are challenges for the designer of distributed systems to effectively share data which are heterogeneity of its components, scalability of resources and users, failure handling, concurrency of components and transparency of the services [1]. The current researches in distributed system focus on efficient search of objects [2,3], object group modeling [4], dynamic replications [5] and load distribution of objects [6,7]. The effective management of resources is critical for coordination of the distributed system components. However, previous researches are lack of knowledge-based methods which provides adaptive system and optimizes the system performance.

* This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 753–762, 2008.

Data mining has attracted a great deal of attention in information industry and in society due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [8]. Data mining techniques are designed to reveal the patterns from data by classifying, clustering or associating methods. These methods are used to extract rules from the data transactions and use these rules to adapt the new constraints of a system. Efficient interaction of the components from the system is necessary for effective adaptive system.

Multi-agent approaches are mostly used by application which requires of distributed task management. The effectiveness of the coordination and communication is considered [8]. Recently, various researchers and industries use software agents to implement intelligent systems [9,10]. Current researches in agent technology focus on providing industries with a new problem solving approaches in a distributed manner, new software tools and automated functions of the system [11]. In multi-agent systems, communications, resource sharing and cooperation for solving problems are considered. Many organization and researcher are working on standards like in FIPA [12] and other technologies of multi-agent system. Also, there are still lots of effort to be done by researchers in contributing new ideas in multi-agent system.

This paper proposes the agent-based data-mining model for the intelligent distributed framework. The data mining model of the intelligent distributed framework is conceptualized from object group models as resources, data mining models as rule generation for adaptive scheme and multi-agent systems as the basis of interaction of each component for efficient communication and coordination. The data mining model uses the clustering, classification and association methods for object grouping, intelligent searching and dynamic replication schemes, respectively, in implementing intelligent adaptive schemes. The proposed data mining model is based on multi-agent for efficient interaction of the components.

2 Research Background

Data mining methods are used to reveal the patterns from transaction data. Mostly, system uses this technique to optimize the configuration base on the new rules generated by data mining. However, modeling the data mining that can be used in the system is a challenge. The data mining functionalities are used to specify the kind of patterns to be found in data mining task [8]. There are three well-known data mining functionalities which are association, classification and clustering mining which is appropriate in implementing adaptive schemes for the proposed intelligent distributed framework. Classification method is a form of data analysis that can be used to extract models describing important data classes or predict future data trends. If the structure is a well modeled for searching, then the method can be efficient to be used by the clients in searching resources. Clustering methods are unsupervised methods which analyzes data without classes. At the initial procedure, the class labels are unknown and the whole procedure helps determine the class label. This method is appropriate on arranging the objects based on properties without apriori knowledge. Association method determines frequent patterns from the data transaction where these frequent patterns analyzed as the trend of the events or transactions. The issue is when does a replication of object is needed and where should the replication takes place. Association rule mining can be used to predict the needed object to be replicated in the system.

Current researches in agent technology focus on providing industries with a new approach of solving problems in a distributed manner, new software tools and automated functions of the system [13]. Agents are provided with intelligence based on information acquisition and rules to function its autonomy in delivering the task. In able to understand the interaction of multi-agents, modeling actions and behaviors of agents are necessary to be studied [9]. The observed phenomenon is translated into form of an abstraction, which can be manipulated through algebraic transformation, numerical computations or simply logical expression to obtain the result which can help understanding or predict the future situations. The advantage of action and behavior modeling is that it can be more explicit, simpler and easier to manipulate.

3 The Intelligent Distributed Framework

Intelligent models are mostly topics of research where recent approaches are solutions to the previous problems experienced in classical systems and tackle the issue of solving future constraints. In distributed systems, concurrency, transparency and scalability are challenging topics to researchers and developers in providing effectiveness of sharing data and resources. This study presents the intelligent distributed framework which implements scalable system and addresses the use of intelligent approaches in distributed environment. The proposed framework is shown in Figure 1.

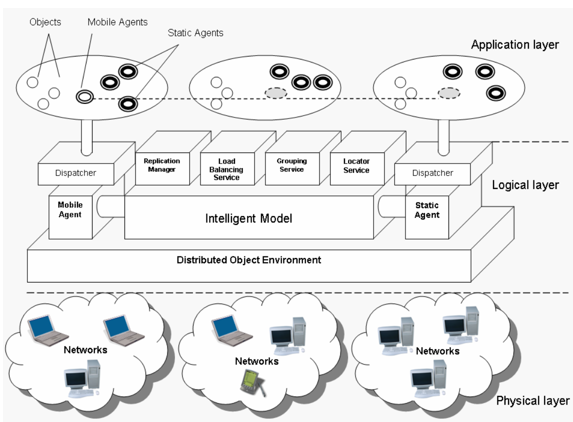


Fig. 1. The intelligent distributed framework

This consists of three layers of software and hardware components: physical, logical and application layers. The proposed system uses multi-agents to implement automation of tasks and distributes task in managing transparently the clients and servers. The physical layer represents networks of different computers, like PC, laptop and personal digital assistants (PDA). The logical layer acts as the middleware where services are transparently operating on serving the clients. Interaction of clients and object services are handled by the logical layer. Users and administrators do not need to know the configuration on how to find, where to find and how to manage the resources but transparently executes the services. Multi-agent components are working for resource finding and optimizing the system performance. Also, the intelligent algorithms are applied in these layers. The application layer is consisted of objects and agents utilizing the distributed environment. The schemes in object group models are improved by the proposed intelligent distributed framework on the following points:

- Intelligent object grouping: Object group binding makes other objects accessible within the groups and enhances the search for objects. Implementing intelligent grouping like data mining algorithms are used for object grouping in the intelligent distributed framework. These techniques form a structure of knowledge that can be easily analyzed and provides ease of search.
- Intelligent search with adaptive load balancing scheme: Locator service, using mobile agents, classifies the request content to choose the appropriate object. In this paper, the intelligent search mechanism considers the adaptive load balancing where after searching for the appropriate object, the loads are weighted to determine which server is the least loaded and forward the request.
- Dynamic replication scheme based on prediction. Replica managers coordinate to other RS on managing the replicated objects. The replication occurs based on the prediction which is the next object is needed to be replicated.

4 Data Mining Model Using Multi-agent System

The proposed data mining model uses the three functions of data mining [8] which are the clustering, classification and association methods to implement knowledge discovery and adapting new system requirements which is based on multi-agent system.

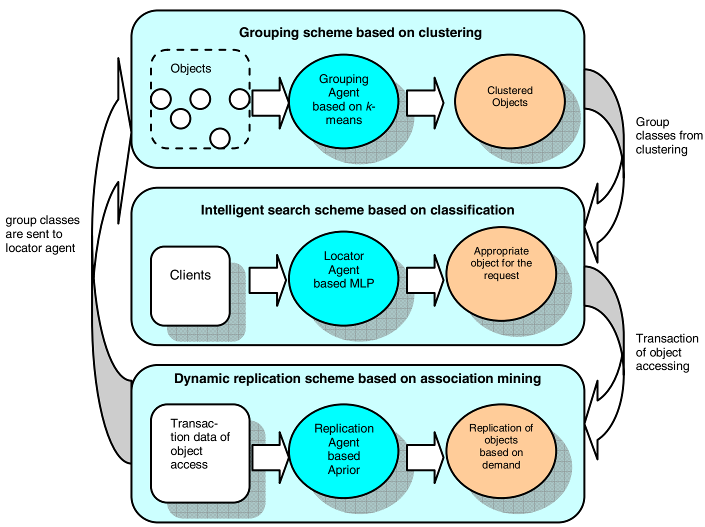


Fig. 2. Data mining model for the intelligent distributed framework based on the communication and coordination of multi-agent

In the first function, the grouping agent uses *k*-means algorithm to cluster the object and then assigns mobile agents to each group which will be used for the intelligent search scheme. The second function is classifying the clients request based on multilayer preceptrons (MLP). This implements the intelligent search with adaptive load balancing. The third function is dynamic replication based on prediction using the patterns generated in association mining. The patterns from the current client transactions are compared to the generated rule in choosing the appropriate pattern for prediction. These interactions of the adaptive schemes are based on multi-agents. The subsections are details of the methods.

4.1 Grouping Scheme Based on Clustering

Cluster analysis divides data into groups such that similar data objects belong to the same cluster and dissimilar data objects to different clusters [14]. Partitioning methods construct *c* partition of data, where each partition represents a cluster. The *k*-means algorithm is one of the popular data mining techniques that clusters data based on minimizing the value from the objective function in Equation 1. In clustering objects in the intelligent distributed framework, we used the computation of the *k*-means.

$$clustering = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\left(\sum_{k=1,\mathbf{u}_k \in C_i} m_{ik}\left\|\mathbf{u}_k - \mathbf{c}_i\right\|^2\right) \quad (1)$$

$$c_i = \{c_1, c_2, \ldots, c_k\} \quad (2)$$

In Equation 1, J_i is the objective function within cluster *i*. The partitioned clusters are typically defined by a $c \times K$ binary characteristic matrix M, called the membership matrix, where each element m_{ik} is 1 if the *k*th data point u_k belongs to cluster i, and 0 otherwise. In choosing the group of object, the property of an object which has the highest value is selected and set it as one of the centers shown in Equation 2. The procedure iterates until a minimum value from *J* is achieved based on the constraint. The basis of object grouping of the proposed clustering is the dominating object which has the highest *k*th property. The clustered objects provide the initial classes which are used for the MLP implemented by intelligent search mechanism. Also, the grouping agent assigns mobile agents to represents the object group.

4.2 Intelligent Search Scheme Based on Classification

The multilayer perceptron is used in data mining for classification techniques. In our proposed model, the MLP is consisted of three layers of processing nodes an input layer which accepts the input variables used in the classification procedure, one hidden layers, and an output layer with one node per class. The data from grouping agent are used to construct the structure of MLP. Every time a request is occurred, the locator agent sends message to the agents containing inputs of the request to mobile agents and processes the MLP-like classification. Equation 3 is the function of the choosing the candidate agent to classify the inputs. A_n are agents *x* are input parameters where the agent that has the maximum value from the aggregation is the candidate agent.

$$classification = candidate\_agent = \max\{\sum_{n=1}^{N} A_n(x)\} \tag{3}$$

$$A_n(x) = \sum_{j=1}^{J} x_j w_{nj} \tag{4}$$

$$object = \min(\sum_{l=1}^{L} objectload_l) \tag{5}$$

Moreover, the processing function of a single agent is presented in Equation 4. The output value from processing the weight in the MLP is sent back to the locator agent. The chosen candidate agent also means that it has the appropriate object for the client requests. The request is sent to the candidate agent and finds the appropriate object using the adaptive load distribution based on round robin and least load algorithm. The object that has the minimum load is the chosen object shown in Equation 5.

4.3 Dynamic Replication Scheme Based on Association Mining

The proposed replication scheme uses the flow balance assumption [15] as indicator of replication. The flow balance assumption states that the completion and arrival rate must have almost same value in able to provide fast response of object services to requests. If the arrival rate is less than the service rate, then replication is occurred. Moreover, the dynamic replication scheme uses association rule mining to construct a pattern from transactions of object request to predict the next objects needed for replication. The Apriori algorithm was used to perform association rule extraction. In this research we used the object replication patterns as the patterns produced by overloading of objects. There are two steps of Apriori algorithm which use to generate the candidate pattern and choose the frequent pattern. First is the join step which finds L_k, a set of candidate k-itemsets by joining L_{k-1} with itself. Second is the prune step where C_k is generated as superset of L_k, and all of the frequent k-itemsets are included in C_k. Choosing the frequent pattern is based on confidence value. The support count determines each pattern frequency shown in Equation 6. After determining the support count, the confidence is determined by getting the ratio of support count and total number of data in Equation 7. After generating all patterns with its respective confidence, the patterns are used in comparing the current object replication patterns.

$$s\_count(A \to B) = P(A \cup B) \tag{6}$$

$$confidence(A \to B) = P(B / A) = \frac{s\_count(A \cup B)}{s\_count(A)} \tag{7}$$

$$prediction(R_k) = If\ O_1\ is\ l\ and,...,and\ O_k\ is\ l\ then\ O_j \tag{8}$$

In Equation 8, all rules are used in comparing the current object replication patterns where O is a type object and l indicates an overload and needs to replicate. If a rule is selected then it follows the next object type that will be overloaded and needs to replicate. Choosing which server performs the replication is considered by determining the server has a least count of objects currently operational. Also, the termination of object replica is done if arrival rate is greater than service rate.

4.4 Action Model for the Agent-Based Data Mining

The classical concepts of actions used in artificial intelligence are based on an approach involving states and transformation of states [9]. It is possible to characterize the set of global state (Σ) of possible states of the world. An action model is designed for the data mining model of the intelligent distributed framework. The operators used were described in Table 1. The agents in the model are the grouping agent, locator agent and replication agent which are represented by GA, LA and RA, respectively.

Table 1. The operators with description of functions used in the data mining model based on multi-agent action models

Operators	Functions
Exec	Executes the data mining algorithm by the local agent. Inputs represents data for the execution required in the algorithm
Informs	The agent communicates through this operator. The input of this operator is the address of the agent.
And	A joint operator with another operators to represent more commands to be done in the state

Table 1 is the reference of the actions from the proposed model. The Σ in the proposed data mining model is consisted of three sub-states which are the grouping (σ_1), searching (σ_2) and replication (σ_3) states. Equation 9 represents the grouping state where GA clusters the *X*, *X* is all data from objects, and communicates with LA. After informing LA, the structure of the MLP in the LA is adjusted. MLP is trained again with the new patterns from clustering. In Equation 10, searching state is described where a request of client represented by *R* is processed in MLP classifier of LA. The LA returns the appropriate object to the client also communicate to RA. In Equation 11 represents the replication state which compares the rule generated by Apriori to be used for predicting the object to be replicated and communicates to GA if successful.

$$\sigma_1 = GA\ Exec(clustering(X)\ And\ Informs(LA)) \tag{9}$$

$$\sigma_2 = LA\ Exec(classification(R)\ And\ Informs(RA)) \tag{10}$$

$$\sigma_3 = RA\ Exec(prediction(P)\ And\ Informs(GA)) \tag{11}$$

5 Experimental Evaluation

The proposed intelligent distributed framework used Visibroker 7.0 to implement the CORBA-based object grouping and JADE to implement the multi-agent interaction. The data mining algorithms which are *k*-means, MLP and Apriori were coded in Java and embedded to the agents. The operating system (OS) platforms used for the experiment were Windows OS, Red Hat Linux and Sun Solaris 8 to simulate the heterogeneity of system. The proposed method was evaluated by determining the delay time, number of objects replicated and load distribution and compared to a normal method.

5.1 Data Mining Results

In our simulation, a synthetic data was used to perform the proposed data mining model. A standard distribution of generating random data, which contains 5 attributes and 100 tuples were done, and then make these data as object properties. The grouping agent performs *k*-means to group the objects into 5 classes. After grouping, the MLP of the locator agent is trained by using the data clustered by grouping agent. Every time a client request occurs in the proposed framework, the locator agent processes the request contents into MLP. In simulating the replication scheme, we used the data transactions gathered from previous simulations and chose 1000 tuples for association mining using Apriori algorithm. Table 2 shows the lists of first 5 results from the association mining. The first rule from Table 2 means that if OBJ2 is least requested then it will imply OBJ3 is most requested with a 97 percent confidence. The same way of explanation can be done to other rules. After generating the rules, the replication agent uses the rules to associate with the patterns of client request.

Table 2. Association rules generated from object transactions

Association rules, showing only 5	Confidence
OBJ2=least req. →OBJ3=most req.	0.97
OBJ4=most req→OBJC3=most req.	0.98
OBJ1=medium req. →OBJC3=most req.	0.97
OBJ1=most req. →OBJ3=most req. OBJ2=least req.	0.96
OBJ5=most req. OBJ3=least req. →OBJ2=most req.	0.99

5.2 Performance of the Proposed Model

The locator agent implements the intelligent search scheme based on MLP with the support of replication agent that uses the flow balance assumption and association mining in performing the dynamic replication scheme for fast response of request. We simulated 1000 client requests and collected the result from the total time delay of clients shown in Figure 3a and the number of replication generated from client request is shown in Figure 3b. The events from request of clients trigger the replication of objects. The result was compared to normal method that has static replicas which means it do not dynamically replicate object. The default number of object replica is 5.

Figure 3a shows that the delay time is minimized as the number of client increases processed by the intelligent search. It shows an average of 38 times better than the normal method. The dynamic replication causes a minimal delay time by serving

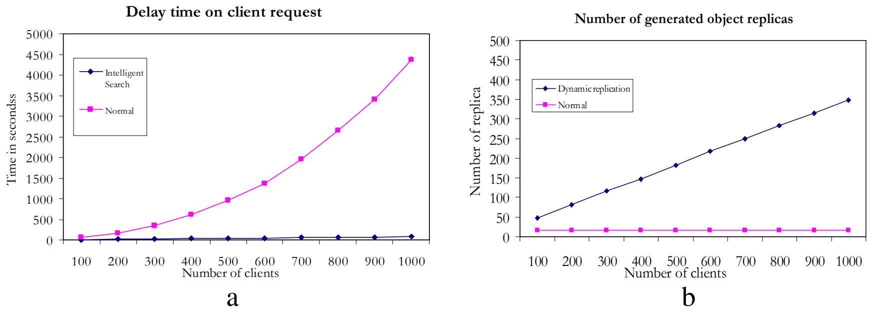


Fig. 3. Delay time (a) and number of object replication (b) based on number of clients

large numbers of client request. Also, in Figure 3b shows the number of object replication increases as the number of client increases. In managing object replications, these can be used on fault tolerance. The load distribution from replication is shown in Figure 4.

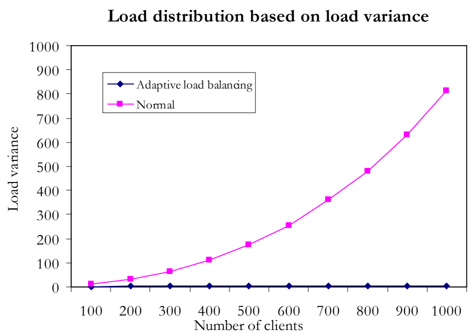


Fig. 4. Load distribution representation in load variance based on number of clients

The load variance in Figure 4 shows that using the proposed method is almost constant in distributing loads within the 5 server nodes with an average of 2.88 and is far better compared to the normal method with an average of 209.14 which exponential increases the variation of loads within the servers. It was observed that the curve from Figure 3a is expected to have the same inverse relationship to Figure 3b, but as the result shows, the replication is optimized by forming a straight line.

6 Conclusions and Future Work

The integration of intelligent system contributes to the efficiency of the system management and effectiveness of providing quality of service to clients. In this paper, the intelligent distributed framework was presented to address the use of intelligent model in the distributed object groups. The data mining model based on multi-agent system was proposed which uses the clustering, classification and association methods for object grouping, intelligent searching and dynamic replication schemes, respectively, in implementing intelligent adaptive schemes. The proposed data mining method is based on the action model of multi-agent system for efficient interaction of the components. Simulation result from proposed model showed 38 times better in minimizing delay time than the normal method and improves the performance by adaptive load distribution of large number of requests.

This paper implemented the data mining model using synthetic data and will use real application data in the future work. Also, other components for the intelligent distributed framework will be implemented.

References

1. Coulouris, G., Dollimore, J., Kinderberg, T.: Distributed Systems Concepts and Design, 4th edn., pp. 1–25. Addison-Wesley, Reading (2005)
2. Badidi, E., Keller, R.K., Kropf, P.G., Van Dongen, V.: The Design of a Trader-based CORBA Load Sharing Service. In: Proc. of the 12th International Conference on Parallel and Distributed Computing Systems, pp. 75–80 (1999)
3. Van Steen, M., Ballintijn, G.: Achieving Scalability in Hierarchical Location Services. In: Proc. of the 26th International Computer Software and Applications Conference (2002)
4. Felber, P., Guerraoui, R.: Programming with Object Groups in CORBA. Concurrency, IEEE 8(1), 48–58 (2000)
5. Felber, P., Guerraoui1, R., Schiper, A.: Replication of CORBA Objects. Advances in Distributed Systems: Advanced Distributed Computing: From Algorithms to Systems, 254 (2000)
6. Othman, O., O'Ryan, C., Schmidt, D.C.: The Design of an Adaptive CORBA Load Balancing Service. IEEE Distributed Systems Online 2 (4) (2001)
7. Kwok, Y.K., Cheung, L.S.: A New Fuzzy-decision based Load Balancing System for Distributed Object Computing. Journal of Parallel and Distributed Computing 2(64), 238–253 (2004)
8. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn., pp. 1–38. Morgan Kaufmann, San Francisco (2006)
9. Ferber, J.: Multi-Agent Systems An Introduction to Distributed Artificial Intelligence, pp. 307–340. Addison-Wesley, Reading (1999)
10. Hendler, J.: Introduction to the Special Issue: AI, agents, and the Web. Intelligent Systems, IEEE 21(1), 11–11 (2006)
11. Yang, L., Wang, F.Y.: Driving into Intelligent Spaces with Pervasive Communications. Intelligent Systems, IEEE 22(1), 12–15 (2007)
12. Marik, V., McFarlane, D.: Industrial Adoption of Agent-based Technologies. Intelligent Systems, IEEE 20(1), 27–35 (2005)
13. Pechoucek, M., Thompson, S.G., Baxter, J.W., Horn, G.S., Kok, K., Warmer, C., Kamphuis, R., Maric, V., Vrba, P., Hall, K.H., Maturana, F.P., Dorer, K., Calisti, M.: Agents in Industry: The Best from The AAMAS 2005 Industry Track. Intelligent Systems, IEEE 21(2), 86–95 (2006)
14. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press, New York (1973)
15. MacDougall, M.H.: Simulating Computer Systems Techniques and Tools, pp. 16–17. The MIT Press, Cambridge (1983)

Recalling the Embodied Meaning of Modal Conjunctions in Artificial Cognitive Agents

Wojciech Lorkiewicz and Radosław Piotr Katarzyniak

Institute of Information Science and Engineering, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-350 Wrocław, Poland
{wojciech.lorkiewicz, radoslaw.katarzyniak}@pwr.wroc.pl

Abstract. An approach to developing the interpretation of incoming modal conjunctions is proposed. It is assumed that an artificial cognitive receiving agent is given. This agent uses modal conjunctions both to inform about its states of knowledge as well as collects incoming conjunctions produced by other external agents. Each single incoming conjunction is transformed into a set of predefined constraints which are used by the receiving agent to modify its private and subjective opinion on a state of an object. The model for single modal conjunction integration is extended to the case of set of incoming conjunctions. The extension is based on consensus approach. Both models are supported by the theory of modal language grounding.

Keywords: Symbol grounding, symbol interpretation, agent.

1 Introduction

Natural cognitive agents use semantic languages to communicate their own knowledge to other agents and to collect the knowledge from other agents in their own knowledge bases. Natural semantic languages that are used by living agents consist of high level representation structures called language symbols. These symbols are knowledge carriers located in external world but they summarize the content of knowledge stored in strictly private and embodied knowledge bases. In case of living agents the nature of external and internal knowledge representations are very different from each other. For instance, external representations can belong to the set of natural language symbols and internal representations can be cognitive patterns based on neural networks. In this work the division into external and internal knowledge representations is applied to a case of artificial cognitive agents and a language of a class of modal formulas. Namely, the knowledge is assumed to be stored in an internal strictly private relational database and communicated by means of logic-like representations called modal conjunctions. The syntax and intuitive semantics for these latter formulas are given in Tab. 1 and Tab. 2.

In [2,6] an original theory of language grounding was introduced. Grounding is understood as relating symbols of communication language to external world and theories of grounding describe the ways this process can be achieved [1]. The theory presented in [2,6] defines all situations in which modal formulas, in particular modal

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 763–772, 2008.

conjunctions, become adequate descriptions of internal knowledge states. It is achieved by the introduction of the so-called epistemic satisfaction relation (or equivalently a grounding relation). Further potential designers of artificial agents are provided with strong a and clear postulates how to organize artificial language that fulfills important set of commonsense constraints originated from natural language processing. As such these postulates were given and deeply discussed in [3-5].

Table 1. Non-modal conjunctions and their commonsense interpretations

Conjunction	Intuitive interpretation
$p(o)\wedge q(o)$	Object *o* exhibits property *P* and exhibits property *Q*.
$p(o)\wedge\neg q(o)$	Object *o* exhibits property *P* and does not exhibit property *Q*.
$\neg p(o)\wedge q(o)$	Object *o* does not exhibit property *P* and exhibits property *Q*.
$\neg p(o)\wedge\neg q(o)$	Object *o* does not exhibit property *P* and does not exhibit property *Q*.

Table 2. Commonsense interpretations of modal extensions of conjunctions uttered by agent *a*, $\varphi\in\{p(o)\wedge q(o), p(o)\wedge\neg q(o), \neg p(o)\wedge q(o), \neg p(o)\wedge\neg q(o)\}$

Modal Extension	Commonsense interpretation
$Know(\varphi)$	I know that φ.
$Bel(\varphi)$	I believe that φ.
$Pos(\varphi)$	It is possible that φ.

The major assumption accepted in the theory of grounding [2,6] states that meaning of modal conjunctions (and other modal formulas) originates from internally stored experiences. These experiences are represented in the so-called base profiles and are taken into account when an artificial cognitive agent verifies applicability of a particular modal conjunction in a current state of its knowledge. However, in none of the aforementioned works [2-6] any process opposite to the process of language generation was discussed. Namely, in none of them any process of interpreting incoming modal conjunctions was studied and any integration of this interpretation with previously existing states of knowledge was proposed and discussed. It is quite obvious that such an opposite process of incoming messages' interpretation and integration is of equal importance for rational language behavior. In this paper an original strategy for interpretation and integration of received modal conjunctions into previously existing and internally stored knowledge bases is outlined. The incoming modal conjunctions, collected by the so-called receiving agent *a*, are interpreted by this agent and then are transformed into a set of internally activated conceptual (cognitive) structures called constraints. These constraints define the needed modification of each state of receiving agent's knowledge existing prior to the arrival of incoming modal conjunctions. It has been assumed that this process can be strongly supported by original contributions proposed in the theory of grounding and has been elaborated in detail in an unpublished thesis [8].

The organization of this paper is as follows. In section 2 the role of epistemic satisfaction relation for processing formulas from Tab. 1 and Tab. 2 is discussed and an original concept of grounding vector is introduced. In section 3 an original model for interpreting a single modal conjunction is presented and in section 4 an extension to this model is proposed by which integration of sets of incoming modal conjunctions is described.

2 The Source of Embodied Meaning and Production of Modal Conjunctions

Let us consider four conjunctions from Tab. 1 and their modal extensions from Tab. 2. It is assumed in the theory of grounding that cognitive agent's ability to produce modal conjunctions in order to inform a potential receiver about current state of an object *o* in regards of properties *P* and *Q* is strongly influenced by previously stored observations of *P* and *Q* in this object. Let the following four sets C^i, i=1..4 be considered to describe the relation between induced meaning and produced modal conjunctions:

Definition 1 (grounding sets)
Let C^1 consist of all internally stored cognitive representations of observations made by the receiving agent a in which object o exhibited property P and exhibited property Q.

Let C^2 consist of all internally stored cognitive representations of observations made by the receiving agent a in which object o exhibited property P and did not exhibit property Q.

Let C^3 consist of all internally stored cognitive representations of observations made by the receiving agent a in which object o did not exhibit property P and exhibited property Q.

Let C^4 consist of all internally stored cognitive representations of observations made by the receiving agent a in which object o did not exhibit property P and did not exhibit property Q.

In the theory of grounding sets C^i, i=1..4 are called grounding sets and their cardinalities are used to determine which modal operator should be used to extend a particular conjunction from Tab. 1, while it is produced to describe internal knowledge states. The strategy of applying grounding sets consists of the following steps:

STEP 1. For each grounding set C^i, i=1..4 the so-called relative grounding strength λ_i is determined according to the following formulae (1-4):

$$\lambda_1 = \frac{|C_1|}{|C_1|+|C_2|+|C_3|+|C_4|}\ , \tag{1}$$

$$\lambda_2 = \frac{|C_2|}{|C_1|+|C_2|+|C_3|+|C_4|} \tag{2}$$

$$\lambda_3 = \frac{|C_3|}{|C_1|+|C_2|+|C_3|+|C_4|} \tag{3}$$

$$\lambda_4 = \frac{|C_4|}{|C_1|+|C_2|+|C_3|+|C_4|} \quad (4)$$

Definition 2 (grounding vector)
The quadruple $<\lambda_1, \lambda_2, \lambda_3, \lambda_4>$ *is called the grounding vector.*

It is assumed in the theory of grounding that sets C^i, i=1..4 are the sources of meaning assigned to conjunctions $p(o) \wedge q(o)$, $p(o) \wedge \neg q(o)$, $\neg p(o) \wedge q(o)$, and $\neg p(o) \wedge \neg q(o)$, respectively. In other words, values λ_i, i=1..4 describe indirectly the subjectively experienced "confidence" with which the meaning communicated by conjunctions $p(o) \wedge q(o)$, $p(o) \wedge \neg q(o)$, $\neg p(o) \wedge q(o)$, $\neg p(o) \wedge \neg q(o)$ is experienced both by the sending as well as receiving agents.

STEP 2. A system of modality thresholds, predefined by the theory of grounding, is activated in and used by the sending agent to evaluate the actual role of λ_i, i=1..4 in grounding related conjunctions. Such system is an integral part of the theory and can be described by a triple $<\lambda^\wedge_{minP}, \lambda^\wedge_{P/B}, \lambda^\wedge_{maxB}>$ consisting of three rational (? numbers. The role of these thresholds is very important and can be explained as follows. At first, if a relative grounding value λ_i i=1..4 fulfills the requirement $\lambda^\wedge_{minP} \leq \lambda_i < \lambda^\wedge_{P/B}$, then the conjunction described by this value can be extended with possibility operator *Pos*. At second, if a relative grounding value λ_i i=1..4 fulfills the requirement $\lambda^\wedge_{P/B} \leq \lambda_i < \lambda^\wedge_{maxB}$, then the conjunction described by this value can be extended with belief operator *Bel*. At third, if a relative grounding value λ_i i=1..4 fulfills the requirement $\lambda_i <= 1$, then the conjunction described by this value can be extended with knowledge operator *Know*. Details of the modality thresholds relations are given in [3-6].

STEP 3. The so-called epistemic satisfaction relation is applied by the sending agent to find out which modal conjunctions should be treated as possible descriptions of its current knowledge state. This verification is supported by the following definition:

Definition 3 (Epistemic satisfaction of modal conjunctions)
Let *KS* denote a current state of agent's knowledge, let *o* be an object, and let *P* and *Q* be two different properties related to language symbols *p* and *q*, respectively. Let $KS \models_G \varphi$ denote epistemic satisfaction relation of formula φ by the agent's knowledge state *KS*. For modal extensions of $p(o) \wedge q(o)$ the epistemic satisfaction relation is given as follows:

- $KS \models_G Pos(p(o) \wedge q(o))$ holds if and only if the current state of properties *P* and *Q* in object *o* is not known to the agent, however at least one previous observation in which object *o* exhibited properties *P* and *Q* has been retrieved by this agent from its embodied knowledge base *KS* and the condition $\lambda_{minP} \leq \lambda_1 \leq \lambda_{P/B}$ holds.
- $KS \models_G Bel(p(o) \wedge q(o))$ holds if and only if the current state of properties *P* and *Q* in object *o* is not known to the agent, however at least one previous observation in which object *o* exhibited properties *P* and *Q* has been retrieved by this agent from its embodied knowledge base *KS* and the condition $\lambda_{P/B} < \lambda_1 \leq \lambda_{maxB}$ holds.
- $KS \models_G Know(p(o) \wedge q(o))$ holds if and only if (a) the current state of properties *P* and *Q* in object *o* is not known to the agent, however at least one previous

observation in which object *o* exhibited properties *P* and *Q* has been retrieved by this agent from its embodied knowledge base *KS* and the condition $\lambda_1=1$ holds or (b) this agent has just observed that object *o* exhibits both properties *P* and *Q,* simultaneously.

Definitions for epistemic satisfaction of modal extensions of $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$ and $\neg p(o)\wedge\neg q(o)$ are similar, provided that necessary modifications are taken into account for applied grounding sets and values λ_i.

All modal conjunctions, satisfied in the above sense, are chosen by the sending agent as possible language representations of its current knowledge state.

Detailed formalizations of the aforementioned definitions is given in [3-6].

Table 3. Constraints definitions for knowledge extensions

Modal Formula	Grounding vector restrictions	Grounding set restrictions
$Know(p\wedge q)$	$\lambda_1=1,\ \lambda_2=\lambda_3=\lambda_4=0$	$\lvert C_2\rvert+\lvert C_3\rvert+\lvert C_4\rvert=0$
$Know(p\wedge\neg q)$	$\lambda_2=1,\ \lambda_1=\lambda_3=\lambda_4=0$	$\lvert C_1\rvert+\lvert C_3\rvert+\lvert C_4\rvert=0$
$Know(\neg p\wedge q)$	$\lambda_3=1,\ \lambda_1=\lambda_2=\lambda_4=0$	$\lvert C_2\rvert+\lvert C_1\rvert+\lvert C_4\rvert=0$
$Know(\neg p\wedge\neg q)$	$\lambda_4=1,\ \lambda_1=\lambda_2=\lambda_3=0$	$\lvert C_2\rvert+\lvert C_3\rvert+\lvert C_1\rvert=0$

Table 4. Constraints definitions for belief extensions

Modal Formula	Grounding vector restrictions / Grounding set restrictions
$Bel(p\wedge q)$	$\lambda_{P/B}<\lambda_1<1,\ 0<\lambda_2<\lambda_{P/B},\ 0<\lambda_3<\lambda_{P/B},\ 0<\lambda_4<\lambda_{P/B}$
	$\frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_2\rvert+\lvert C_3\rvert+\lvert C_4\rvert)<\lvert C_1\rvert$
$Bel(p\wedge\neg q)$	$\lambda_{P/B}<\lambda_2<1,\ 0<\lambda_1<\lambda_{P/B},\ 0<\lambda_3<\lambda_{P/B},\ 0<\lambda_4<\lambda_{P/B}$
	$\frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_1\rvert+\lvert C_3\rvert+\lvert C_4\rvert)<\lvert C_1\rvert$
$Bel(\neg p\wedge q)$	$\lambda_{P/B}<\lambda_3<1,\ 0<\lambda_2<\lambda_{P/B},\ 0<\lambda_1<\lambda_{P/B},\ 0<\lambda_4<\lambda_{P/B}$
	$\frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_2\rvert+\lvert C_1\rvert+\lvert C_4\rvert)<\lvert C_1\rvert$
$Bel(\neg p\wedge\neg q)$	$\lambda_{P/B}<\lambda_4<1,\ 0<\lambda_2<\lambda_{P/B},\ 0<\lambda_3<\lambda_{P/B},\ 0<\lambda_1<\lambda_{P/B}$
	$\frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_2\rvert+\lvert C_3\rvert+\lvert C_1\rvert)<\lvert C_1\rvert$

Table 5. Constraints definitions for possibility extensions

Modal Formula	Grounding vector restrictions / Grounding set restrictions
$Pos(p \wedge q)$	$\lambda_{\min P} < \lambda_1 < \lambda_{P/B}$, $0 < \lambda_2 < 1-\lambda_{\min P}$, $0 < \lambda_3 < 1-\lambda_{\min P}$, $0 < \lambda_4 < 1-\lambda_{\min P}$
	$\frac{\lambda_{\min P}}{1-\lambda_{\min P}}(\lvert C_2 \rvert + \lvert C_3 \rvert + \lvert C_4 \rvert) < \lvert C_1 \rvert < \frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_2 \rvert + \lvert C_3 \rvert + \lvert C_4 \rvert)$
$Pos(p \wedge \neg q)$	$\lambda_{\min P} < \lambda_2 < \lambda_{P/B}$, $0 < \lambda_1 < 1-\lambda_{\min P}$, $0 < \lambda_3 < 1-\lambda_{\min P}$, $0 < \lambda_4 < 1-\lambda_{\min P}$
	$\frac{\lambda_{\min P}}{1-\lambda_{\min P}}(\lvert C_1 \rvert + \lvert C_3 \rvert + \lvert C_4 \rvert) < \lvert C_1 \rvert < \frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_1 \rvert + \lvert C_3 \rvert + \lvert C_4 \rvert)$
$Pos(\neg p \wedge q)$	$\lambda_{\min P} < \lambda_3 < \lambda_{P/B}$, $0 < \lambda_1 < 1-\lambda_{\min P}$, $0 < \lambda_2 < 1-\lambda_{\min P}$, $0 < \lambda_4 < 1-\lambda_{\min P}$
	$\frac{\lambda_{\min P}}{1-\lambda_{\min P}}(\lvert C_2 \rvert + \lvert C_1 \rvert + \lvert C_4 \rvert) < \lvert C_1 \rvert < \frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_2 \rvert + \lvert C_1 \rvert + \lvert C_4 \rvert)$
$Pos(\neg p \wedge \neg q)$	$\lambda_{\min P} < \lambda_4 < \lambda_{P/B}$, $0 < \lambda_1 < 1-\lambda_{\min P}$, $0 < \lambda_3 < 1-\lambda_{\min P}$, $0 < \lambda_2 < 1-\lambda_{\min P}$
	$\frac{\lambda_{\min P}}{1-\lambda_{\min P}}(\lvert C_2 \rvert + \lvert C_3 \rvert + \lvert C_1 \rvert) < \lvert C_1 \rvert < \frac{\lambda_{P/B}}{1-\lambda_{P/B}}(\lvert C_2 \rvert + \lvert C_3 \rvert + \lvert C_1 \rvert)$

3 Interpretation of Single Incoming Modal Conjunction

It has already been explained above in what way a single conjunction should be related to grounding restrictions (epistemic satisfaction) in an (artificial) mind of the sender. The opposite process of interpretation is realized when this modal conjunction is received by the receiving agent and needs to be integrated into its knowledge state. In the unpublished thesis [8] the following scenario for interpretation of single modal conjunction and their integration was proposed.

STEP 1. The receiving agent performs an analysis of encapsulated knowledge base and determines related grounding sets C^i, i=1..4, as well as related coordinates of grounding vector $\lambda = <\lambda_1, \lambda_2, \lambda_3, \lambda_4>$ (see Def. 2). This way the agent computes its internal and purely subjective standpoint to be used in further interpretation of received conjunction.

STEP 2. The receiving agent activates a set of cognitive constraints induced by received modal conjunction in its knowledge base. It chooses this set of constraints from Tab. 3-5, where they are defined both over grounding vector coordinates as well as over grounding cardinalities values what is equivalent.

All constraints given in Tab. 3-5 result directly from the theory of grounding [2,6].The influence of activated constraints on modal conjunctions' interpretation and integration is as follows: The activated set of constraints defines the way the grounding set assigned to this modal conjunction that was received, should be related to the other three grounding sets to make the generation of such conjunction possible, where this possibility is defined by rules of grounding theory [2,6]. For instance, if a received modal conjunction is based on $p(\mathrm{o})\wedge q(\mathrm{o})$, then the activated set of constrains defines all necessary relations required for cardinality of set C^1 and the remaining three cardinalities of sets C^2, C^3 and C^4. However, if the actual cardinalities of these grounding sets, encapsulated in the receiving agent, do not fulfill the activated set of restrictions, then all necessary minimal modifications to cardinalities of sets C^1, C^2, C^3 and C^4 need to be determined to make the received modal conjunction properly grounded (according to Def. 3). In other words, by determining these minimal modifications to the current state the receiving agents wants to get to know what minimal changes to its current knowledge state need to be introduced to make the received modal conjunction satisfied in the epistemic sense (or properly grounded what is equivalent).

STEP 3. By confronting the current values of actual grounding vector coordinates and actual grounding cardinalities with activated set of constraints, the receiving agent determines a required modification vector $m^*=<m^1,m^2,m^3,m^4,>$, where each m^i, i=1..4 is a minimal modification of cardinality $|C^i|$ required to make the activated set of constraints fulfilled. Rules for determining coordinate values of m^* for the case of all modal extensions of $p(o)\wedge q(o)$ are given as follows:

CASE $Know(p(o)\wedge q(o))$: Required modification is applied to cardinalities $|C_2|, |C_3|, |C_4|$ and is described by vector: $m^* =< 0, -|C_2|, -|C_3|, -|C_4|>$.

CASE $Bel(p(o)\wedge q(o))$: required modification is applied to cardinality $|C_1|$ according to the following rules:

IF $\frac{\lambda_{P/B}}{1-\lambda_{P/B}}(|C_2|+|C_3|+|C_4|) > |C_1|$, THEN

$$m^* =< \frac{\lambda_{P/B}}{1-\lambda_{P/B}}(|C_2|+|C_3|+|C_4|) - |C_1|, 0,0,0 >$$

OTHERWISE $m^* =< 0,0,0,0 >$.

CASE $Pos(p(o)\wedge q(o))$: required modification applies to cardinality $|C_1|$ according to the following rules:

IF $\frac{\lambda_{\min P}}{1-\lambda_{\min P}}(|C_2|+|C_3|+|C_4|) > |C_1|$, THEN

$$m^* = < \frac{\lambda_{\min P}}{1-\lambda_{\min P}}(|C_2|+|C_3|+|C_4|) - |C_1|, 0,0,0 > ;$$

IF $\frac{\lambda_{P/B}}{1-\lambda_{P/B}}(|C_2|+|C_3|+|C_4|) < |C_1|$, THEN

$$m^* = < \frac{\lambda_{P/B}}{1-\lambda_{P/B}}(|C_2|+|C_3|+|C_4|) - |C_1|, 0,0,0 >$$

OTHERWISE $m^* = < 0,0,0,0 >$.

Similar modifications can be proposed for the remaining cases of modal conjunctions $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$ *and* $\neg p(o)\wedge\neg q(o)$. Necessary changes are obvious.

STEP 4. The receiving agent applies computed modification m^* to change its state of knowledge. It is assumed in [8] that the intensity with which vector m^* influences a current state of receiving agent's knowledge depends on an additional factor called the willingness towards induced modifications. This concept is represented by a value $\gamma: \gamma\in[0,1]$, where $\gamma=0$ denotes the tendency of the receiving agent to reject all suggested modifications and $\gamma=1$ denotes the opposite highest level of willingness to accept them. For a given modification vector $m^*=<m_1,m_2,m_3,m_4>$, a vector of current grounding set cardinalities $g=<|C_1|,|C_2|,|C_3|,|C_4|>$ and an assumed level γ, a modification function $g^*=F(m^*,g,\gamma)$ is then applied, where F fulfills additional requirements $F(m^*,g,0)=g^*=g$ and $F(0,g,\gamma)=g^*=g$.

It is argued in [6] that there exists a class of functions F fulfill the above mentioned requirements. However the simplest one seems to be the following function:

$$g^* = g + \gamma m = < g_1 + \gamma m_1, g_2 + \gamma m_2, g_3 + \gamma m_3, g_4 + \gamma m_4 >, \quad (5)$$

which implies that in case of an agent with the highest willingness to apply the induced modifications, each state of knowledge described by vector $<|C_1|,|C_2|,|C_3|,|C_4|>$ is substituted by vector $<m_1+|C_1|, m_2+|C_2|, m_3+|C_3|, m_4+|C_4|>$.

The latter completes the whole process of interpretation of a single incoming modal conjunction and its integration with previously developed state of knowledge. Obviously, this interpretation and integration of new and external incoming opinion may strongly influence the set of modal conjunctions that are satisfied in the sense of Def. 3.

4 Integration of Multiple Incoming Modal Conjunctions

There are natural extensions possible to the above strategy proposed for processing of a single incoming modal conjunction. The most obvious captures the situation in which the receiving agent is handled with sets of multiple modal conjunctions coming from external senders. In [8] a consensus based strategy is proposed for interpreting and integrating multiple incoming modal conjunctions. This strategy assumes that each incoming modal conjunction is processed separately to yield related individually induced vectors of modifications. However a consensus modification is derived from the set of all modification vectors. This consensus modification is then applied to realize induced changes in knowledge state. The strategy is presented as follows.

STEP 1. For each received modal conjunction the receiving agent carries out STEP 1, STEP 2 and STEP 3 (defined in section 3) in order to determine the resulting set *M* of all individual modifications induced by single received modal conjunctions.

STEP 2. The receiving agent derives a consensus modification from the set *M*. For a given set of all interpretations *M*, the possible consensus modification *Con*(*M*) can be defined as a vector $m^*=<m_1,m_2,m_3,m_4> \in \Re^4$ that fulfills the postulate of minimal summarized distance. This solution seems to be the simplest and treats all individual modifications in an equal manner. Let $M=\{m^1,m^2,...,m^L\}$, where $m^i=<m^i_1,m^i_2,m^i_3,m^i_4>$, i=1..L be the set of all determined modifications. The suggested consensus structure can be given as:

$$Con(M^*) = m^* =< m_1^*, m_2^*, m_3^*, m_4^* >,$$

where $m_1^* = \frac{1}{L}\sum_{i=1}^{L} m_1^i$, $m_2^* = \frac{1}{L}\sum_{i=1}^{L} m_2^i$, $m_3^* = \frac{1}{L}\sum_{i=1}^{L} m_3^i$ and $m_4^* = \frac{1}{L}\sum_{i=1}^{L} m_4^i$.

This structure exhibits desirable properties, where $\|.\|$ denotes an assumed vector norm:

Property 1

If $Con(M^*) = m^*$, then for $\| m' \| > \| m^* \|$ the inequality

$\| Con(M^* \cup m') \| \geq \| Con(M^*) \| = \| m^* \|$ holds.

Property 2

If $Con(M^*) = m^*$, then for $\| m' \| < \| m^* \|$ the inequality

$\| Con(M^* \cup m') \| \leq \| Con(M^*) \| = \| m^* \|$ holds.

Property 3

If $Con(M^*) = m^*$, then for $m' = m^*$ the equality

$Con(M^* \cup m') = Con(M^*) = m^*$ holds.

Property 4

$Con(\{m\}) = m$ and $Con(\emptyset) =< 0,0,0,0 >$

Other models for deriving consensus modification are also possible [8].

5 Final Remarks

In this work an original approach to developing the interpretation of incoming modal conjunctions has been presented. This process has been described for the case of modal conjunctions. The proposed model shows how opinions of other agents can be used to influence private opinions of individual agents, provided that these opinions have been communicated by means of modal extensions of conjunctions $p(o)\wedge q(o)$, $p(o)\wedge\neg q(o)$, $\neg p(o)\wedge q(o)$, $\neg p(o)\wedge\neg q(o)$.

The proposed approach has been based on the theory of grounding a modal language of communication consisting of simple modalities, modal conjunctions and modal alternatives (inclusive and exclusive) [6]. This theory has been used to derive sets of cognitive constraints that need to be activated by an artificial cognitive agent when it receives a concrete modal formula and tries to integrate it with its existing knowledge state. The problem of interpretation and integration of incoming modal formulas has already been discussed in more detail in an unpublished work [8]. However, some results related to interpreting and integrating simple modalities have already been presented in [9]. Forthcoming works should cover the case of modal alternatives, both inclusive and exclusive ones.

An alternative approach to integration of incoming modal formulas has been proposed in [7], too. However, in this contribution a stronger theoretical model for grounding has been applied and a reference to a particular class of artificial cognitive agents has been made.

References

1. Harnad, S.: The symbol grounding problem. Physica D 42, 335–346
2. Katarzyniak, R.P.: The language grounding problem and its relation to the internal structure of cognitive agents. Journal of Universal Computer Science 11(2), 357–374 (2005)
3. Katarzyniak, R.P.: On some properties of grounding non-uniform sets of modal conjunctions. Int. Journal of Applied Mathematics and Computer Science 16(3), 399–412 (2006)
4. Katarzyniak, R.P.: On some properties of grounding uniform sets of modal conjunctions. Journal of Intelligent & Fuzzy Systems 17(3), 209–218 (2006)
5. Katarzyniak, R.P.: Some notes on grounding singletons of modal conjunctions. Systems Science 32(1), 45–55 (2005)
6. Katarzyniak, R.P.: Grounding modal languages of communication in artificial agents, EXIT Warsaw (in Polish) (2007)
7. Katarzyniak, R.P., Pieczyńska-Kuchtiak, A.: An approach to resolving semantic inconsistency of multiple prepositional attitudes. Journal of Intelligent & Fuzzy Systems 17(3), 231–238 (2006)
8. Lorkiewicz, W.: Methods for knowledge integration in believable agents. MSc Thesis, Wrocław University of Technology, Faculty of Computer Science and Management, Wrocław (in Polish) (2007)
9. Lorkiewicz, W.: An approach to the integration of simple modalities into the knowledge bases encapsulated in artificial cognitive agents. In: Proc. Int. Workshop on Knowledge Processing and Reasoning for Information Society, February 18–19, Wrocław Poland (in press, 2008)

Performance Evaluation of Multiagent Systems: Communication Criterion

Faten Ben Hmida, Wided Lejouad Chaari, and Moncef Tagina

University of Manouba, National School of Computer Studies, SOIE / LI3 Laboratory, 2010 Manouba, TUNISIA
{Faten.BenHmida, Wided.Chaari, Moncef.Tagina}@ensi.rnu.tn

Abstract. Many MultiAgent Systems (MAS) have been developed in various application domains such as computer networks, Internet, industrial applications, automation, process control, air traffic, robotic, simulation, etc. In spite of the rapid growth of the international interest in MAS field and the importance of the number of developed MAS, there is still a lack related to their performance evaluation. In fact, there is no measurement tool that allows evaluating the performance of a MAS or comparing two MAS. The existing works on systems performance evaluation deal principally with classic computer systems. In this paper, we try to identify the MAS' special features and properties which have an impact on their performance and we propose a measurement model to evaluate one of these properties: communication. This model is based on the graph theory. An experimentation of the proposed evaluation model is carried out and tested on a diagnosis application.

Keywords: Multiagent Systems, Communication, Performance Evaluation, Measurement, Methodology, Graph Theory.

1 Introduction

With the continuous extension of MAS application fields, there are more and more issues concerning particularly standardization of many concepts governing the development of this new technology. General issues deal with MAS terminology and incite scientists to agree on a consensus defining basic notions relative to MAS. Other issues are concerned about more precise questions such as the enumeration of the MAS's common characteristics and their estimation within the framework of evaluation prospects [1][2].

Now let's focus on the latter point. Consider the situation where we are faced to a given problem and we have several computer systems destined to resolve this problem. In this case, how could we know the most appropriate one? It is essential that we have a mean allowing us to choose the most appropriate system while we are convinced of the reasons of our choice. Yet, we should refer to a standard methodology to evaluate and analyze the performance of each system and thus make our choice.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 773–782, 2008.

However, the problem with the MAS is that there is no mean to evaluate the efficiency of a MAS or to compare two MAS. In fact, performance evaluation was rarely considered within the framework of MAS [3]. Consequently, until now, we don't have a standard method to evaluate and quantify the common MAS's characteristics.

Our work precisely deals with this point. It is question of establishing a methodology to analyze and measure the characteristics of MAS from an organizational point of view. In this paper, we don't look on all these characteristics at the same time, but we start by establishing a mean to evaluate one characteristic which is *communication*.

This paper is organized as follow: In section 2, we briefly present some basic notions relative to system performance evaluation. In section 3 we explain our approach to evaluate MAS communication. Our evaluation system's architecture is described in detail in section 4. Section 5 shows the results obtained after the experimentation of our approach on a multiagent application. Results are followed by some interpretations and explanations. A conclusion and a look at future work are presented in section 6.

2 System Performance Evaluation

Performance evaluation is the most critical part in any system construction. It consists in determining the different aspects of a system performance and estimating them quantitatively and/or qualitatively. There is no general definition of a performance metric, it is system dependent and its definition requires understanding the system and its usage well [4]. According to Jain, it's an art of computer systems engineering [5]. There are three basic techniques through which performance evaluation can be performed [6]:

- **Analytical modeling:** consists in using abstract model based on mathematical notions to describe certain aspects of the real system. This model is analyzed numerically to evaluate the real system's performance.
- **Simulation:** consists in implementing a simplified model that reproduces the real system behavior in software.
- **Measurement:** consists in fitting the system with specific instruments that allow picking up the relevant values in order to measure the system's performance.

Most of the evaluation methodologies present in the literature take inspiration from Jain's one [5]. This methodology is composed of the following steps:

1. Define the evaluation objective and determine the system components which will be considered during the evaluation process.
2. List all the services provided by the system and the possible results following the solicitation of each service.
3. List the system criteria useful to the performance evaluation.
4. Select the system parameters having an impact on its performance.
5. Select the factors to make vary in order to observe the consequences of their variation on the system performance.

6. Choose the appropriate evaluation technique.
7. Fix the system workload (the number and the nature of requests submitted to it).
8. Carry out the experimentation.
9. Analyze and interpret the obtained results.
10. Display the conclusions to the user.

3 Evaluation Methodology

To evaluate the MAS's performance, two modifications were made on Jain's methodology in order to adapt it to this context: The first modification consists in reorganizing the different steps. In fact, we noticed that some steps are independent from the application domain. We gathered these steps together in a first phase that we call *choice phase*. The second phase that we call *implementation phase* includes the steps to realize in the presence of the application to evaluate. The second modification concerns steps 3, 4 and 5 of Jain's methodology. Here we confuse parameters, criteria and factors and we propose to identify the common *characteristics* of the MAS. Thus, the evaluation methodology we use is described below:

- First phase : Choice phase (application independent)
  - Define the objective and determine the system components.
  - List the system's characteristics.
  - Select the characteristic to evaluate.
  - Choose the evaluation technique to apply.
- Second phase : implementation phase (application dependent)
  - List all the system's services and the possible results.
  - Fix the system's workload.
  - Carry out the experimentation.
  - Analyze and interpret the obtained results.
  - Display conclusions to the user.

First of all, the evaluation objective must be the clearest possible, so that, the evaluation process deals only with the pertinent data necessary to attend this objective. We have identified many abstraction levels according to which performance data can be classified:

- The lowest level concerns data relative to the effects that the MAS has on the host computer system such as CPU utilization or memory consumption.
- The highest level concerns generic data relative to the MAS's proper characteristics which are generally used by designers to evaluate the efficiency of a design approach or to compare different design approaches.

Our objective is to evaluate the performance of MAS from an organizational point of view. This supposes to deal with the highest abstract level data. So, we aren't interested in material constraints and hardware implementations of the MAS, we are rather interested in the common characteristics of MAS. According to a study made by Boissier and al. in [2], MAS have 13 characteristics which are: Autonomy, Distribution, Decentralization, Communication, Interaction, Organization, Situation in an environment, Openness, Emergence, Adaptation, Delegation, Personalization, and

Intelligibility. The second point we should tackle is the fact of defining the system components which will be considered during the evaluation. In this work, the whole MAS is considered. However, we don't take care about the internal architectures of the several agents.

Communication is one of the most important characteristics of MAS. In fact, it is a central aspect, on the basis of the agents' interaction, and essential to realize the social attribute of the MAS [7]. Thus, in this work, we focus on the evaluation of the different aspects of communication in MAS. For this purpose, two evaluation techniques are used: analytical modeling and direct measurements on the system.

4 Communication Evaluation System

The evaluation system proposed is composed of two main modules: an observation module and a measurement module (see Fig.1 below).

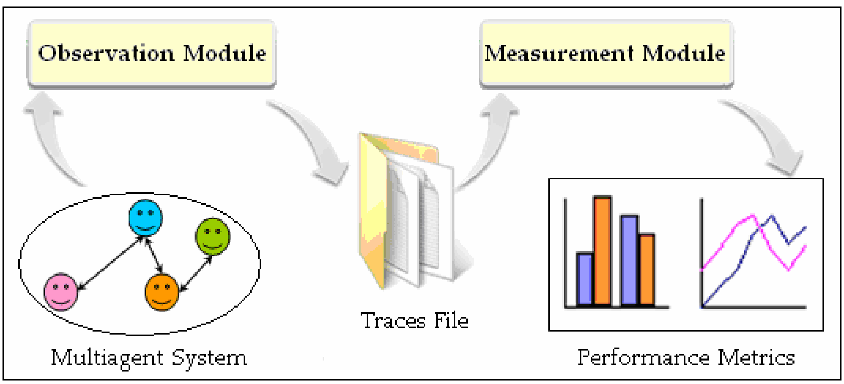


Fig. 1. Global architecture of the evaluation system

4.1 Observation Module

There are three different ways in which a MAS could be observed:

- The first consists in observing the system implicitly, i.e. spying it without interfering with it, so that there is no impact on its performance.
- The second consists in observing the system explicitly. This can be performed thanks to check points, captors or counters, etc.
- The third consists in submitting requests to the system to test its performance.

The two latter methods suppose to interfere with the system under study and consequently, the obtained results could be distorted. That's why we propose to observe our system activity using the first method i.e. implicitly, so that we don't need to analyze the effect of the observation on the system performance.

4.2 Measurement Module

Communication could be evaluated according to three main aspects which are:

- The structural aspect concerning the network communication topology.
- The syntactical aspect concerning the messages typology and complexity.
- The statistical aspect concerning the quantification of usual data such as the number of exchanged messages and their sizes.

Our study covers these three aspects. However, greater importance is accorded to the communication structural aspect. This is due to the nature of the MAS model used. In fact, our system is modeled by an oriented graph where the nodes represent the system agents and the arcs represent the communication links between those agents. Each arc is weighted by the number of messages exchanged. According to [8], an oriented graph $G=[X,U]$ is composed of:

- A set X of nodes, where $|X|=N$.
- A set U of oriented node pairs called arcs, where $|U|=M$.

There are several possible representations of a graph. Here, the adjacency matrix is used.

$$A=\left(A_{ij}\right)_{\substack{i=1..N\\ j=1..N}} \tag{1}$$

In an adjacency matrix, each line (column) corresponds to a node in the graph.

$$A_{ij}=1 \;\; \textit{if and only if} \;\; \left(i,j\right)\in U \;\; \left(A_{ij}=0 \; \textit{otherwise}\right). \tag{2}$$

In order to refine this model, each arc is weighted by the number of messages exchanged by the corresponding nodes. For this purpose, another matrix called *weight matrix* is used.

$$P_{ij}=p(u) \;\; \textit{if and only if} \;\; u=\left(i,j\right)\in U \;\; \left(P_{ij}=0 \; \textit{otherwise}\right). \tag{3}$$

Structural properties of the communication graph

Once the MAS's communication graph is generated, we have to analyze its properties. Our analysis is based on graph theory which is a branch of mathematics concerned about networks encoding and properties [9]. To apprehend the graph communication structure, we start by measuring the two half-degrees of each node.

- External half-degree $d^{+}\left(n\right)$: it is the number of links coming from n . It reflects the participation degree of the corresponding agent to the communication act.
- Internal half-degree $d^{-}\left(n\right)$: it is the number of links coming into n . It reflects the solicitation degree of the corresponding agent to the communication act.

$$d^{+}(n)=\sum_{j=1}^{N} A_{nj} \quad and \quad d^{-}(n)=\sum_{i=1}^{N} A_{in} \tag{4}$$

There are other criteria useful to describe the graph's structure, namely, *indices*. Indices where developed by K. J. Kansky in 1963 in order to evaluate transport networks [9]. Some of Kansky's indices are used, which are:

- Beta index β: it is expressed by the relationship between the number of links (M) and the number of nodes (N). The higher is β the more complex is the network. So, Beta index reflects the complexity of the communication network connecting the agents.

$$\beta=\frac{M}{N} \tag{5}$$

- Gamma index γ: it is expressed by the relationship between the number of observed links and the number of possible links. Its value is between 0 and 1. A value of 1 indicates a completely connected network. In our case, Gamma index allows us to measure the communication degree in the MAS.

$$\gamma=\frac{E}{N^{2}} \text{ where } E=\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij} \tag{6}$$

- Theta index θ: it measures the average amount of traffic per node. The higher is θ the greater is the load of the network.

$$\theta=\frac{Q(G)}{N} \text{ where } Q(G)=\sum_{i=1}^{N}\sum_{j=1}^{N} P_{ij} \tag{7}$$

Also, the load $Q(n)$ of each node can be measured.

$$Q(n)=\sum_{\substack{i=1\\ i\neq n}}^{N} P_{in}+P_{ni} \tag{8}$$

- Connectivity: A graph is said to be connected if for all its distinct pairs of nodes there is a linking chain. The study of the communication graph's connectivity enables us to have an idea about the organization of the MAS's agents. Identifying the different connected sub-graphs comes to identifying the different acquaintances in the MAS. To this end, Tarjan's algorithm [8], which allows finding the connected components in a graph, is used.
- Articulation points: In a connected graph, a node is said to be an articulation point if its suppression increases the number of connected components in the graph. The existence of such nodes in the MAS's communication network reflects some centralization in the communication. Tarjan's algorithm [8] is used to find the articulation points in the communication graph.

Syntactical properties of the communication

There are two criteria according to which the syntactical aspect of the communication is evaluated. These criteria are:

- Messages typology: The communication between agents is based on speech act theory which treats communication as action and claims that speech acts could change the state of the world just like physical actions [10]. All agent communication languages define a list of performatives corresponding to various communication acts. To study the messages typology, the performative field is extracted from each message and the number of the performatives used is counted.
- Messages complexity: Depending on the message structure, to each captured message one of these three qualifications (*simple*, *medium*, *complex*) is attributed. The qualification attributed depends on whether the content of the message is:
  - A string: in this case, the content is *simple* since it does not need to be encoded and decoded to be interpreted.
  - An ontology based clause: in this case, the content is *medium* since it is encoded according to a common ontology and needs to be decoded by the agent to be interpreted.
  - A protocol based message: in this case, the content is *complex* since a protocol specifies a number of rules and behaviors to be performed in addition to the common ontology.

Statistics

The analysis of the communication is completed by some statistics. In fact, we count the total number of exchanged messages, the messages sizes, the number of agents involved in the communication, and the percent of agents involved in the communication.

5 Experimentation

Our evaluation system has been tested on a multiagent application designed in order to detect and localize failures in an industrial system [11][12]. This application was implemented on the JADE multiagent platform. It consists of the following agents:

- The detection agents D1, D2, D3, D4 and D5, whose role is to detect the failures.
- The localization agent LOC, whose role is to localize the failure.
- The interface agent INT, whose role is to coordinate the other agents processing's and to display results to the application user.

As we mentioned before, our evaluation system is composed of two modules: the observation module and the measurement module: The observation module uses a Jade spy agent called *Sniffer* [13] to collect the several messages exchanged by the application's agents. All the messages captured by the Sniffer are saved in a traces file. The measurement module takes this file as input. An ACL parser parses it and progressively fills a table with the information contained in this file. Then this table is used to draw the communication graph and to calculate the performance metrics. Fig.2 shows the communication graph of the diagnosis application.

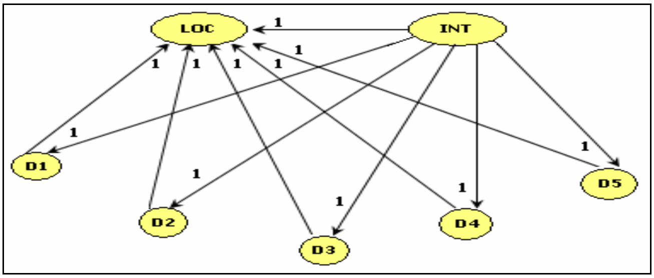


Fig. 2. The multiagent system's communication graph

According to the communication graph, we notice that the detection agents D1-5 don't communicate with each other. They rely on the localization agent LOC which is solicited by all the other agents. This is due to the fact that while detecting a failure, the detection agents send the residues' values to the localization agent. On the contrary, the agent INT sends messages and doesn't receive any one. This arises from the fact that it is an organizer agent; it transmits to the detection agents a diagnosis request and informs the localization agent to receive the residues values from the detection agents. Table 1 illustrates the several performance metrics values obtained.

Table 1. Performance metrics

Property	***Measure***	***Value***
network communication complexity	Index β	1.5714
communication degree	Index γ	0.2245
network mean traffic	Index θ	1.5714
connected components	*NBcc*	1
articulation points	*NBpa*	0
number of agents involved in communication	*NBac*	11
percent of agents involved in communication	*Pac*	100%

According to the metrics values presented in Table.1, the communication network is not complex (β index is low). γ is nearer to 0 from 1, so the MAS is characterized by a low degree of communication. In addition, the whole system is composed of only one acquaintance (only one connected component) in which all the agents participate to the communication (100% of the agents are implicated in the communication).

According to results illustrated by Fig.3, Fig.4 and Fig.5, we notice that the load isn't equitably shared and that the agents LOC and INT are involved in the communication more than the others.

Like it is shown in Fig.6, the message typology is poor; the agents use only one type of messages which is represented by the communication act INFORM. This seems to be natural according to the application needs. In fact, the communication between the several agents consists in an exchange of information; there are no sophisticated or complex conversations. According to Fig.7, the messages' sizes are almost the same. All the messages' contents have a medium complexity given that the agents send encoded matrices in accordance with the ontology defined by the application designer.

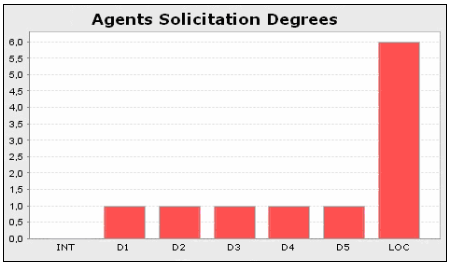


Fig. 3. The agents' solicitation degrees

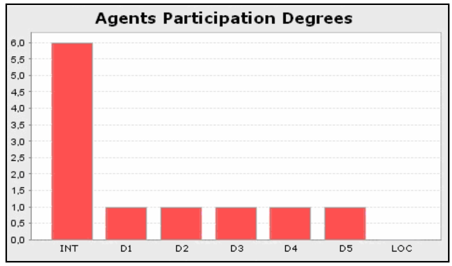


Fig. 4. The agents' participation degrees

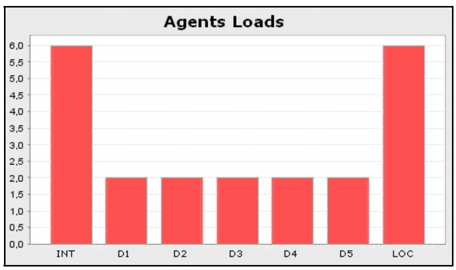


Fig. 5. The agents' loads

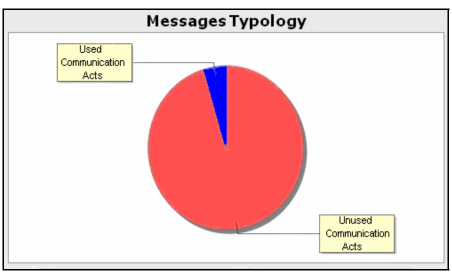


Fig. 6. Messages' sizes

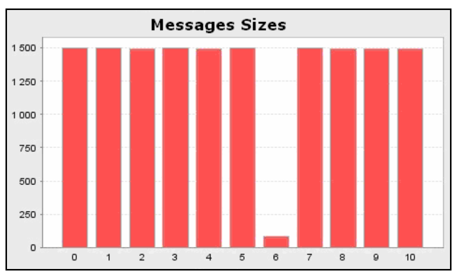


Fig. 7. Messages Typology

6 Conclusion

In this paper, an experimental approach to evaluate communication in multiagent systems was proposed. The implemented evaluation tool is composed of two modules: observation module and measurement module. The observation module consists in a spy agent that captures messages exchanged by the agents and saves them in a file. This file is then exploited by the measurement module to draw the communication graph and to calculate the performance metrics. Our evaluation covered several aspects of the communication in MAS which are: the structural, the syntactical and the statistical ones. The obtained results allowed us to validate our approach. In the future, we will focus more on the observation module and we will try to find a more generic solution to the MAS's activity observation. In addition, we will be interested in some other characteristics of the MAS.

References

1. Boissier, O., Guessoum, Z., Occello, M.: Un essai de définition de critères pour l'étude comparative de plates-formes multi-agents. In: Technique et Science Informatiques, vol 21. Hermès, Paris (2002)
2. French Observatory of Advanced Technologies: Systèmes Multi-Agents. Arago 29, pp. 27–44. Tec & Doc, Paris (2004)
3. Jurasovic, K., Jezic, G., Kuzek, M.: A Performance Analysis of Multiagent Systems. International Transactions on Systems Science and Applications 1(4) (2006)
4. Le Boudec, J.Y.: Performance Evaluation of Computer and Communication Systems. EPFL (2007)
5. Jain, R.: The Art of Computer Systems Performance Analysis. John Wiley & Sons, England (1991)
6. Willig, A.: Performance Evaluation Techniques, Fundamentals and Big Picture. Telecommunication Networks Group, Technical University Berlin (2005)
7. Lejouad-Chaari, W., Mouria-Beji, F.: High-Level Communication Protocol in a Distributed Multiagent System. In: Moonis, A., Mira, J., de Pobil, A.P. (eds.) IEA/AIE 1998. LNCS, vol. 1415, pp. 606–616. Springer, Heidelberg (1998)
8. Gondran, M., Minoux, M.: Graphes et Algorithmes., Eyrolles, Paris (1995)
9. Graph Theory: Measures and Indices, http://people.hofstra.edu/geotrans/index.html
10. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons, England (2002)
11. Saddem, R.: Conception et Implémentation d'un Système de Diagnostic à Base d'Agents. Research Master Report. National School of Computer Studies, Tunis (2006)
12. Bouabdallah, B.S., Saddem, R., Tagina, M.: A Multiagent Architecture for Fault Detection and Isolation. In: 1st International Workshop on Applications with Artificial Intelligence. Patras (2007)
13. JADE Sniffer Online Documentation, http://jade.tilab.com/doc/tools/sniffer/index.html

Using Multiple Models to Imitate the YMCA

Axel Tidemann

IDI, Norwegian University of Science and Technology
tidemann@idi.ntnu.no

Abstract. Learning by imitation enables people to program robots simply by showing them what to do, instead of having to specify the motor commands of the robot. To achieve imitative behaviour in a simulated robot, a modular connectionist architecture for motor learning and control was implemented. The architecture was used to imitate human dance movements. The architecture self-organizes the decomposition of the movement to be imitated across different modules. The results show that the decomposition of the movement tends to be both competitive (i.e. one module dominates the others for a part of the movement) and collaborative (i.e. modules cooperate in controlling the robot).

Keywords: Human Robot Interaction, Neural Networks.

1 Introduction

The ability to imitate has proven to be an important part of human cognition. Learning by imitation allows for easy transfer of learnt motor skills between individuals. Using imitation as means to program robots has become increasingly popular among artificial intelligence researchers [1]. Learning by imitation is evident when humans learn to dance. Students use visual input as teaching signals for their motor system when observing the teacher. This paper presents a modular connectionist architecture that enables learning by imitation in a simulated robot. The research focus is to understand how complex movements can be achieved from combining smaller motor primitives, and how these motor primitives are acquired. The idea is that motor primitives are learned and stored in different modules, which can be thought of as experts. The emphasis is on self-organization of these modules during execution and learning of motor primitives. The core of the architecture consists of multiple paired inverse/forward models, implemented by Echo State Networks.

2 Imitation in Psychology, Neuroscience and AI

How infants learn from adults by imitating them has been studied extensively in developmental psychology. Piaget [2] relates adaptation of sensory-motor schemas to imitation. The infant constantly tries to adapt the schemas (i.e. motor and perception stimuli) to the external world. Meltzoff and Moore [3]

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 783–792, 2008.

propose a framework called *active intermodal mapping* (AIM) that is capable of combining perception and production of actions. Crucial to the AIM is its ability to correct the behaviour of the child according to the behaviour of the teacher. In neuroscience, mirror neurons [4] are thought to enable imitation in the brain. Mirror neurons are active both when observing and performing the same action, and they are considered to form a neural basis for imitation [1], language [5] and mind reading [6]. Within the AI community, research on imitation learning is divided in two groups: solving the *correspondence problem* (i.e. the transformation of coordinate systems that occurs when visual input is used to guide motor systems) or focusing on the perception-action system (perceptual stimuli has already been processed to a meaningful representation to the agent) [1]. According to Schaal, *model-based learning* (which this brief background section focuses on) is the most interesting way of implementing imitation learning. Model-based learning consists of pairing an inverse model (i.e. controller) with a forward model (i.e. predictor). Schaal sees the paired models approach as a way to implement the AIM framework, in addition to matching the simulation theory of mind-reading [1]. Demiris [7,8] and Wolpert [9] use multiple inverse/forward models in their architectures to implement imitation learning. Demiris implement the models using different techniques (e.g. PID controllers and Bayesian belief networks), whereas Wolpert focuses on neural networks (in [10] Wolpert argues that multiple paired inverse and forward models are located in the cerebellum). The architecture of Matarić [11] has several modules that have specific pre-defined roles (e.g. visual processing, attention, motor primitives and learning modules), and has been deployed on several robot testbeds. The modules are implemented using various techniques.

3 The Multiple Paired Models Architecture

This paper presents the implementation of a multiple paired models architecture (abbreviated MPMA) inspired from Demiris' HAMMER [8] and Wolpert's MOSAIC [9] architectures. The goal is to combine the best of HAMMER and MOSAIC; the consistent inverse/forward ordering of the former (MOSAIC varies how the inverse/forward models are organized depending on whether it is observing or executing an action) and the responsibility predictor (explained shortly) and focus on self-organization from the latter.

At the core of MOSAIC and HAMMER are multiple paired inverse and forward models, which in principle are similar to Jacobs' mixtures of experts [12]. Using several models to represent different motor knowledge coincides with the modular structure of the brain. The brain may not share the actual inverse/forward coupling that is explicitly defined in HAMMER or MOSAIC, but from a computer science perspective it is an approach that is well understood for motor control [13]. Multiple models can be encode redundancy, which is crucial to ensure robustness and an important part of intelligent systems [14]. Spreading motor knowledge across several models can also be a solution to *catastrophic forgetting* [15], which occurs when the acquisition of new concepts destroys previously acquired knowledge. The representational capacity of the system can be expanded by adding new modules (although this is not done in the current implementation). A modular architecture allows for studying the decomposition of a movement into smaller submovements, and how and why these are represented by the modules. By using neural networks to implement the models, the MPMA takes advantage of both local and distributed representations. The local representation corresponds to a module, making it possible to tell where a certain concept is stored. The distributed representation (i.e. the neural networks used as building blocks of the architecture) is tolerant to noise and faulty network nodes. The architecture can be seen in figure 1. The models will be explained in more detail in the following section. Note that the term *module* is used to group the inverse model, the forward model and the responsibility predictor together.

3.1 The Concept of the MPMA

The forward model learns to output the predicted next state $\hat{x}^i_{t+1}$ given the current state x_t and the motor command u^i_t applied to the robot. The dashed arrow shows the error signal used to train the forward model; it is the difference between the predicted and actual next state.

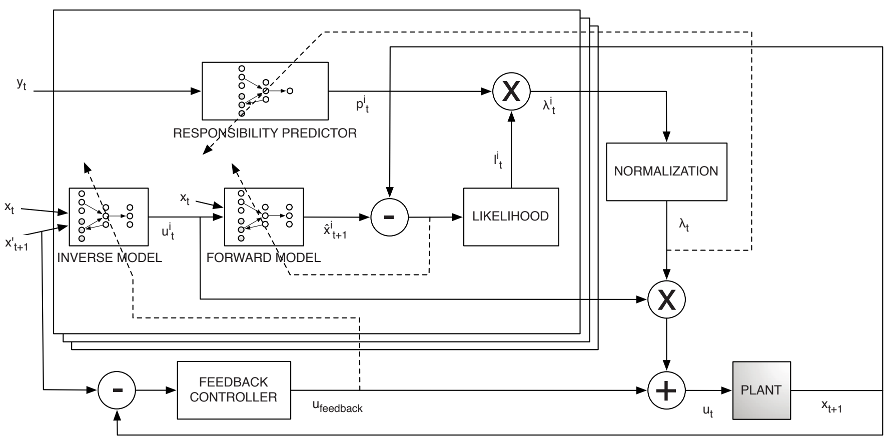


Fig. 1. The multiple paired models architecture. See the text for details. Inspired from [9] and [8].

The inverse model learns what motor commands u^i_t will achieve a desired state $x^{'}_{t+1}$ given the current state x_t. The error signal is the *feedback motor error command*, u_{feedback}. The feedback controller [16] is a simple way of pulling the system in the right direction when bad motor commands are issued. This is achieved by computing the difference between the target state $x^{'}_{t+1}$ and actual state x_{t+1}, and using the difference in *state* to issue *motor commands*. Differences in state and motor commands are normally not equal; motor commands are typically velocities or torques, whereas state are coordinates. In most robot systems there is a degree of redundancy which implies that there are many ways to achieve a desired state [13]. This crude approach ensures that a solution will be found to training an inverse model. In addition it will increase the robustness of the system, since it will act as an online correction controller when there is a mismatch between the desired state and the actual state. The influence on the system from the feedback controller will decrease with increasing performance from the system.

The responsibility predictor (RP) learns to predict how well the module is suited to control the robot *prior* to movement. This prediction is shown in the signal p_t^i, based on the context signal y_t. The RP along with the likelihood forms the λ_t^i signal. The likelihood l_t^i is a measure of how well the forward model predicts the next state. It is computed as the difference between the predicted next state and the actual next state, assuming the presence of gaussian noise with a standard deviation σ, i.e. $l_t^i = (2\pi\sigma^2)^{-1/2} e^{-|x_t - \hat{x}_t^i|^2/2\sigma^2}$. The λ_t^i values are normalized in the final λ_t vector (i.e. $\lambda_t^i / \Sigma_j \lambda_t^j$), and the normalized value serves as the error signal of the RP. The λ_t vector represents how the different modules will influence the total motor command u_t sent to the plant (i.e. robot). Before summing the motor commands of the inverse models they are multiplied with their corresponding λ value, which ensures that modules that make good predictions will have more influence over the robot than those that make bad predictions. The λ_t effectively enables switching between modules.

3.2 Input/Output of the MPMA

The desired state was the normalized elbow and wrist positions (i.e. 3D coordinates) of the demonstrator. The elbow coordinates had the shoulder as origo, and were in the range of $[-1, 1]$. The wrist coordinates were relative to the elbow coordinates in the same manner. This is also how the state of the robot was defined. This simplification of visual input was done to overcome the correspondence problem [17]. Findings in neuroscience indicate that there might exist a geometric stage where sensory input is transformed to postural trajectories that are meaningful to the motor system [18]. Demiris and Hayes [7] and previous work by Tidemann and Öztürk [19] have used the same approach, albeit with *joint angles* as input instead of position. Using coordinates as opposed to joint angles are more challenging for the inverse models, since they must learn the transformation from coordinates to joint angle velocities.

The context signal aids the modules in selecting which module is more suitable prior to movement. Wolpert uses the example of a full or empty cup as a context signal [9]. If MOSAIC had one inverse model for each case, the context signal would help MOSAIC to select the appropriate inverse model prior to movement. The experiment presented in this paper uses the context signal to represent the melody playing while dancing. Dancing in the real world requires the dancer to actively listen to the music while dancing, making the choice of context signal appropriate. The outputs of the system are motor commands to the robot, given as *joint angle velocities*, since the simulator uses joint angle velocities as motor commands (not forces, as real-world robots do). However, since the direction and speed of the motor commands are defined, replacing the simulator with an inverse dynamics controller that calculates forces to be applied on a real-world robot should be straightforward. Many robot models use joint angle velocity in addition to coordinates to describe the state of the robot. In the MPMA velocity is not used as state, since the inverse/forward models are dynamic systems with memory (see the next section), thus able to represent the changes in coordinates (e.g. velocity) internally.

4 Experimental Setup

The experiment was to imitate the dance from a song called YMCA by the Village People, a 70s disco group from New York (figure 2 shows the dance movements). This dance was chosen because it is a rather well-known dance in addition to being easy to understand when described verbally (e.g. as forming letters with your arms), with enough complexity to make it interesting for an imitation task. Movement data was gathered using the Pro Reflex tracking system at the Developmental Neuropsychology lab at the Institute of Psychology, NTNU. Five infrared cameras span a volume in which the position of fluorescent markers can be tracked over time. The markers were placed on the shoulder, elbow and wrist of the dancer. Pro Reflex sampled with a frequency of 200Hz; each 10th sample was extracted from the noisy data and used as training signals, thus the models had to predict 0.05 seconds into the future. Since the experiment consisted solely of imitating arm movements, a four degree of freedom (DOF) model of a human arm was implemented as described by Tolani and Badler [20], with a 3DOF spherical shoulder joint and a 1DOF elbow revolute joint. The simulated robot with two arms was thus described by eight degrees of freedom. Echo State Networks (ESNs) [21] were used to implement the inverse/forward models and the RP. An ESN is a recurrent neural network (RNN) that is characterized by a large sparsely connected hidden layer (typically hundreds of neurons) and

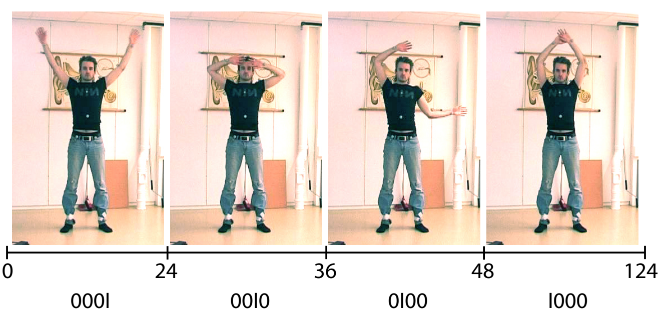


Fig. 2. Dancing the YMCA, by forming the letters Y M C A using arm movements. The numbers show at which timestep the next letter is formed in the sequence. The four-digit vector is the context signal.

by only modifying the output layer weights during training. The randomly generated input layer weights remain fixed. Achieving the desired output values is then reduced to a simple linear regression task, whereas in traditional RNNs the error is backpropagated through all the layers of the network, which is computationally more expensive. This approach exploits the massive memory capacity in the huge hidden layer of the ESN. There were 24 inputs to the inverse model: 12 signals for the current state (6 signals for each arm, corresponding to the 3D coordinates), same for the desired state. The outputs of the inverse model were the eight degrees of freedom to control both arms. The forward model had 20 inputs: 12 signals for the current state, and 8 signals from the inverse model. The forward model output was 12 signals, corresponding to the predicted next state. The responsibility predictor had four inputs, corresponding to the context signal (the 124 timestep epoch is shown in figure 2). It had a single output, predicting the suitability of the module to control the robot. The output of the inverse and forward models were in the range $[-1, 1]$ whereas the RP output was in the range $[0, 1]$. How the system performed with different number of hidden nodes was examined. All networks had spectral radius $\alpha = 0.1$ (determining the length of the memory) and the noise level $v = 0.2$ (adding extra noise internally in the network). To ensure convergence in such a high dimensional system, good error signals are crucial. An advantage of using the human arm model described by Tolani and Badler is that the joint angle rotations can be found analytically from positions of the elbow and wrist. The difference in the desired state and actual state can then be expressed as difference in rotational angles, which the u_{feedback} uses on the motor commands (which are joint angle velocities) to pull the system in the right direction. Furthermore, the desired joint angle velocities can be calculated from the desired state. This was used as a training signal for the inverse models. Note that this differs slightly from the general MPMA layout for training the inverse model (see figure 1), where only the feedback error is used. This training approach was taken to ensure that the system would converge. However, adding 1% random noise to both the input and output values during training and using a relatively high noise level in the ESN itself prohibits that the inverse models become equal. To further force the system into the desired trajectory in the early phase of training, the u_{feedback} gain K was much stronger than the output gain L of the inverse models during the first epoch (e.g. $K = 1$, whereas $L = 0.01$). L was subsequently increased with 0.1 at each epoch until it had equal influence over the motor commands as the u_{feedback}. The parameter of the likelihood function was $\sigma = 0.25$. There were two stopping criteria: 1) the mean difference between p_i^t and λ_i^t throughout an epoch had to be lower than 0.04 for all modules, and 2) the performance error $p_e = \frac{1}{ts}\Sigma_t\Sigma_s|x_t^{\prime s} - x_t^s|$ (t timesteps, s signals) had to be less than 0.06. The maximum number of epochs was set to 20. The fast training of the ESNs allowed the maximum number of epochs to be so low. The architecture had four modules. It was the designer's intention that the modules would decompose the movement according to the context signal. The system (visualization included) was implemented in MatLab.

5 Results

Four different sizes of hidden layer were examined (50, 100, 200 and 400 nodes). Each of the network configurations were run 12 times. Table 1 shows the results. See figure 3 for the imitation performed by the simulated robot. Figure 4 shows how the modules collaborate and compete when controlling the robot, and that the architecture is capable of decomposing the movement into smaller submovements. Closer examination of the λ values of each experiment reveals that *every* module had been active throughout the final epoch for all network configurations (active being defined as $\lambda_t^i > 0.1$), thus the *dominance* value is a good indicator of the degree of collaboration between modules. Particularly networks with 100, 200 and 400 nodes in the hidden layer tends to collaborate on controlling the robot. Figure 5 shows that the imitation matched closely the desired trajectory captured from the human dancer.

Table 1. The p_e is the mean of the absolute value of the difference between the desired state and actual state at the final epoch. The $\Sigma u_{\text{feedback}}/\Sigma u_t$ ratio is an indication of how strong the feedback error motor signal was relative to the total motor command at the last epoch. The *dominance* tells how often one of the modules dominated the control of the robot at the last epoch. A module is defined to dominate the others when $\lambda_t^i > 0.7$. Although all the runs for the 400 node configuration went to the maximum number of epochs, the performance was high (the average p_e was second best), and the runs were very close to the stopping criteria.

Nodes in hidden layer	min/max/avg epochs	min/max/avg p_e	min/max/avg $\Sigma u_{\text{feedback}}/\Sigma u_t$	min/max/avg dominance
50	15/20/17	0.021/0.040/0.028	0.344/0.389/0.363	0.202/0.944/0.752
100	15/20/17	0.022/0.141/0.043	0.347/0.419/0.374	0.137/0.927/0.590
200	15/20/18	0.025/0.082/0.041	0.335/0.420/0.368	0.073/0.927/0.521
400	20/20/20	0.023/0.068/0.034	0.329/0.380/0.353	0.169/0.758/0.4745

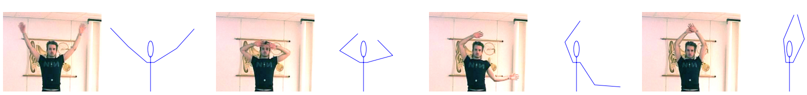


Fig. 3. Imitation of the YMCA, demonstrator on the left, imitator on the right

6 Discussion

The degree of collaboration between modules is higher than what was found in previous work by Tidemann and Öztürk [19], where it was showed that the modules tended to divide the submovements in an almost exclusive-other fashion. It could be the modules collaborate more in this experiment due to the increase in memory capacity in the ESNs which allows for storage of subtler nuances of motor control ([19] used RNNs with fewer nodes in the hidden layer). Not only is the dimensionality of this experiment higher compared to [19] (state was

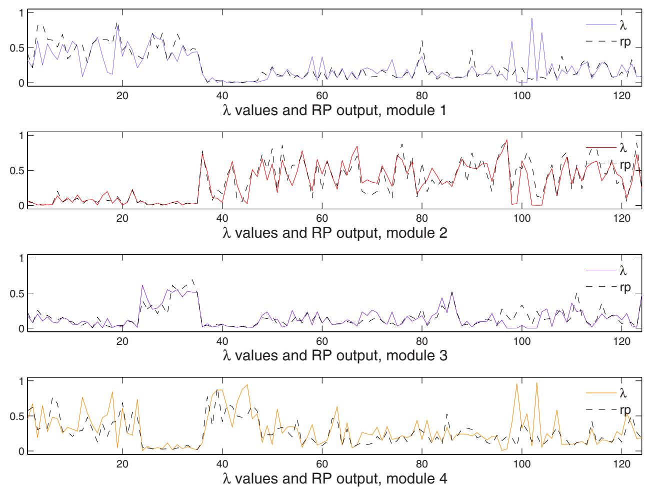


Fig. 4. Collaboration and competition between modules when controlling the robot, using 400 nodes in the hidden layer. For the Y movement (timesteps 0-23), modules 1 and 4 share most of the control. The M movement (timesteps 24-35) is controlled by modules 1 and 3, C (timesteps 36-47) by modules 2 and 4, and finally A (timesteps 48-124) where module 2 dominates with smaller (but still significant) influence from the other modules. This shows that the MPMA is capable of self-organizing the decomposition of a movement into smaller submovements. Overlapping RP and λ indicates stability, since the module correctly predicted how much it would influence the control of the robot.

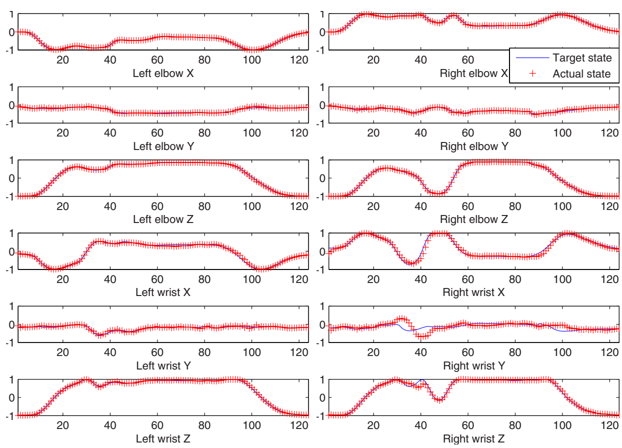


Fig. 5. Actual trajectory versus desired trajectory (same experiment as in figure 4)

described by four joint angles, and the robot had four degrees of freedom), it is also harder to learn the mapping from 3D coordinates to joint angle velocities than from joint angles (which would be simply the derivative), which could be another reason why the modules need to cooperate to a greater extent in order to successfully control the robot. It has been suggested that specific muscle synergies are shared across motor behaviours [22], a similar process could be what is occuring during the collaboration among modules in the MPMA. It should be noted that the $\Sigma u_{\text{feedback}}/\Sigma u_t$ ratio was quite high for all network configurations, roughly 1/3. The high ratio could be attributed to the noise added on input and output during training as well as the relatively high noise level used in the ESN, however this needs further investigation. Still, it is clear from the good performance of the system that using a feedback error controller helps ensure robustness. Trying to define motor primitives is tempting, although hard to do [11]. In figure 4 module 1 shares control with two modules at different times when controlling Y and M, but *how* they share control is not apparent. Perhaps the combination of the two modules defines a motor primitive, or other structures in the sensory flow that are not easily observable; figure 4 shows a decomposition of motor control that is fairly coincident with the context signal, however this is not always the case. One possible disadvantage of the YMCA movement is the symmetry of the Y, M and A movements, only the C movement breaks this symmetry. Closer examination of when transitions between modules occur reveal a dominance for switching around C, indicating that it is harder to learn. Neuroscientific findings support this claim; nonsymmetric action in bimanual movements leads to interference and takes longer time compared to symmetrical ones [23], thus understanding motor primitives may require more low-level examination.

7 Future Work

Studying how repetition influences the self-organization of submovements is of great interest. This will coincide with examining how the architecture self-organizes when longer and more complex movements are to be imitated. This might lead to discovery of a saturation level of the architecture, i.e. where more neural resources must be added to deal with the increase in complexity. Another focus is investigating how the MPMA captures attractors through self-organization (as in [24]), especially for modules that capture several primitives.

Acknowledgements

The author wishes to thank Ruud van der Weel for using the Pro Reflex system.

References

1. Schaal, S.: Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences 3(6), 233–242 (1999)
2. Piaget, J.: Play, dreams and imitation in childhood. W. W. Norton, New York (1962)
3. Meltzoff, A.N., Moore, M.K.: Explaining facial imitation: A theoretical model. Early Development and Parenting 6, 179–192 (1997)

4. Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L.: Premotor cortex and the recognition of motor actions. Cognitive Brain Research 3, 131–141 (1996)
5. Arbib, M.: The Mirror System, Imitation, and the Evolution of Language. In: Imitation in animals and artifacts, pp. 229–280. MIT Press, Cambridge (2002)
6. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mindreading. Trends in Cognitive Sciences 2(12) (1998)
7. Demiris, Y., Hayes, G.: Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model. In: Imitation in animals and artifacts, pp. 327–361. MIT Press, Cambridge (2002)
8. Demiris, Y., Khadhouri, B.: Hierarchical attentive multiple models for execution and recognition of actions. Robotics and Autonomous Systems 54, 361–369 (2006)
9. Wolpert, D.M., Doya, K., Kawato, M.: A unifying computational framework for motor control and social interaction. Philosophical Transactions: Biological Sciences 358(1431), 593–602 (2003)
10. Wolpert, D.M., Miall, R.C., Kawato, M.: Internal models in the cerebellum. Trends in Cognitive Sciences 2(9) (1998)
11. Matarić, M.J.: Sensory-Motor Primitives as a Basis for Learning by Imitation: Linking Perception to Action and Biology to Robotics. In: Imitation in animals and artifacts, pp. 392–422. MIT Press, Cambridge (2002)
12. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation 3, 79–87 (1991)
13. Jordan, M.I., Rumelhart, D.E.: Forward models: Supervised learning with a distal teacher. Cognitive Science 16, 307–354 (1992)
14. Pfeifer, R., Scheier, C.: Understanding Intelligence, Illustrator-Isabelle Follath. MIT Press, Cambridge, MA, USA (2001)
15. Ans, B., Rousset, S., French, R.M., Musca, S.: Self-refreshing memory in artificial neural networks: learning temporal structures without catastrophic forgetting. Connection Science 16(2), 71–99 (2004)
16. Kawato, M.: Feedback-error-learning neural network for supervised motor learning. In: Eckmiller, R. (ed.) Advanced neural computers, pp. 365–372 (1990)
17. Nehaniv, C.L., Dautenhahn, K.: The Correspondence Problem. In: Imitation in Animals and Artifacts, pp. 41–63. MIT Press, Cambridge (2002)
18. Torres, E.B., Zipser, D.: Simultaneous control of hand displacements and rotations in orientation-matching experiments. J. Appl. Physiol. 96(5), 1978–1987 (2004)
19. Tidemann, A., Öztürk, P.: Self-organizing multiple models for imitation: Teaching a robot to dance the YMCA. In: Okuno, H.G., Ali, M. (eds.) IEA/AIE 2007. LNCS (LNAI), vol. 4570, pp. 291–302. Springer, Heidelberg (2007)
20. Tolani, D., Badler, N.I.: Real-time inverse kinematics of the human arm. Presence 5(4), 393–401 (1996)
21. Jaeger, H., Haas, H.: Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. Science 304(5667), 78–80 (2004)
22. d'Avella, A., Bizzi, E.: Shared and specific muscle synergies in natural motor behaviors. PNAS 102(8), 3076–3081 (2005)
23. Diedrichsen, J., Hazeltine, E., Kennerley, S., Ivry, R.B.: Moving to directly cued locations abolishes spatial interference during bimanual actions. Psychological Science 12(6), 493–498 (2001)
24. Kuniyoshi, Y., Yorozu, Y., Ohmura, Y., Terada, K., Otani, T., Nagakubo, A., Yamamoto, T.: From humanoid embodiment to theory of mind. In: Pierre, S., Barbeau, M., Kranakis, E. (eds.) ADHOC-NOW 2003. LNCS, vol. 2865, pp. 202–218. Springer, Heidelberg (2003)

A Comparison of Performance-Evaluating Strategies for Data Exchange in Multi-agent System

Dariusz Król and Michał Zelmozer

Wrocław University of Technology, Institute of Applied Informatics,
Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, Poland
dariusz.krol@pwr.wroc.pl

Abstract. The growing complexity of agent systems has made network management very difficult. The paper presents the comparative study of two different open-source MAS architectures. The main output of this paper can be summarized within two statements. Firstly, we focus on the description of selected features in Aglets and Jade. Secondly, we aim to provide the programming implementation of the same distributed data exchange experiment in both systems. We illustrate this by specifying a large number of experiments within different environment configurations to build a base for comparison Aglets and Jade and to put across valuation of network configuration impact on system performance.

Keywords: Aglets, Jade, network quality, RMI, message passing, communication process.

1 Introduction

The main advantages of Multi-Agent System (hereafter MAS) are scalability, transparency, simultaneousness, robustness and interoperability. These characteristics are the main reason for insight into those systems. Of course there are some drawbacks like sensitivity to network bandwidth and openness for numerous kinds of attacks.

In this paper, we consider that the complexity of agent systems is still growing and we face decision which framework will be the most suitable for our architecture. The choice from the numerous open-source implementations developed by many universities, groups and government projects, which are accessible from the Internet, was done to pick Aglets [14] and Jade [1]. The main reasons why those two solutions were chosen are their capabilities, quality of documentations and range of provided examples.

The research is being conducted within the implemented experiments. The general idea is to send message through all agents taking part in the experiment but only once through each agent. It is similar to Hamilton's Cycle with the exception that each time agent receives message it has to ask agent broker for the next message receiver. This basic scenario is altered by network configuration like hosts distance, combinations of number of agents from 12 to 1024 and number of simultaneous messages send from 1 to 10.

By the variety of experiments we try to simulate many common environments [13] where complete system can be placed in. The other reason is the need of checking

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 793–802, 2008.

how those solutions fit to experiments in different configurations. All of these features are supported with graphs. In Section 4 we can find results of the experiments. The paper is a source of many conclusions about Aglets and Jade as well the network's configuration impact on MAS quality which are placed in Section 5.

2 Insights into Architectures

Aglets and Jade represent the same Java RMI architecture, so we compare them from the point of the developer. There are couple of issues that has influence on quality and facility. The first thing we would like to compare is the structure of the objects on the server, how agents are hold in containers. How process of finding agents is supported by each solution and how is developed the way of controlling the agents. Other thing is the cycle of the agent's life from creation through dispatching to disposal. The next issue that has great impact on informality of the developer is the complexity of different events handling. There are many different kinds of events connected not only with creation, dispatching and destroying the agents but also connected with timer or handling messages events. The way messages are handled is most crucial part of this comparison.

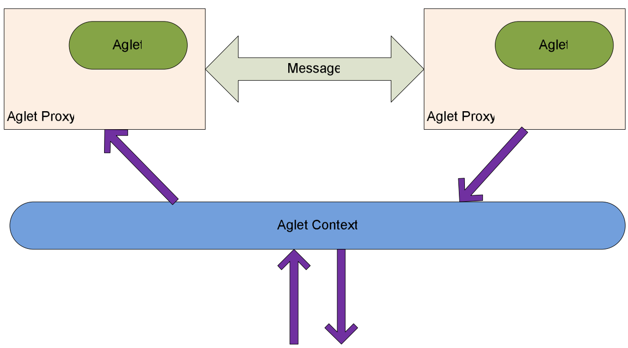


Fig. 1. Aglets architecture

In Fig. 1 Aglets' architecture is presented. It consists of Aglets which are the core of the agent. Aglet Proxy provides the access to Aglet and provides security that stops other aglets from calling some of public methods. It is some kind of firewall that provides protection from malicious agents. The other task is to provide transparency of Aglet residence. There is no difference in use of Proxy no matter if Aglet is on the same node or on a different one.

Aglet Context is an object that provides runtime environment that handles Aglets. Aglet Context supports Aglets with possibility of creation of new Aglets. When Aglet is dispatched it leaves the Context and is no longer available from this Context. But there can be created Remote Proxy that will keep the Aglet in control.

There are four types of messages that can be sent by Aglet: Now-type, Future-type, One-way-type and Delegation-type. The Now-type messages block the actions performed by sender to the time the answer is received. Future-type message is asynchronous, that means that it does not inflict on current work of the sender. The One-way-type message also does not block the execution of the sender but it differs that it is put at the tail of the queue. The last type is Delegation-type message, which is about to get information from the receiver whether message was handled or not. This information is passed without participation of developer.

As a basic manual we would recommend [7] and as a complete work about Aglets [14].

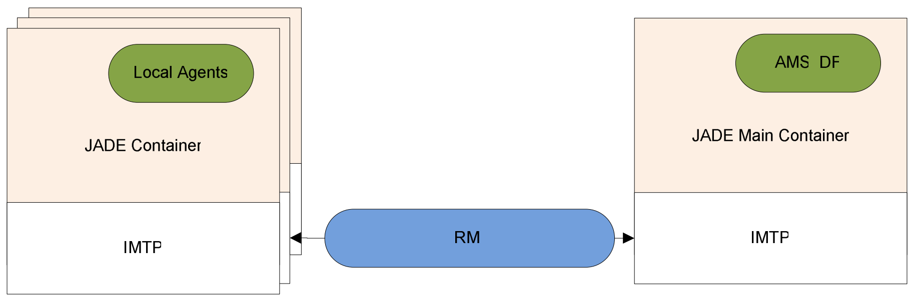


Fig. 2. Jade architecture

Jade's architecture consists of Main Container which is created automatically with runtime environment. It consists as well Agent Management System (AMS) and Directory Facilitator (DF) which support agent existence. As the Fig. 2 shows there can be created additional containers but they store only local object as an agent. As a proxy Jade has an Agent Controller.

The service registry in Jade is well developed. There are several operations that can help controlling the services. First step is registration of the service; it is usually done by the agent. It is always possible to de-register the service. Then other agents can register to the service and make the DF Lookup which returns the list of existing agents registered in this service. The last thing is DF subscription service which works to deliver the information about registration of new agents.

Structure of the massages in Jade allows getting and setting all of the attributes. One massage can have many receivers. Multicasting can be done also by registering in the appropriate service and sending message to the service. Sending messages does not stop the sender execution in any time when it waits for the reply. Massages in Jade are passed by Message Transport Service (MTS).

Jade supports the developer with lots of tools available from graphical Interface. It allows creating Agent Sniffer that shows all of the interactions between agents. The other feature is Introspector Agent. GUI also allows us to add the remote platform from remote host and supports all of the methods available from the command line and Jade API.

We would recommend [2] as a complete programmers guide. The general look at the Jade can be found in [1].

3 Testbed for Message Exchange

The main purpose of the tests is to find out how characteristics of the different architectures inflict on the results. The other reason for testing is comparison of two different MAS implementations: Aglets and Jade. It would be crucial to find the differences between these two architectures mainly in effectiveness, scalability and robustness. Bellow some experiment scenarios and network configurations are given. There are three types of the Jade agents taking part in experiments:

- Agent Manager is the broker and is responsible for general management of all processes: sends initial message in order to create agents on the local host; just before beginning of the communication process sends message to destroy randomly selected agent to simulate breakdown; as well during communication process sends destroy messages to agent to simulate breakdown in underway of the experiment; there is only one Agent Manager needed for the whole experiment.
- Agent Father is responsible for receiving messages from Agent Manager in order to create all of the Child Agents on the local host with specified parameters; it is also responsible for destruction of the agents before start and during the communication process; each host taking part in experiment needs one running Agent Father.
- Agent Child is responsible for receiving messages from Agent Manager; it sends messages to other agents and asks Agent Manager for the next agent to deliver the message.

There are only two types of the Aglets agents taking part in experiments. Agent Father is not needed because all of the Child Agents have to be created locally and then dispatch to remote locations.

- Agent Manager is the broker and is responsible for general management of all of the processes. Agent Manager creates all of the Child Agents and dispatches them to specific URL addresses. Than it starts the communication process.
- Agent Child is responsible for receiving messages from Agent Manager, sends messages to other agents and asks Agent Manager for the next agent to deliver the message.

Unified process of communication for two systems is presented in Fig. 3. There are eight steps in the communication act between three agents:

(1) Agent Manager randomly selects one from all of the Agent Child (Agent 3) and sends it the Get next message.
(2) When Agent Child (Agent 3) receives Get next message it sends Get next message to the Agent Manager to get the IP address of the host and name of next Agent Child.
(3) Agent Manager sends the reply to Agent 3. Agent 3 receives the reply from the Agent Manager with the coordinates of the next agent.
(4) Agent 3 sends message Move on to the next agent (Agent 2).

(5) Agent 2 receives message from Agent 3 and sends message Get next to the Agent Manager.
(6) Agent Manager sends the message to Agent 2. Agent 2 receives the reply from the Agent Manager with the coordinates of the next agent.
(7) Agent 2 sends message Move on to the next agent (Agent 1).
(8) Agent 1 receives message from Agent 2 and sends message Get next to the Agent Manager. Agent Manager ends the process.

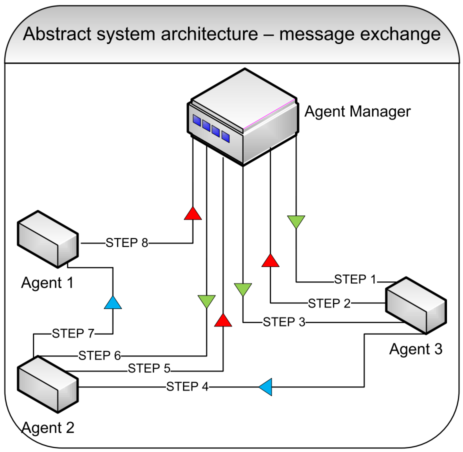


Fig. 3. Process of communication

There are four main characteristics of the experiments that are combined with each other to achieve a great number of different configurations:

- Different locations of the host: local network, remote network, presence of firewalls. We try to find out how different network configurations as well as firewalls affect the experiments results and how Aglets and Jade cope with different configurations.
- Number of hosts taking part in experiment varies from one to three. We try to observe how distribution of agents between hosts can make system more efficient (e.g. what is better two hosts with 100 agents on each host or one host with 200 agents).
- Total number of agents taking part in experiment was selected from the set: 12, 48, 128, 512, and 1024. We also observe how increasing the number of agents effects usage of resources.
- Numbers of massages are selected from set: 1, 2, 5, and 10. We try to incriminate the hosts not only by increasing number of agents but also by sending simultaneous messages. We want to find out how concurrent agents share processor, memory and network of occupied host.

Experiments are organized in groups. First group is done by the presence of firewall. Then experiments have specific number of hosts taking part in experiment and different network configuration. All of the groups that are presented below come as a result from combination of the total number of agents and the number of messages.

There are five different scenarios: one host, two hosts in local network, two hosts in remote locations, and two hosts in local network with remote broker, host with local broker and remote host. All scenarios were performed without random Child Agents destruction. Results of the experiments are presented in the next Section. Below are presented just two of those scenarios to make an outline of them.

Two remote hosts
In this configuration Agent Manager shares the host with Agent Father and Child Agents; there is another host Agent Father and half of the Child Agents placed in the remote location. This scenario is presented in Fig. 4. There are two kinds of scenarios: the worst and the best. The worst scenario is when agents communicate one by one between two hosts, so there is much longer time of information exchange. The best scenario is an ideal situation when first the agent on one host communicates with each other and then all of the agents on the other host send the messages.

One host with local broker and one remote host
This configuration is similar to the previous configuration but this time Agent Manager does not share the host with other agents. Agent Manager is placed in the local network with one of the host with half of the agents; the other half is placed on the remote host. This scenario is presented in Fig. 5. There are two kinds of scenarios: the worst and the best either and nothing has changed in this case.

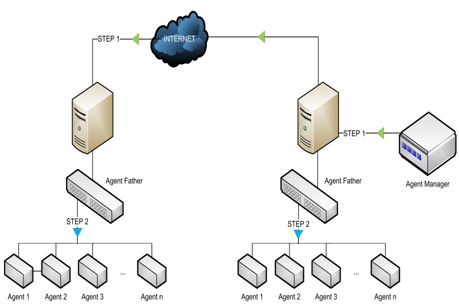


Fig. 4. Two remote hosts scenario

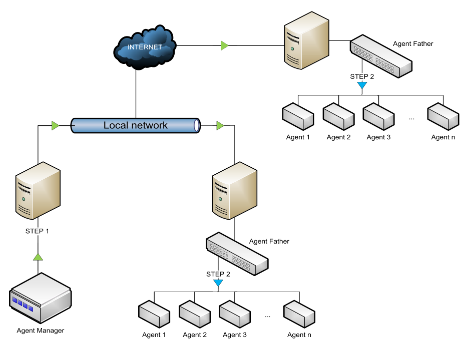


Fig. 5. One host with local broker and remote host

4 Performance Evaluation

In order to prove the differences between frameworks the following experiments performed on Aglets and Jade are prepared: two remote hosts, one host with local broker and one remote host, local network and remote network.

Two remote hosts
The results of two remote hosts experiment are presented in Fig. 6. In this case it is easy to notice that Jade has performed better. We can also find out that in some cases,

when number of messages are large, two different hosts behave differently. This may be effect of various memory and processor configuration. As it was with 48 agents (not shown here) it is the same for 1024 agents. As we can see Jade results are better that Aglets'. It is worth to remind that in local network testbed Aglets present better results for lower number of messages and worse results for higher number. Here we cannot notice this dependence. Jade presents better in all conditions.

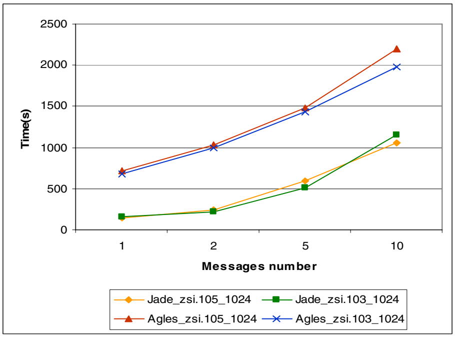


Fig. 6. Aglets, Jade - two remote hosts, 1024 agents

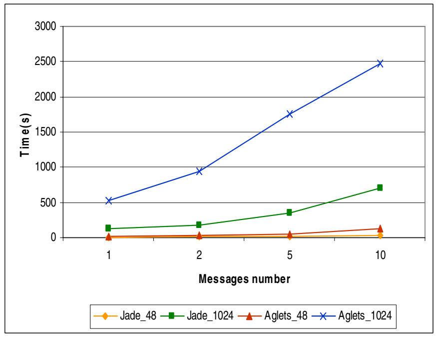


Fig. 7. Aglets, Jade - one host with local broker and remote host

One host with local broker and one remote host

As we can see in Fig. 7 Aglets behave much worst when 1024 agents are working. If we consider 1024 agent: Jade's time of 10 messages is 3.5 times lower than Aglets'. If we consider 48 agents: Jade's time of 10 messages is not so different, this time it is about 3 times lower.

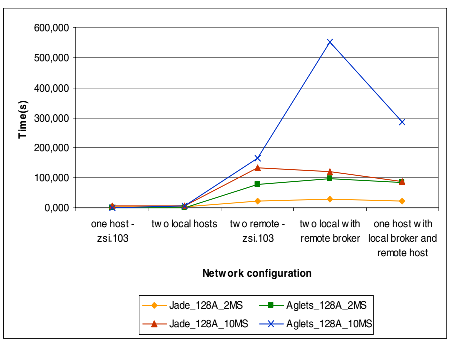


Fig. 8. Aglets, Jade - 128 agents, 2 and 10 messages

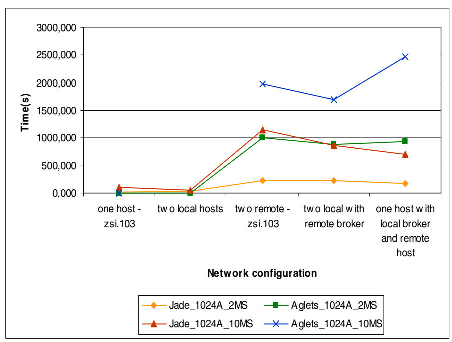


Fig. 9. Aglets, Jade - 1024 agents, 2 and 10 messages

Local network

It was very easy to predict that as Fig. 8 shows one host and two local hosts are the best configurations. It is very interesting that two remote hosts is the worst configuration despite in the last two: two local hosts with remote broker and one host with local broker and remote host there are three hosts communicating. As we can see for two first configurations Aglets are generally equal to Jade but with growth of distance and number of host the difference starts to increase. For the last configuration 128 Jade's agent with 10 messages has equalized 128 Aglets' agent and 2 messages. That means that in this case Jade was able to exchange five times more massages in the same time.

Situation is similar if we increase number of agents. As we see in Fig. 9 nothing changes if we add agents.

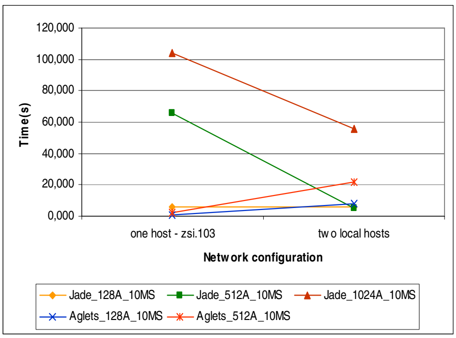


Fig. 10. Local network

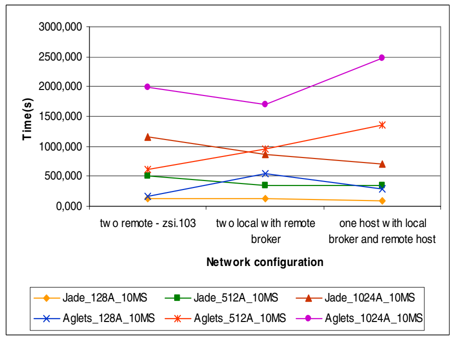


Fig. 11. Remote network

Remote network

For local solutions results seem to be equal (SEE Fig. 10) but for remote network configurations (see Fig. 11) Jade presents a big advantage. We can also observe the same case as it was with 128 agents, namely for last two configuration Jade's 1024 agents with 10 messages finished test equals to Aglets' 1024 with 2 messages. That means that Jade is again 5 times faster than Aglets in this case.

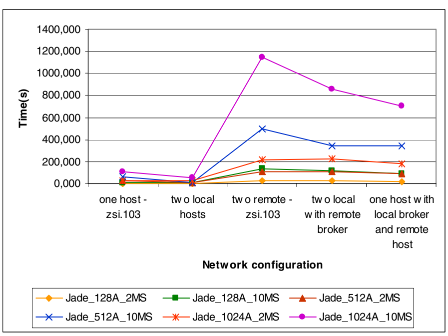


Fig. 12. Jade - 128, 512, 1024 agents, 2, 10 messages

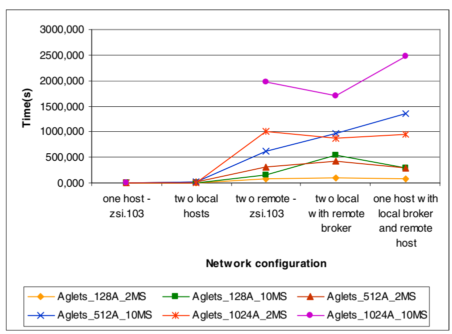


Fig. 13. Aglets - 128, 512, 1024 agents, 2, 10 messages

At the end it would be interesting to compare these results in two smaller groups: one host with two hosts in local network and other three configurations together. In

Fig. 10 we can spot that Jade performs better when there are two local hosts in configuration. Results of Aglets tend to rise with number of hosts. If we consider remote locations Jade seems to present best results in one host with local broker and remote host configuration (see Fig. 11).

Despite the encouraging results already achieved, we are aware that our approach still has some limitations. Aglets behave more changeably, we cannot generalize which configuration is the most suitable as it was in Jade. If we consider Jade separately we can notice the tendency between different configurations as in Fig. 12. If we consider Aglets separately in Fig. 13 we can notice that there is no tendency between different configurations as it was in Jade. In this case some configurations behave better in some conditions and it is hard to say which one is the best.

5 Conclusions

All of the developed experiments have proved main characteristics of MAS like transparency, scalability and robustness. The simplest way to exemplify the level of complexity is to point that the most intricate scenario consisted of 1024 agents placed on three hosts in different locations that exchanged 10 simultaneous messages between each other. Concurrency in messaging is the only feature that has not been fully satisfying in all experiments. Number of simultaneous messages send has almost linear influence on time of all messages exchange. It does not prove that these solutions are not concurrent but it shows that massage exchange has some unknown constraints. It is probably caused by weak messaging protocols that are shared by agents what generates queues.

In this paper, we focus on comparison of two Java RMI implementations: Aglets and Jade. The broad range of different scenarios allows us to make many conclusions about those two solutions. If we consider distribution of agents between local and remote architecture it is easy to notice that Jade performs better in remote solutions and Aglets in local. The biggest drawback of Aglets is communication between faraway hosts, because time of connection during tests seems to vary more than it is with Jade. If we look at dependence on number of agents and system overload Aglets notes better result and is more immune to this kind of aggravation than Jade. The last issue that differences between those two implementations is vulnerability on overload of the systems. Aglets seem to be less vulnerable on this matter.

If we look at the presented experiment we might make some conclusions about the impact of the network configuration that agents were placed. If one of the hardest experiment configurations is considered, as it was mentioned before, 1024 agents and 10 messages, configuration of two hosts in local network has presented the best results. If we consider configuration including remote hosts configuration of one host with local network broker and remote host has accomplish better results than two remote hosts configurations and two local host with remote broker. This shows that distribution of the tasks is crucial in this kind of solutions. The other case that experiments have proved is the slight influence of firewall on results. We can observe just a minimal but regular increase in all compared test.

References

1. Bellifemine, F., Caire, G., Greenwood, D.: Developing Multi-Agent Systems with JADE. Wiley Series in Agent Technology (2007)
2. Bellifemine, F., Caire, G., Trucco, T., Rimassa, G.: Jade Programmer's Guide, JADE 3.5 (2007)
3. Ciobanu, G., Lucanu, D.: A Specification Language for Coordinated Objects. In: Specification and Verification of Component-Based Systems Workshop (2005)
4. Ciobanu, G., Lucanu, D.: Specification and Verification of Synchronizing Concurrent Objects. In: Integrated Formal Methods 2004. LNCS, Springer, Heidelberg (2004)
5. Ciobanu, G.: Coordination and Self-Organization in Multi-Agent Systems. In: SDA 2006, Jinan, China (2006)
6. Day, K., Al-Ayyoub, A.E.: Topological properties of OTIS-networks. Parallel and Distributed Systems, IEEE Transactions on 13(4), 359–366 (2002)
7. Ferrari, L.: The Aglets 2.0.2. User's Manual (2004)
8. Frelechoux, L., Osborne, M., Haas, R.: Topology optimization of IP over ATM. In: Universal Multiservice Networks, ECUMN 2000, pp. 122 – 131 (2000)
9. Grama, A., Gupta, A., Kumar, V.: Isoefficiency Function: A Scalability Metric for Parallel Algorithms and Architectures. Department of Computer Science, University of Minnesota (1993)
10. Gruer, P., Hilaire, V., Koukam, A., Cetnarowicz, K.: A formal framework for multi-agent systems analysis and design. Expert Systems with Applications 23, 349–355 (2002)
11. Haeuser, J., et al.: A test suite for high-performance parallel Java. Advances in Engineering Software 31, 687–696 (2000)
12. Jamakovic, A., Uhlig, S.: On the relationship between the algebraic connectivity and graph's robustness to node and link failures. In: Next Generation Internet Networks, 3rd EuroNGI Conference on 21-23, pp. 96–102 (2007)
13. Kolp, M., Giorgini, P., Mylopoulos, J.: Multi-agent architectures as organizational structures. Auton Agent Multi-Agent Sys 13, 3–25 (2006)
14. Lange, D.B., Oshima, M.: Programming and Deploying Java Mobile Agents with Aglets. Addison-Wesley, Reading (1998)

The Emergence and Collapse of the Self Monitoring Center in Multi-agent Systems

Asaki Nishikawa

Graduate School of Science and Technology, and, Department of Earth and Planetary Science, Faculty of Science, Kobe University, Nada, Kobe, 657,Japan
asakiasakiasaki@gmail.com

Abstract. We studied the emergence and collapse of the central agent in a system which consists of homogeneous agents. To do this study, we defined a simple model of a situation, i.e., cooperations with the exchange of operational demands. As a result, an agent which gives orders as a central power and has a microcosm of the whole system appeared. Next we pointed out an efficiency problem and argued how to resolve it. We also shows its side effects relating to collapses of the central agents.

Keywords: multi-agent system, self reference, emergence, consciousness, centerity, parallel computation.

1 Introduction

Traditionally, in researches of complex systems, researchers focus their attention on functional emergences in decentralized distributed systems[1][2]. However in the many real complex systems, there is a single center which operates a large amount of control. For example such systems are a brain in a body, a nucleus in a cell, a leader in an organization, etc... There are common features in such systems, i.e., 1) the center in the system controls cooperation of the whole system, and 2) the center has a microcosm of the whole system.

A model which shows the emergence of such centers is not presented as far as we know. There is some related research [3][4][5][6]. We will omit their description due to space limitation. We focused on the problem of how to model the emergence of centers with the above features in a multi-agent system by self-organization.

2 Model

2.1 Our Model for Origin of Center

In our model, a society is composed of N agents. Each agent has one ability and one demand. An ability of each agent is executed when the execution is demanded by another agent. A demand is a request for executing the ability of another agent. We model a situation where each agent demands other agents' ability. We call this situation "exchange of demand" based on the model[7]

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 803–812, 2008.

As an exchange, we pick up agent A and search for agent B that has the ability agent A demands. In the case in which agent A coincidentally has the ability agent B demands (we call this situation “double coincidence of demands” like [8]), the exchange is successful, and they exchange their abilities. However, exchanges are not formed frequently because double coincidences of demands are rare.

We add another concept “authority” into our model to solve this situation. First, the authority of agent A from the viewpoint of B increases when the following condition is met: Agent A is the executor of B’s demand (we call this a “server”) and B is the requester (we call this a “client”), and agent A refuses to execute the demand of B. Authority of A for B increases in a “one-sided love”-like situation. That is to say, A is able to execute B’s demand but B is not able to execute A’s demand. In addition, in this paper “client” means an active agent searching for an agent that can execute his demand, and “server” means a passive agent that is sought for by a client.

Furthermore, the steps below are added to the demand exchange process. If authority of client Y for server X is higher than x(a real number constant shared by all agents), then server X executes client Y’s demand, but server X’s demand is suspended. Agent X delegates its demand and ability to client Y, expecting that in the future the demand will be executed.

We call this delegated demand suspended demand(SD) and the ability proxied ability(PA). These are memorized by the agent until executions. PA is used as follows. A client Y that has some PAs can execute a server’s demand by using one of its PAs instead of its own ability. Conversely, SD is used when the agent Y is a server. The agent Y that has SD can choose a demand from SDs instead of its own demand, when a client can execute the SD. When a SD is consumed, a notification is sent to the agent that has delegated the SD, and its demand is reset because its demand is executed.

We add another condition to this model to make it more realistic. The condition is that when an client agent executes its server demand by using PA, then a notification is sent to the agent that delegated its demand and ability. If notification of using SD did not come after PA execution notification until certain period (CT, shared by the system), then the agent that delegated its PA and SD punishes the agent abusing the PA(This is “abuse” because expectation of execution of its suspended demand is betrayed. The agent only used the delegater’s ability without reward). Punishment is implemented as a constant decrement of the delegater’s authority over the receiver agent. SD is retrieved from the agent.

We call one exchange *step*, and N times of steps a *turn*, as seen below.

2.2 Details of the Model

We make a detail description of our model in this section.

Variables. Our model consists of N agents. An agent i has several attributes.

– Demand $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{in}) : d_{ij} = 0$ or 1, $\sum_j^N d_{ij} = 1$
  Here j is the index of ability that an agent i demands. If $d_{ij} = 1$ then agent

i demands ability j and if 0 then does not. In this model, an agent only demands one ability at a time, so the sum of d_i is one. Demand vector is randomly initialized, i.e., $d_{ij'} = 1$ for random j'.

- Ability $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{in}) : a_{ij} = 0$(in the case $i \neq j$), $a_{ij} = 1$(in the case $i = j$)
  Ability vector represents an operation that an agent i can execute. If $a_{ij} = 1$ then agent i can reply to the demand that needs index j. In our model each agent has one ability. Here for ease, agent i has ability i. Ability is fixed and does not decrease, so this vector does not change in a time series.
- Authority $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{in}) : 0 \leq v_{ij} \leq N$, $0 \leq \sum_j^N v_{ij} \leq N$
  This vector represents authority of agent j for i. In other words, if $v_{ij} = r$ then agent i feels authority intensity r(positive real number) from j. Because parameter x that we define below has relative value to N and we compare v with x, we impose summation limitation to v. Initial values are all zero.
- Suspended Demand $\mathbf{sd}_i = (sd_{i1}, sd_{i2}, \ldots, sd_{in}) : sd_{ij} = 0$ or 1
  This vector represents suspended demands delegated to an agent i. If $sd_{ij} = 1$ then the agent i has suspended demand that wants ability index j. Where this suspended demand comes from is represented by another data. Suspended demand j is consumed when agent i chooses sd_{ij} as alternative to its own demand. Initial values are all zero.
- Proxied Ability $\mathbf{pa}_i = (pa_{i1}, pa_{i2}, \ldots, pa_{in}) : pa_{ij} = 0$ or 1
  This vector represents proxied abilities an agent i has. If $pa_{ij} = 1$ then agent i has proxied ability from j. Because agent i has ability i in our model, we have no need to consider where this ability came from. Proxied abilities are a temporary copy of the ability from other agents, so if the agent i uses pa_{ij} then its value is reset to zero. Initial values are zero.

Dynamics

One Step of Model One exchange of demands is modeled as below.

1. **Selection of Client**

   The system selects agent c randomly. We call this agent a “client”.

2. **Selection of Server**

   The client selects agent s that can respond to its demand. We call this agent a “server”. The client selects the server from all agents in the system by the criterion below. The client searches for agents that have an ability or a proxied ability w that it wants i.e., the client selects an agent s that satisfies $d_{cw} = 1$ and ($a_{sw} = 1$ or $pa_{sw} = 1$).
   If both conditions are satisfied, then we choose the agent selected by pa.

3. **Exchanging of Authority Vectors**

   Before an exchange, the client and the server exchange their views of authority. This is modeled by averaging their authority vectors, i.e.,

$$v'_{cj} = v'_{sj} = \frac{v_{cj} + v_{sj}}{2}, \text{ for each j}$$

We normalize authority vectors as below.

$$\text{if } \sum_{j}^{N} v'_{cj} \geq N, \text{ then } v''_{cj} = v'_{cj} \frac{N}{\sum_{j}^{N} v'_{cj}}, \text{ for each j}$$

We also normalize v_s in the same way.

4. Judging Cooperativity and Execution of Demand

From the definition, the server has the ability to meet the client's demand, however reverse conditions are not tested yet.

The first case is the client can meet the server's demand or the suspended demand i.e., the client's ability can meet the server's demand (or the suspended demand) by chance or the client's proxied abilities meet the demanded ability $((d_{sw'} = 1 \text{ or } sd_{sw'} = 1) \text{ and } (a_{cw'} = 1 \text{ or } pa_{cw'} = 1))$.
In this case, the exchange of demands succeeds, and the agents execute their abilities. Because there is no implementation of abilities, "real" executions do not occur in our model. We count the executions formally. We call this quantity $e(t)$. t means "turn". It unites N times of exchanges.

(a) Using Ability

Each agent executes its ability, the system increments $e(t)$ by 2.

(b) Using Proxied Ability

The client uses its proxied ability. The system resets $pa_{cw'}$ to zero. It increments $e(t)$ by 2.

If the proxied ability is used, then a waiting agent that delegates this ability starts a countdown for punishment. If its suspended demand has already been executed, the countdown does not start. Because agents that share the same ability do not exist in our model, only one waiting agent exists.

The client satisfies its demand and its demand is reset to zero. The server needs more complicated procedures. If the server's suspended demand met, the server's own demand still persists. Of course, if the server's demand was satisfied, both are reset to zero ($d_{cw} = d_{sw'} = 0$).

If the server's $sd_{sw'}$ is used, the following procedures will be applied. 1) $sd_{sw'}$ is reset to zero. 2) We call the agents "tracers" of w' when they wait for the suspended demand $sd_{sw'}$ to be satisfied. The tracers memorize that its suspended demand to the agent s was satisfied, and there is no need to punish s. All suspended demands are satisfied simultaneously here. If countdowns have already started, the tracers stop them. 3) The tracers' own demand is reset to zero.

Next, we treat the case where the client cannot meet both the server's demand and the suspended demand.

(a) Delegating Demand and Ability

If the server feels authority of the client more than x(i.e., $v''_{sc} \geq x$), the server suspends and delegates its demand to the client (i.e., $sd_{cw'} = 1$).

And the server also delegates its ability to the client c. i.e., $pa_{cs} = 1$. Note that x is a constant shared by all agents.
In this case, only the server executes the client's demand, so the system increments $e(t)$ by 1. The client's demand is reset to zero(i.e., $d_{cw} = 0$).

(b) Declining Demand

If the server feels authority of the client less than x(i.e., $v''_{sc}x$), the server declines the client's demand. In this case, the authority of the server increases by 1.0 (i.e., $v''_{cs} + 1.0$). Because no execution occurred, $e(t)$ does not change.

5. Updating Demands

As a consequence of these processes, d_c or d_s is reset to all zero, then for randomly chosen index k, d_{ck} or d_{sk} is updated to 1.
When update of the demand of the agent u happens(i.e., d_{uo} became $d_{uo'}$), all suspended demands of u also change simultaneously as follows. 1) The system searches all agents that have the suspended demand including the old demand index o of u. 2) For each agent t of 1), the system checks sd_{to} has other waiting agents than agent u. 3) If no other agent is waiting sd_{to}, then $sd_{to} = 0$. 4) sd_{tk} is updated to 1 to synchronize the new demand of u.

One Turn of Model. We call these five processes one 'step'. As mentioned above, N steps consist of one turn. With each turn, the system chooses one agent k and changes its demand randomly. This prevents the system falling into equilibrium with no exchange.

1. Starting Countdown

As mentioned above, if a proxied ability is used, an agent p that delegates the ability counts down until reaching the tolerance limit CT. CT is a constant value shared by all agents (unit of CT is 'turn'). This countdown halts after an execution of the suspended demand from p.

2. Punishment

If the countdown reaches zero, then the agent p punishes the agent q that abuses p's proxied ability. The procedure of punishment is as follows. 1) Authority of q for p is decreased by pv, a constant value of the system (i.e., $v_{pq} - pv$). 2)pa_{qp} is reset to zero. 3)sd_{qm} is updated by the same way of updating demands. m represents the agent q's current demand index. Note that d_{pm} is still 1 because this demand has not been satisfied by anyone yet.

3 Results

3.1 Emergence of a Central Agent

After certain turns of our model's execution, a special agent that can be called a center emerges. Fig.1 shows an example of time series.

In this condition, around 500 turns, the agent that has a very large amount of pa,sd,v appears. This means only this agent has a microcosm of the system in a view of demands and abilities. Simultaneously, execution amount of the system

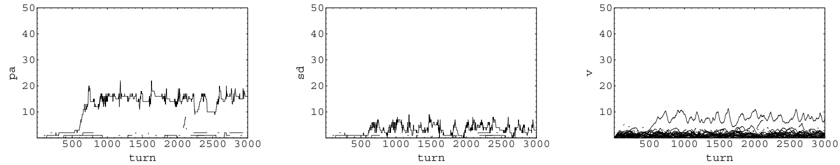


Fig. 1. We draw a time series of pa, sd, v for all agents in the system. Parameters are $N = 50, x = 10, CT = 2, pv = 1.0$. From here, pa, sd mean $\sum_j^N pa_{ij}$, i.e., the size of these for each agent i. v represents $\frac{\sum_j^N v_{ij}}{N}$, i.e., the average of v_i from all agents.

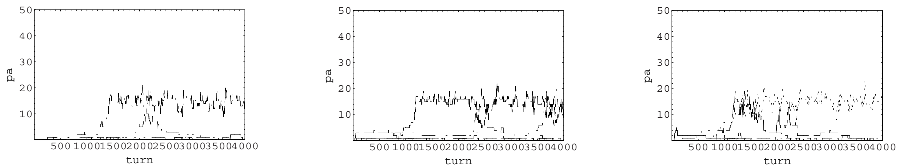


Fig. 2. Several patterns of time series of pa in the same parameter condition as Fig.1, however with different random seeds. A small center emerges(*left*). Two centers appear alternatively(*center*). Multiple centers coexist(*right*).

also increases. That is to say, the difficulty of double coincidence of wants is resolved after the center agent arises. We call an agent with such a property a "center". Here it is a qualitative notion; everybody who sees these graphs may detect the central agent. However, for making statistical graphs we will define the criterion to judge a center later.

Fig.2 shows several time series with different random seeds. With these graphs, we notice that the central agent may be multiple, and collapse after a certain period. However in this condition, for a relatively large amount of period, a single central agent is dominant.

Fig.3 shows contents of center agents as a microcosm of the system. In Fig.4, we investigate the initial delay for central agents to emerge by using statistical view of several time series, and the distribution of center durations until collapse. By distribution figure, we can see long-tail structure of durations of center.

3.2 Efficiency of the Central Agent

In Fig.5, we show a system efficiency achieved by the emergence of central agents. We can understand from this graph, around 2500 turns, that efficiency of the system increases. However we can also know that this efficiency improvement does not continue if the system's size N became large. But we have already found the way to improve this inefficiency.

New Selection Methods. We suggest two methods that resolve this inefficiency. Cause of this inefficiency relates selection probability of a center agent as

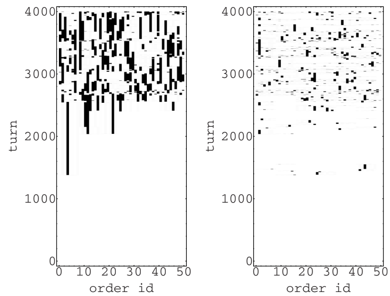


Fig. 3. The contents of center agent c. X-axis is a demand id. Y-axis is a turn. We draw a black square when $pa_{cj} \neq 0$ or $sd_{cj} \neq 0$ for each turn's end. So, these pictures represent demands and abilities which the central agent gathered. In this picture, after 2500 turns a central agent appears.

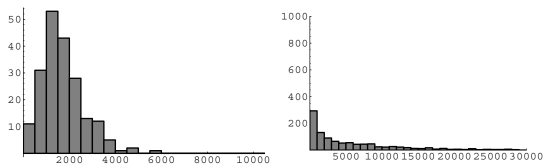


Fig. 4. *Left* is the histogram of turns that pass until a first center agent emerges. X-axis is the turn before the center agent appears. Y-axis is the number of the agents included in a bin. Parameters are the same as Fig.1. One time series is 30000 turns, and we accumulate 200 time series with different random seeds. Threshold for defining the center is $\frac{1}{100}\sum_{t-100}^{t} pa(t)_i \geq \frac{N}{6}$. The bin is 500 turns. *Right* is the duration statistics of center agents. X-axis is the duration (the bin is 1000 turns). Y-axis is an amount of centers included in the bin. One duration is measured as follows. The judging method of appearance for centers is the same as above. After an appearance of the center, we increment the duration count until the center disappears. Disappearance threshold is $\frac{1}{100}\sum_{t-100}^{t} pa(t)_i \leq \frac{N}{6}$.

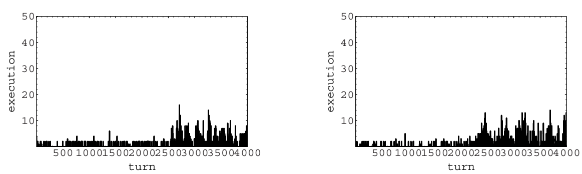


Fig. 5. The time series of the sum for all executions in the system per turn. *Left* is $N = 50, x = 10$, *right* is $N = 100, x = 14$. All parameters except N and x are the same as Fig.1.

client. If the center is selected as a server, the center's demand is rarely met by a client. However if the center is selected as a client, the probability of a server agent's demand met by the center is higher than the reverse condition.

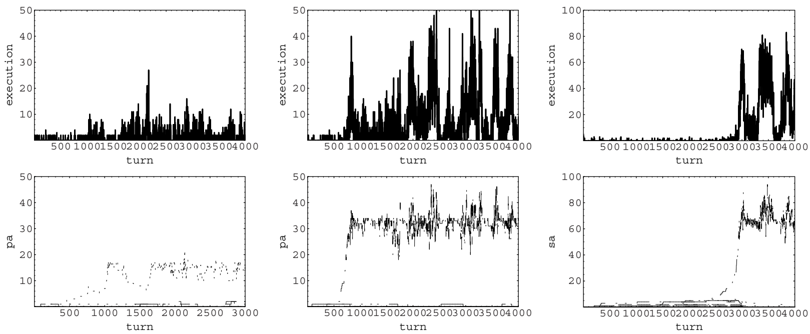


Fig. 6. Execution series and *pa* series for the system with SD selection. *Left* is $N = 50, x = 10$, *center* is $N = 100, x = 14$, *right* is $N = 200, x = 15$. All parameters except N and x are the same as Fig. 1. Note that only *right* figure's X-axis limit is 100, others are 50.

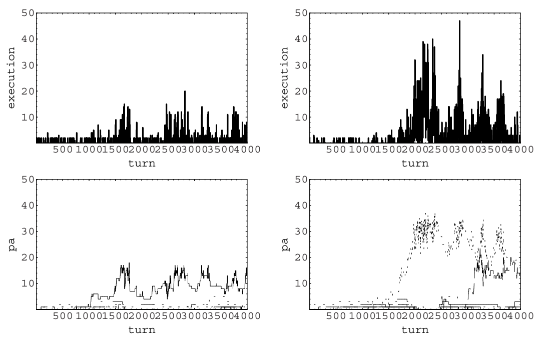


Fig. 7. Execution series and *pa* series for the system with PA selection. *Left* is $N = 50, x = 10$, *right* is $N = 100, x = 14$. All parameters except N and x are the same as Fig. 1.

To resolve this problem, we modify the selection method of clients as follows. Instead of random selection of clients, we choose a client according to the probability function $S_{sd}(i) = \frac{sd_i}{\sum_i^N sd_i}$. We call this selection method SD selection. In a similar way we can define PA selection by function $S_{pa}(i) = \frac{pa_i}{\sum_i^N pa_i}$.

Result of New Selection Methods. In Fig. 6 and Fig. 7 we show the effect of new selection methods. Both methods show prominent improvement of efficiency with functionality of the central agents with large N.

It may seem both selection methods have the same effect. However we have a criteria to distinguish the two methods. Fig. 8 shows distribution of center agent duration until collapse, with new selection methods. We can understand that by PA selection, durations of center become very short and this method may be destructive. However with SD selection, long-tail durations of center still

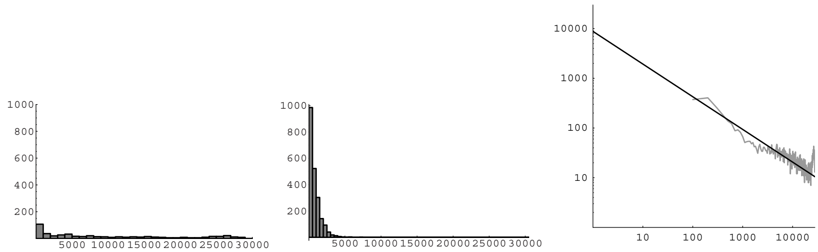


Fig. 8. The duration statistics of center agents with different selection methods and the power law of duration statistics for SD selection method. Parameters are the same as Fig.4. *Left* is SD selection, and *center* is PA selection. *Right* is a graph of power law. X-axis represents duration of the bins. Y-axis represents the amount of centers included in the bin. Parameters are the same as Fig.8. However, the bin is 100 turns and 3200 trials are accumulated here. *Solid line* represents αx^{γ} where $\gamma = 0.66$.

persist, even improve. Fig.8 shows a power law structure behind the durations distribution of SD selection method.

4 Conclusions and Discussion

We modeled the demand exchange system to investigate an emergence of a center agent. It shows the emergence and collapse of center agents as a monitor and an ordering power in the system. We encountered an efficiency problem and proposed methods to resolve the problem.

The assumption of simulating our model by real parallel processors is attractive in the view of Dennett's model on consciousness[9]. With Dennett, consciousness is a sequential computation emulated by the parallel computer like a brain. However why he thought so is not clear. In our model, before an emergence of a center agent, computing with parallel processors may have no problem. However after the emergence of a center, the amount of calls to the center by agents would become large, so if we try to keep the center's memory consistent (relations of microcosm in the center to agents' real status), we must introduce exclusive access control, i.e., partial sequential computation. If we use many copies of centers to avoid it, we must maintain consistency of these copies. Furthermore, if a center "thought" like a symbolic paradigm A.I with symbols in microcosm, then a certain part of the center must be sequential. So, situations like our model must have a sequential computation part. These situations may be found in several complex systems. We may also regard this "sequentiality" in the system as an emergence of some property that needs maintenance of consistency. In Dennett's sense, we may call it "consciousness". Probably, a very important problem of future A.I is connecting sequential symbolic operations with consistency and parallel computation like Brooks' machines[2]. Our research is a brief introduction to such investigations on systems with self-organization.

References

1. Johnson, S.: Emergence: The Connected Lives of Ants, Brains, Cities, and Software. Scribner (2001)
2. Brooks, R.: A robust layered control system for a mobile robot. Robotics. and. Automation, IEEE. Journal. of. 2(1), 14–23 (1986)
3. Cheon, T.: Evolutionary stability of ecological hierarchy. Physical. Review. Letters. 90(25), 258105 (2003)
4. Anghel, M., Toroczkai, Z., Bassler, K.E., Korniss, G.: Competition-driven network dynamics: Emergence of a scale-free leadership structure and collective efficiency. Physical. Review. Letters 92(5), 58701 (2004)
5. Ehresmann, A.C., Vanbremeersch, J.P.: Hierarchical evolutive systems: A mathematical model for complex systems. Bulletin. of. Mathematical. Biology 49(1), 13–50 (1987)
6. Brown, R., Paton, R.C., Porter, T.M.: Categorical language and hierarchical models for cell systems. In: Computation. in. Cells and Tissues: perspectives and tool of thought, Springer, Heidelberg (2004)
7. Yasutomi, A.: The emergence and collapse of money. Physica. D 82(1), 180–194 (1995)
8. Menger, C.: Principles of Economics. University Press, New York (1981)
9. Dennett, D.C.: Consciousness explained. Penguin (1993)

The Design of Model Checking Agent for SMS Management System

Ali Selamat[1] and Siti Dianah Abdul Bujang[2]

[1,2] Department of Software Engineering,
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia (UTM), 81310 Skudai, Johor, Malaysia
aselamat@utm.my

Abstract. Nowadays the short messaging service (SMS) has been used in direct sales and network marketing which is a part of e-business. Many systems integrate information system using mobile phone as a medium of communication in ensuring business processes become more effective and conducive. The intention of the system is to enhance the connection and networking in management system for direct sales activities. However, as the SMS services have been used extensively by the mobile users, there are several exceptional issues occurred during the runtime of each state of the e-business processes. Consequently, the correctness and secureness of the system performance have not been confidently trusted. Therefore, in this paper we propose a design of model checking agent for SMS management system to address the issue. Agents become a platform for model checking in handling the verification and validation processes in SMS management system. We have developed a model checking agent as a reliable tool for evaluating the failure of protocols in SMS management system. We have also analyzed the propose model which has contributed to the quality of SMS messages and increase the failure detection in e-business process.

Keywords: model checking agent, SMS management system, e-business, direct sales and network marketing.

1 Introduction

SMS management system is a concept of integrating mobile technology with information management system. Short messaging services (SMS) in mobile technology have been used as a tool to integrate the SMS management system for direct sales and network marketing. We have selected a case study on direct sales and network marketing sector as it is one of the most popular concepts in network marketing. The connection with distributors is an opportunity to the introduction of SMS as a fast and cost saving tools that enable the e-business activities. The SMS management system in direct sales and network marketing has used the SMS as a tool in improving the e-business processes. Direct sales and network marketing require a good relationship management within distributors [3]. They are encouraged to build and manage their sales force by recruiting,

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 813–821, 2008.

motivating, and training others to sell the product or services of the organization. A percentage based on the person sales force would be rewarded as their compensation, in addition to personal sales [4]. The SMS management system can create one to one communications within a multi-level market. The SMS services have been used effectively to improve market response that will add value to the distributor relationship management strategy and increase recognition of product marketing. Direct sales and network marketing might take the advantage of the SMS to use it in order to increase their sales and connecting people to people strategy. However, we are concerned with several issues that might be occurred towards the architecture of the SMS management system as the complexity of the database design is limited and unreliable to the distributed environment system. In fact, during the transition, the existing systems may not have unique identifications in controlling the security of the information. Therefore, in order to ensure the correctness of the data transfer within the systems in direct sales and network marketing, we propose a model checking agent (MCA) to verify the messages being sent and received within the SMS management system [10]. This paper is organized as follows: Section 2, the SMS technology in direct sales and network marketing and the concepts of model checking have been discussed. Section 3, the architecture of model checking agents and SMS marketing is discussed. The conclusion and summarization of the overall system are discussed in Sections 4 and 5, respectively.

2 Related Research Works

2.1 SMS Technology

The SMS technology have been created in Europe, by Global System for Mobile communications (GSM) pioneers [1]. The standardization process has been lead by the European Telecommunications Standards Institute (ETSI) [1]. The SMS technology has been created to provide an infrastructure for the transportation of short messages containing a maximum of 140 bytes of useful data in mobile telecommunication networks. The transportation is done in the GSM signalling path in traditional GSM networks, and as General Packet Radio Service (GPRS) packets in (GPRS) networks. The messages are composed using the protocol data unit (PDU) specification. An SMS is a binary string that contains the necessary information to form the message header needed for transportation and the message body that contains the payload. The basic addressing scheme of SMS messages based on the mobile telephone number that is Mobile Station Integrated Services Digital Network (MSISDN) [8].

2.2 Direct Sales and Network Marketing

Network marketing is a business in which a distributor network is needed to build the business. Usually such businesses are also multi level market in nature whereby the payout also occurs at more than one level. Network marketing is also used incorrectly to indicate that the business uses a network of product suppliers in order to offer a broader selection of products. It is usually used to differentiate the services as a way to suggest that their program is superior to other similar programs [4]. Multilevel marketing, or MLM, is a system used to sell goods or services through a network of distributors [2]. Direct sales and network marketing adopt the concept of MLM. The typical MLM program works through recruitment [2]. A person is invited to become a distributor, through another distributor of the MLM company's products and through an advertised meeting [3]. Distributor will earn money both through the sales of the MLM's products and through recruiting other distributors by receiving a portion of the income generate by these distributors [4].

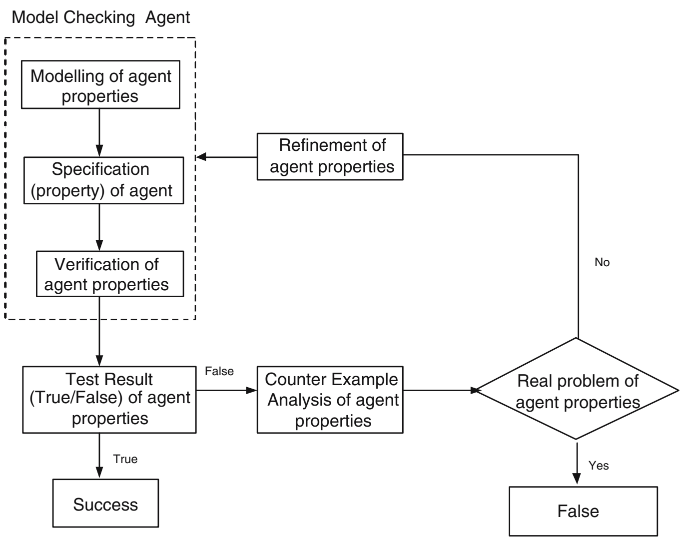


Fig. 1. The framework of model checking agent (MCA) processes in the SMS management system

2.3 Agent Formulation for SMS Marketing

Model checking is a method used to prove a digital system satisfies a user-defined specification. Both the system and the specification must be formally specified. The model of the systems must have a finite number of states and the specification or property, that is often expressed in temporal logics. The model and the property of the agents are represented by the Kripke structure and a temporal logic formula [6]. Formal model of the given digital system is known as Kripke structure. It is an annotated finite-state transition graph and is defined in 5 tuples as follows:

$$K = (S, S_0, T, A, \wedge) \quad (1)$$

where, S is a finite number of states;
$S_0 \subseteq S \times S$ is the set of initial states;
$T \subseteq S \times S$ is the transition relation;
A is a finite alphabet for which a set P of atomic propositions is given and $A = 2_p$,
$\wedge : S \rightarrow A$ is the labelling function for the states.

There are three main steps that we have used to create a model checking agent (MCA) activities as shown in Fig. 1:

Step 1: Modeling Agent

In SMS management system, there are four main activities of agents in the business process. The first step is a registration process whereby a customer will register as a distributor. The second activity is the purchase of a product. The third activity is a bonus calculation which means the profit of calculation percentage in sales force on MLM for the MLM agents. The final activity is a process of getting an information on activities and seminar. For example, the SMS management system has involved four actors that will be coordinating the processes; Distributor Agent(DA), Model Checking Agent (MCA), Order Entry Agent (OEA) and Product Server Agent(PSA)[5].

i)Distributor Agent (DA)

Distributor agent will request products for purchase through the Short Messaging Services (SMS). The distributor agent can purchase products, recruit new distributors, view the current bonus and retrieve current activities organized by the MLM company.

ii)Model Checking Agent (MCA)

The model checking agent is implemented to automate the sales process where the sales agent will be able to forward and deliver distributor requests of information on the purchase or bonus. The model checking agent gives a response to the product status and verify the distributor authenticity using model checking algorithm. This process is done in order to ensure the correct information is delivered to the distributor agent. If the payment is rejected, the model checking agent will send an alert message 'process failure' to the distributor agent and the transaction automatically be canceled.

iii)Order Entry Agent (OEA)

An order entry agent will coordinate with the purchasing agent in order to ensure the orders from the sales agent could be fulfilled in a given time frame.

iv)Product Server Agent (PSA)

The product server agent will organize the contents of the products according to their types in a SMS database.

Step 2: Specification

Specification requirement;
IF NOT transaction trace = STOP
OR {

Distributor agent (DA) sends SMS to model checking agent (MCA);
Order entry agent (OEA) receives SMS from model checking agent (MCA);
STOP;

}
OR {

Distributor agent sends (DA) SMS to model checking agent (MCA);
Order entry agent (OEA) receives aborted message from model checking agent (MCA);
STOP;

}
THEN
Requirement is violated;
End requirement;

Step 3: Verification
During the verification stage, the proposed model checking agent will use the SPIN model checker and PROMELA (PROcess MEta LAnguage)[5] in order to ensure the effectiveness of the agent validations. The SPIN model checker will perform a depth-first search on the state space and only considers all possible transitions that lead to the states that have not been visited. The SPIN model checker can detect such a livelock if the reservation and cancelation of the transitions have failed [5]. After the verification process, the model checking agent will test the results whether the system is in the correct condition. If an error is being identified, a counter-example agent will generate the errors report according to circumstances happened in the system. The model checking agent will provide identification if the model is failed and needs to be modified in order to overcome the uncertainties. But if the error is not found, the model checking agent may refine the model and make it reliable to the specification. The verification process will continue until all states are completely checked.

3 The Architecture of SMS Management System

As described in Section 2, there are four main activities in SMS management system as shown in Fig. 2. The proposed architecture is based on consideration of verification and validation of model checking agent in SMS management system for purchasing products. The architecture consists of a distributor agent, model checking agent, order entry agent and account management agent that communicates for verification and validation in handling the SMS management system. Further description on each of the components are described as follows:

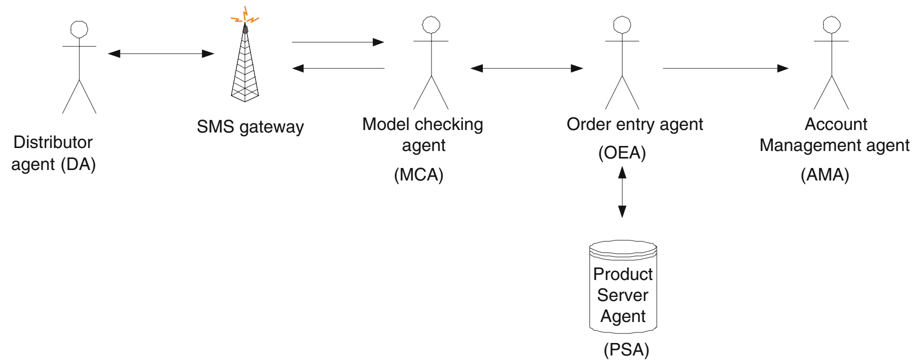


Fig. 2. The architecture of model checking agent (MCA) interactions in SMS management system

3.1 Distributor Agent (DA)

A user interface in mobile phone has used as a medium for a distributor to interact with distributor agent to the SMS management system. The distributor agent(DA) will receive the input of the SMS with the following format in order to purchase the items as shown in Fig. 3.

PUR#<ItemID>*<Quantity>#<Bank in Reference No>

Fig. 3. The syntax of SMS management system to purchase item

3.2 Model Checking Agent (MCA)

Once the SMS is received via SMS gateway, the distributor agent (DA) will send the SMS to model checking agent (MCA) for verification and validation of the format using a SPIN model checker. If the model checking agent (MCA) detects the activity has an error in an exceptional syntax or protocol, the SMS will be aborted. The MCA will send acknowledgement to the DA for notification. If the SMS meets the syntax requirements and protocols then the SMS will be sent to order entry agent (OEA) for order processing. At the same time, the DA will receive the acknowledgement immediately direct from MCA when the order has successfully been delivered.

3.3 Order Entry Agent (OEA)

Order Entry Agent (OEA) will manage the purchase order once the SMS has successfully been verified by the Model Checking Agent (MCA). The OEA will place the items ordered by the Account Management Agent (AMA) for purchasing

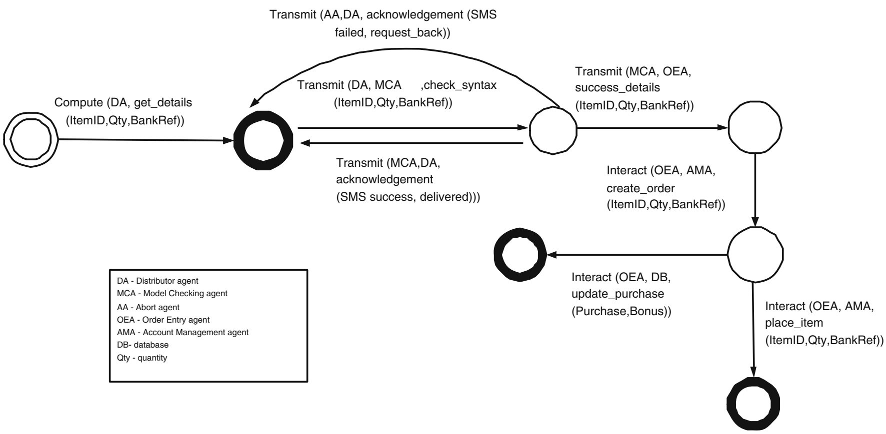


Fig. 4. The interactions of multi-agents in the SMS Management System

process to take place. The item purchased by the distributor agent (DA) will be given a bonus increment on each item sold by the sales agents.

3.4 Account Management Agent (AMA)

Account Management Agent (AMA)will receive the details of the requirements for purchasing the item from the OEA. The purchasing through SMS management system will be handled by the account management agent.

In particular, when many processes are distributed and automated, there are many states within these processes that needed to be evaluated. The reliability of the information transmitted among agents with the correct syntaxes and protocols using the SMS needs to be checked efficiently. Though, the SPIN model checker provides an effective method to verify whether the processes in each state transition of the SMS management system has no error, bugs and deadlock during the execution of the processes [9]. The failure of the SMS system can be detected at the earlier phase of development.

The architecture of the multi-agent and model checking algorithms used to check and verify the model of the SMS system is shown in Figs. 3 and 4, respectively. Based on the interactions of model checking agents in Fig. 2, we have further verified the agents interaction as shown in fig. 4. The Distributor Agent (DA) will send the SMS message with the details requirement that contains an item identifier (IT), quantity of items to be purchased (Q) and a bank reference number (R) in order to purchase the order to the Model Checking Agent (MCA). The message will be verified and validated by MCA in terms of protocol and syntax of the input message as shown in Fig. 3. The message received from the DA with the specified syntax can be accepted by the SMS management system. If the MCA detects the message is invalid, then it will send an acknowledgement to the DA to resend the message in a correct format. Otherwise the message is void. If the message then is valid, then the MCA will transmit the details of the

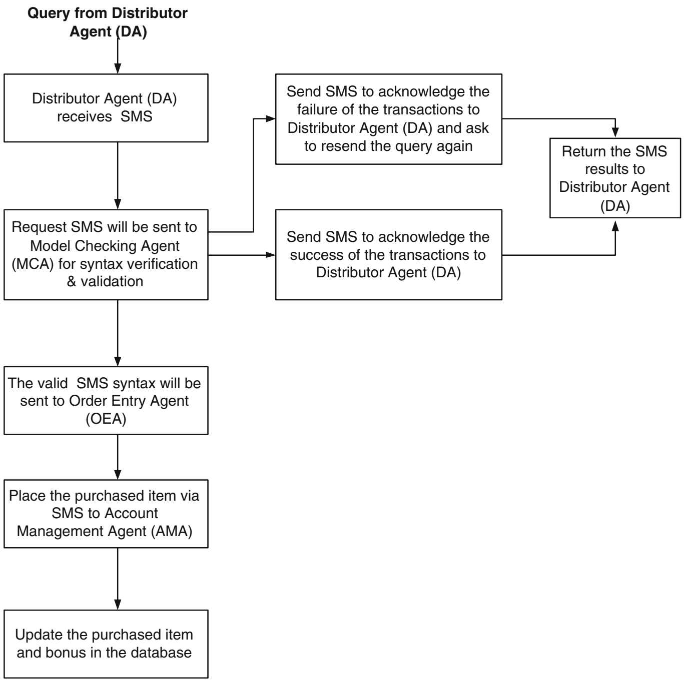


Fig. 5. The workflow of multi-agent interactions during product purchasing activities using SMS management system

message to the Order Entry Agent (OEA) for creating order and place the item with Account Management Agent (AMA). At the same time the MCA will send the acknowledgement to the DA that the message is successfully delivered. The database will then be updated after the OEA has completed the process.

The MCA plays an important role in ensuring the data transaction is manageable and releasable during the delivery of the messages. At the early state, the SMS message has to be verified by the MCA in order to ensure the requirement is satisfied the specification given by the system. If the specification of the particular state has not been satisfied, the MCA will generate a counterexample of an error. It decides whether the system need to do a refinement or an automatic violated message will be sent to the distributor. In return, the SMS system has to be modified in order to resolved the problem. The SMS system has used the SPIN model checker in order to evaluate the validity of the data and verify the correctness of the data given by the distributors and merchants. This is to ensure that the SMS management system is non-tolerable to errors as shown in Fig. 5.

4 Conclusions

The model checking agent is presented in this paper is mainly focusing on the problems in SMS management system for multi-level marketing. We have shown that the model checking agent architecture for the SMS management system. We have also validated the messages sent to the agents using SPIN model checking tools. As a result, it is a promisable area to be explored in verifying the messages sent by various agents in dealing with the purchasing of items in e-business activities.

Acknowledgement

This work is supported by the Ministry of Science & Technology and Innovation (MOSTI), Malaysia and Research Management Center, Universiti Teknologi Malaysia (UTM) under the Vot 78234 & 79200.

References

1. European Telecommunications Standards Institute (2004) http://www.etsi.org
2. Multilevel Marketing International Association (2007), http://www.mlmia.com
3. Direct Sales Association of Malaysia (1978), http://www.dsam.org.my
4. Stone, B.: Successful direct marketing method, Chairman, Stone & Adler, Inc. (1988)
5. Wang, W., Hidvegi, Z., Bailey, A., Whinston, A.: Model checking - A rigorous and efficient tool for e-commerce internal control and assurance, in Working Paper. Atlanta, GA: Gozuita School Bus., Emory Univ., pp. 5–10 (2001)
6. Wang, C., Hachtel, G.D., Somenzi, F.: Abstraction refinement for large scale model checking, Springer.com (2006)
7. Schuldt, H., Popovici, H.,, A.: Execution guarantees in electronic commerce payments. In: Proc. 8th Int. Workshop Foundations Models Languages Data Objects, New York. LNCS, pp. 189–198 (1999)
8. Mee, V.N., Selamat, A.: SMS management system for direct sales and network marketing. In: Proceedings of CITA 2007, Fifth International Conference on Information Technology in Asia, Malaysia (2007)
9. Bonnie Brinton Anderson, B.B., Hansen, J.V., Lowry, P.B., Summers, S.L.: Model checking for E-Business control and assurance. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews 35 (2005)
10. Selamat, A., Selamat, M.H.: Analysis on the performance of mobile agents for query retrieval from internet. Int. Information Sciences 172, 281–307 (2005)
11. Weiss, G.: Multi Agent System - A modern approach to distributed artificial intelligence. The MIT Press, Cambridge, Massachusetts London, England (2000)
12. Sciascio, E.D., Donini, F.M., Mongiello, M., Piscitelli, G.: Web applications design and maintenance using symbolic model checking. In: Proceedings of the Seventh European Conference on Software Maintenance and Reeingineering (CSMR 2003), IEEE, Los Alamitos (2003)
13. Selamat, A., Bujang, S.D.: Model checking agent for e-commerce system. In: The 3rd Malaysian Software Engineering Conference 2007, Kuala Lumpur, Malaysia, pp. 147–151 (2007)

An Intelligent Agent-Based System for Multilingual Financial News Digest

James N.K. Liu and M.K. Ho

Department of Computing, Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
csnkliu@comp.polyu.edu.hk

Abstract. Nowadays, online financial news from different sources is widely available on the Internet and there exists systems to help investors extracting and analyzing the financial news. However, many of these systems present news articles in a way without categorization and do not provide enough query options to search or get the specific aspect of news that they want. In this paper, we extend our previous work to develop an intelligent agent-based system for multi-lingual news extraction. We adopt a document categorization approach based on fuzzy keyword classification. The system applies fuzzy clustering to obtain a classification of keywords by concepts of the category. A category profile is developed and worked as a searching interface for document browsing. Experimental results show that the proposed Categorize News Agent is capable of categorizing news documents with a reasonable rate of accuracy.

Keywords: Intelligent agents, multi-lingual news extraction, document categorization, fuzzy classification.

1 Introduction

Nowadays, the World Wide Web becomes a major channel for information delivery. Online financial news from different sources is widely available on the Internet. The web provides real time and updated financial news whereas traditional newspapers are updated once or twice a day. It is difficult for investors to analyze such dynamic and large amount of information. This leads to a necessity for some intelligent multilingual system that is able to extract and analyze large amount of financial news written in different languages such as Chinese and English for our regional needs. Although similar systems (e.g. [3], [4]) have been developed recently, there are still some problems to be concerned.

The major concern of such system to users is how to get the news information they are interested in. Finance covers a huge area in which the users may want to search within a specific domain of news such as those in the stock market. Most existing financial news digest systems present news articles in a way without categorization and do not provide enough query options to search or get the specific aspect of news that they want. For example, if the user wants to analyze the recent stock market situation and types in the query “Stock”, the result list will probably show the news articles that contain the keyword “Stock” and miss many news articles that are relevant to the stock market.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 822–831, 2008.

The main focus of the paper is to present an intelligent system that can extract financial news from online English and Chinese financial newspaper websites and provide user-friendly web interfaces for users to retrieve and query information through Internet browser. As this paper is an extension of our early paper [4] which gave detailed discussion on many basic functions such as fetching news articles from the web, segmenting Chinese sentences, etc, we would like to concentrate the study on news categorization. Section 2 explores the means of categorizing financial news. Section 3 presents an overview of the agent-based system. Section 4 describes the experiments and results. Conclusions and future work will follow in the last section.

2 Related Work in News Categorization

News categorization aims to automatically classify news articles into a fixed number of predefined categories. Using machine learning techniques, we build classifiers to learn from examples and perform category assignments automatically. A collection of news articles are first downloaded as training set for the process of keyword classification. The keywords in the training set are then classified into clusters with corresponding weight and stored as a category profile. Any incoming uncategorized news articles will then be classified into categories by calculating the similarity between the clusters and news articles. The system flow of news categorization is shown in Fig. 1.

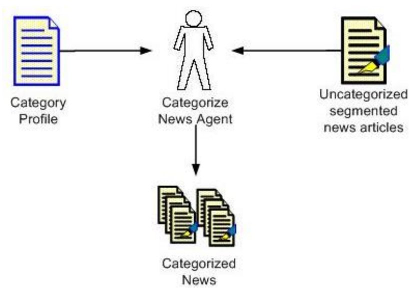


Fig. 1. The workflow of news categorization

2.1 Fuzzy Keyword Classification

This process aims to generate category profile that contains the keywords to be classified and its membership values in that category. It analyzed the occurrences of the keywords in different training documents and calculated the membership value of the keywords in each cluster. A fuzzy c-means algorithm (FCM) [1] is applied to sort a set of keywords collected from a corpus of news articles into clusters. A collection of news articles would be used as the training data for generating the category profile.

Fuzzy c-means algorithm (FCM) is applied to calculate the membership values of each keyword in each cluster and therefore determine the semantic meaning of keyword to each category. The steps are outlined as follows:

Step 1. Construct a keyword-document matrix

The keyword-document matrix, KD, is used to represent K keywords extracted from D training documents. The weight W_{KD} of the K keywords in D documents is the keyword frequency in the document.

$$KD = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1D} \\ W_{21} & W_{22} & \cdots & W_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ W_{K1} & \mu_{K2} & \cdots & W_{KD} \end{bmatrix}$$

Step 2. Initialize membership values of each keyword

Initialize the membership values μ_{ik} of k keywords t_k to each of the cluster v_i randomly such that

$$\sum_{i=1}^{c} \mu_{ik} = 1 \qquad \forall k = 1,\dots,K \text{ and } \mu_{ik} \in [0,1] \qquad \forall i = 1,..c; \forall k = 1,...K$$

where c is the number of clusters (categories) and K is the number of keywords extracted from the document collection.

Step 3. Calculate cluster centers

The cluster centers v_i are calculated using the membership values μ_{ik}:

$$v_i = \frac{\sum_{k=1}^{K} (\mu_{ik})^m \cdot t_k}{\sum_{k=1}^{K} (\mu_{ik})^m} \qquad \forall i = 1,\dots,c$$

where m is the fuzziness index. Usually, m = 2 is chosen [3].

Step 4. Calculate new membership values using cluster centers

The new membership value μ_{ik}^{new} of each keyword is then calculated using the cluster centers v_i:

$$\mu_{ik}^{new} = \frac{1}{\sum_{j=1}^{c} (\frac{\|v_i - t_k\|}{\|v_j - t_k\|})^{\frac{2}{m-1}}} \qquad \forall i = 1,\dots,c; \forall k = 1,\dots,K$$

Step 5. Repeat Step 2 to 5 until new membership values μ_{ik}^{new} is less than convergence threshold ε (small positive number):

If $\left\|\mu^{new} - \mu\right\| < \varepsilon$, let $\mu = \mu^{new}$ and go to Step 2.

Otherwise, stop.

We have set the convergence threshold ε to 0.001. After completing the calculation of membership values, each keyword is mapped to the found clusters with different membership values and all existing category profiles are represented by a matrix S as:

$$S = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1k} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{c1} & \mu_{c2} & \cdots & \mu_{ck} \end{bmatrix}$$

After applying FCM to classify keywords in the training articles into clusters, a category profile is generated. The core keywords of a cluster are assigned relatively high membership values indicating their significance in this category. For example, in Property market category, the keyword "property 物業' with membership value 0.9025 which has a relatively high membership value in this category and indicates that this keyword is important and significant in the category of property market. The keyword with broader scope of meaning such as "investment 投資'will be assigned moderate membership values across different clusters as this keyword occurs in most of the news articles and it is somewhat related to each category. The sample of category profile is shown in Table 1.

Table 1. Sample of category profile

Keyword	D1	D2	D3	D4	D5	D6	D7	D8	D9	Stock Market	Prop. Market	Finance Market
Property 物業	0	0	9	1	6	2	0	0	0	0.0276	0.9025	0.0699
Investment 投資	15	2	0	0	0	16	0	1	10	0.3280	0.2529	0.4191
Property price 樓價	0	0	10	1	4	0	0	2	0	0.0249	0.9126	0.0625
Gold price 金價	5	3	0	0	0	5	3	4	3	0.1217	0.0983	0.7800
Exchange rate 匯價	2	0	0	0	0	1	0	2	4	0.0664	0.1375	0.7962
Stock price 股價	1	0	0	0	0	6	15	2	5	0.8994	0.0387	0.0619

2.2 Document Classification

In this process, a document classification model based on vector space model [8] is applied to classify uncategorized news documents into categories by calculating the similarity between news documents and clusters. The similarity is calculated based on the cosine similarity function (see Fig. 2).

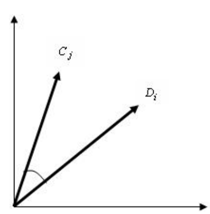


Fig. 2. Cosine of angle between two vectors

Similarity between news document and clusters

Each cluster C_j, resulting from the previous fuzzy keyword classification process, is represented by a vector of weighted keywords as follows:

$$C_j = \{W_{j1}, W_{j2}, ..., W_{jt}\}$$

where W_{jt} represents the weight of keyword T_t being assigned to cluster C_j.

Each news document D_i is represented by:

$$D_i = \{W_{i1}, W_{i2}, ..., W_{it}\}$$

where W_{it} represents the weight of keyword T_t in news document D_i.

The weight W_{it} is determined by using Salton's term weighting formula [8]:

$$W_{it} = \frac{tf_{it} \bullet \log\left(\frac{N}{n_t}\right)}{\sqrt{\sum (tf_{it})^2 \bullet \left(\log\left(\frac{N}{n_t}\right)\right)^2}}$$

where:

tf_{it} is the frequency of occurrence of keywords T_t in news document D_i.

N is the size of the document collection.

n_t is the number of documents in the collection with keyword T_t.

$\frac{1}{\sqrt{\sum (tf_{it})^2 \bullet \left(\log\left(\frac{N}{n_t}\right)\right)^2}}$ is used for content normalization to avoid long documents with a broad content coverage to be assigned with a high similarity degree to clusters involved. This ensures that only documents specifically relevant to a cluster are ranked first.

With reference to the cosine similarity function [9], the similarity between document and cluster is the cosine of the angle between the cluster vector and the document vector.

$$sim(C_j, D_i) = \frac{\sum_{k=1}^{t} W_{jk} \bullet W_{ik}}{\sqrt{\sum_{k=1}^{t} {W_{jk}}^2 \bullet \sum_{k=1}^{t} {W_{ik}}^2}}$$

These similarity values will be between 0 and 1, and depend on the weight of matching keywords in the document vector. The larger the similarity values of the document to the cluster, the higher the chance to be classified to that cluster.

Category assignment

To classify the news documents into categories, we applied a threshold strategy, Scut, which is commonly used in the field of document categorization [10]. Scut stands for score-based classification and it is suitable for classifying objects into categories using scores such as membership values, weights.

We have set a single global threshold θ for all the categories. If the similarity between the news document and the category (cluster) is larger than the threshold, the news document will be assigned to that category. Otherwise, that category is discarded for that news document. The document can be assigned to more than one category and it depends on the similarity values of the documents to each category. If all of the similarity values of the document to the 3 predefined categories are less than the threshold, that document will be assigned to the category of "Others 其他". Therefore, the category of "Others" contains the documents which do not belong to the category "Property market", "Finance market" and "Stock market". Table 2 shows the news documents and respective similarity value.

Table 2. News documents and its similarity with respect to each category

Document	**Similarity value with respect to category**		
	Property Market	Finance Market	Stock Market
D1	0.2502	0.3091	0.1204
D2	0.1687	0.3838	0.3326
D3	0.1501	0.3521	0.6578

Table 3. News document classification

Rank	**Property Market**	**Finance Market**	**Stock Market**
1	D1	D2	D3
2		D3	D2
3		D1	

Based on our model of document classification, the news documents are classified and their ranking with respect to each category and corresponding similarity values are summarized in Table 3. D2 and D3 are not assigned in the "Property Market" category and D1 is not assigned in the "Stock Market" category, as their similarities with respect to that category are lower than the preset threshold.

3 System Overview and Design

As an extension from [4] a new agent, Categorize News Agent, is added to perform news categorization. We mainly focus on the agents which are involved in news categorization processes. They are Categorize News Agent, Web Controller, News Agent, Chinese Word Segmentation Agent and Fetch News Agent, as shown in Fig. 3 below:

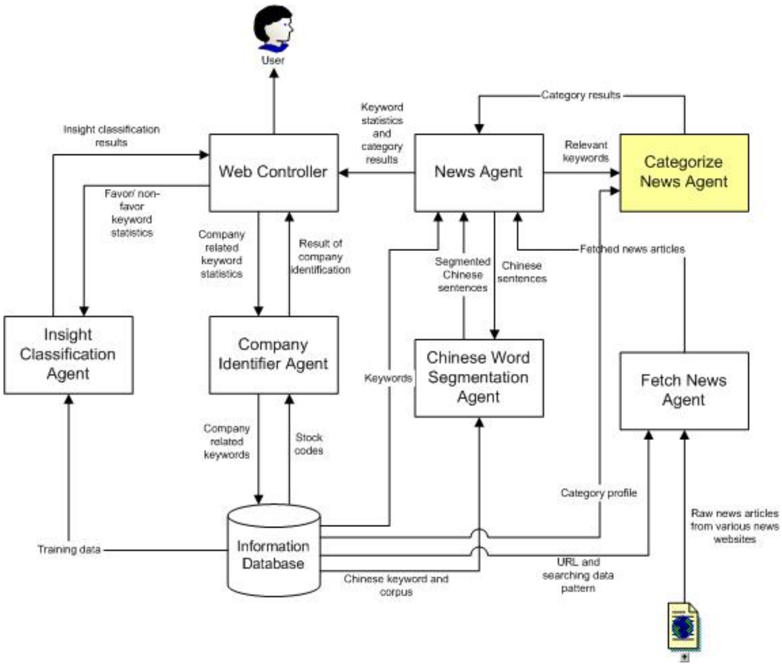


Fig. 3. Overview of the system

1. Web Controller receives the user request and send message to Fetch News Agent for fetching news.
2. Fetch News Agent downloads news articles from various newspaper websites and sends the fetched news articles to News Agent.
3. News Agent reads news contents and sends them to Chinese Word Segmentation Agent if the news content is written in Chinese.
4. Chinese Word Segmentation Agent segments the Chinese sentences into Chinese words and sends back to News Agent.
5. The relevant keywords are sent to Categorize News Agent through News Agent and Categorize News Agent classifies news articles into categories and sends back the results to News Agent.
6. News Agent sends the category results to Web Controller.
7. Web Controller forwards the response back to the user..

The news articles are defined according to category assignment of the new articles from most of the newspaper websites (see Table 4). In order to gather a set of training keywords for generating category profile, a set of online financial news, written in Chinese, was collected from the financial newspaper websites. 10 news articles were collected from each category and a total of 30 news articles were collected. The keywords are then extracted from these 30 articles by using the Chinese word segmentation method developed in [4].

We have applied intelligent agent using iJADK [11] to perform the tasks including fetching news articles from the Internet, segmenting Chinese sentences into Chinese

Table 4. Definition of category

Category	Definition
Property Market (物業市場)	Rental, housing selling and buying situation
Finance Market (金融市場)	Services and competition from banks and alternative financial institutions, the exchange rates of the bank.
Stock Market (股票市場)	Price movements of stocks
Others (其他)	Anything related to financial activities such as Hong Kong economy, the company current situation, etc.

words, etc. This saves time from client and saves the cost of client device, as all these jobs are done by designated agents. This greatly increases the feasibility and mobility of application.

4 Experimental Results and Discussion

The experimental tests are divided into 2 parts. In the first part, we conducted a set of experiments to choose the best value for the threshold. A total of 172 Chinese news articles published in March 2006 were downloaded from various newspaper websites for determining the threshold used in the system. In the second part, we evaluate the performance of Categorize News Agent in classifying news articles into the most relevant category they belong to. A total of 209 financial news articles written in Chinese are fetched from various newspaper websites in April 2006. To evaluate the performance of Categorize News Agent that classifies news articles into categories, three metrics [2] are employed and defined as follows:

$$\mathrm{Re}\,call = \frac{\text{No of news articles correctly assigned to category}}{\text{No of news articles belonging to category}}$$

$$\mathrm{Pr}\,ecision = \frac{\text{No of news articles correctly assigned to category}}{\text{No of news articles assigned to category}}$$

$$Accuracy = \frac{\text{Total no of news articles correctly categorized}}{\text{Total no of news articles in the collection}}$$

Before evaluating the performance of the Categorize News Agent, We tested the similarity function's threshold θ for classifying news articles into categories, and determined the best value of θ based on the average recall, precision and the overall accuracy of the results. Tables 5-6 show the distribution of news articles assigned to the actual and predicted categories and the accuracy in each category:

Table 5. Distribution of news articles assigned to the actual and predicted categories and its accuracy

Predicted Category (No of news)	**Actual Category (No of news)** Property Market (79)	Finance Market (58)	Stock Market (42)	Others (30)
Property Market (58)	57	0	0	1
Finance Market (51)	3	35	9	4
Stock Market (39)	2	8	25	4
Others (61)	17	15	8	21
Accuracy (%)	72	60	60	70

Table 6. Summarization of test result of Categorize News Agent

Category	Recall (%)	Precision (%)	Accuracy (%)
Property Market	71	98	72
Finance Market	63	69	60
Stock Market	60	64	60
Others	68	34	70
Overall	66	66	66

From the above tables and Fig. 4, we can summarize that:

- The category "Property Market" has the highest precision rate of 98%.
- The category "Others" has the lowest precision rate of 34%.
- The overall recall, precision and accuracy are 66 %.

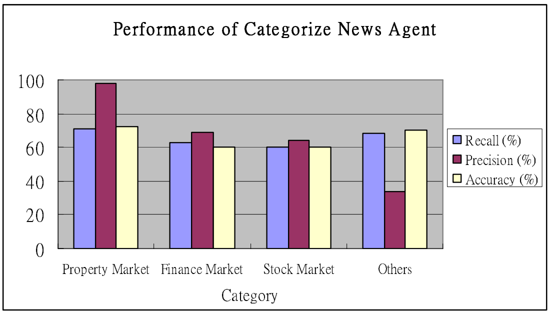


Fig. 4. Performance of Categorize News Agent

5 Conclusion and Future Work

In this paper, we focused on classifying the keywords into clusters with different weights using fuzzy clustering method and compare the similarity of the news documents and the keyword cluster. Our preliminary findings show the following:

1. High precision in the category of "Property Market"
   The precision rate of the category "Property Market" reaches 98% and the accuracy is 72%. We note that the category profile can cover most of the significant keywords of news articles in "Property Market"category where most of the news articles have the keywords "Unit rate 呎價', "Property value 樓價' which are the significant words in the "Property market" category and have less number of keywords which are significant in other categories.
2. Low precision in the category of "Others"
   The precision rate of the category "Others" is relatively low but it has a fair recall rate. The reason of low precision rate in this category is that the category profile cannot cover such keywords which are significant to "Finance Market" and "Stock Market". Therefore, the similarity to these two categories would be lower than the threshold and many documents which belong to other categories are assigned to "Others".

3. Fair overall recall, precision and accuracy of the Categorize News Agent
   The overall recall, precision and accuracy are 66% which implies there is still much room for improvement. The performance is not good enough as there is not enough training data set for training the category profile. The incorrectly categorized news articles are often assigned to the "Others" category. We have examined those news articles and discovered that keywords in those news articles cannot be found in the category profile indicating the need to get more data.

The results described show that the Categorize News Agent is capable of categorizing news documents with a reasonable rate of accuracy. Also, according to the analysis on the performance of Categorize News Agent, it is shown that the performance is somehow affected by the category profile which requires large set of news articles for training. To improve the matching result of keywords in Chinese news further, we need to analyze the semantic meaning of Chinese sentences. For example, we can apply the Chinese synonym collocation extraction method added on the keyword matching method. The information can be used to build the ontology structure of the domain categories of news [6]. On the other hand, we wish to incorporate the system with automatic summarization algorithm and distributed crawlers that can generate news abstraction automatically (e.g. [5], [7]). All these will be the subject of future research.

Acknowledgement

The authors would like to acknowledge the partial support of departmental grant 1-ZV41 of The Hong Kong Polytechnic University.

References

1. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithm. Plenum Press, New York (1981)
2. Junker, M., Hoch, R., Dengle, A.: On the Evaluation of Document Analysis Components by Recall, Precision and Accuracy. In: International Conference on Document Analysis and Recognition, pp. 713–716 (1999)
3. Lim, S.J., Ng, Y.K.: Categorizing and Extracting Information from Multilingual HTML Documents. In: International Database Engineering and Application Symposium, pp. 415–422 (2005)
4. Liu, J.N.K., Dai, H.H., Zhou, L.: Intelligent Financial News Digest System. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 112–120. Springer, Heidelberg (2005)
5. Liu, J.N.K., Luo, W.D., Chan, E.M.C.: Design and Implement a Web News Retrieval System. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 149–156. Springer, Heidelberg (2005)
6. Middleton, S.E., Shadbolt, N.R., Roure, D.C.: Ontological user profiling in recommender systems. ACM Trans. Information Systems 22(1), 54–88 (2004)
7. Mukherjee, S., Yang, G.Z., Tan, W.F., Ramakrishnan, I.V.: Automatic Discovery of Semantic Structures in HTML Documents. In: International Conference on Document Analysis and Recognition, pp. 245–249 (2003)
8. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)
9. Salton, G., Buckley, C.: Term-weighting approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
10. Yang, Y.M.: A study on thresholding strategies for text categorization. In: ACM SIGIR Conference on Research and Development in Information Retrieval (2001)
11. iJADK – Intelligent Java Agent Development Kit: `http://www.ijadk.com`

An Agent-Based Diabetic Patient Simulation

Sara Ghoreishi Nejad[1,2], Robert Martens[3], and Raman Paranjape[1,2]

[1] TRLabs Regina
[2] Electronic Systems Engineering
University of Regina
[3] SaskTel Research and Development Inc.
Regina, Canada
Raman.Paranjape@uregina.ca

Abstract. This paper presents a new paradigm for modeling illness in the human population. In this work we propose the development of a patient model using a Mobile Software Agent. We concentrate on Diabetes Mellitus because of the prevalence of this disease and the reality that many citizens must learn to manage their disease through some simple guidelines on their diet, exercise and medication. This form of modeling illness has the potential to forecast outcomes for diabetic patients depending on their lifestyle. We further believe that the Patient Agent could be an effective tool in assisting patients to understand their prognosis if they are not meticulous in controlling their blood sugar and insulin levels. Additionally simulation results may be used to exercise physiological data collection and presentation systems. The Patient Agent is developed in accordance with the general parameters used in archetypal Diabetes medical tests. Conventional formulae have been applied to transform input variables such as Food, Exercise, and Medications, as well as other risk factors like Age, Ethnicity, and Gender, into output variables such as Blood Glucose and Blood Pressure. The time evolution of the Patient Agent is represented through the outputs which deteriorate over the long term period.

Keywords: Agent-based modelling, simulation, patient agent, diabetes, health care system.

1 Introduction

The modeling of health care components and systems, in order to develop a complete understanding of component interactions, is one of the more challenging simulation and modeling problems for software agent systems [1]. Clearly, one of the most important components in health care is the patient. In this work, we present our initial development strategy and preliminary results on the creation of a diabetic patient agent model.

1.1 Overview of Diabetes

This model has the potential to present various phases of the illness, in which the severity may change with time. The patient can have specific symptoms depending on

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 832–841, 2008.

the nature of his/her condition. There are a number of explicit causes and factors that affect patients' convalescence and improvement period. We focus particularly on the Diabetes Mellitus illness and try to delineate this disease through the attributes of a Mobile Software Agent.

Diabetes is a disease in which the body is not capable of producing or properly using insulin. Insulin is a hormone required to convert sugar and other carbohydrates into necessary energy for the body. The main cause of Diabetes has not been discovered but genetics and environmental factors, such as obesity and lack of exercise, seem to influence the occurrence of the disease [2, 3].

Diabetes is common in North America. More than 2 million people suffer from Diabetes in Canada alone [2]. In the US, 7% of the population (20.8 million people) have diabetes. About 14.6 million people have been diagnosed, while 6.2 million people are uninformed of their disease [3].

There are three main types of diabetes: Type I, Type II, and Gestational Diabetes. Type I Diabetes occurs when the pancreas is not capable of producing insulin, and is typically more common among children and adolescents. In Type II, the pancreas cannot produce enough insulin or use insulin properly, which generally develops in adulthood. The majority of diabetic patients suffer from Type II Diabetes; therefore the study of this type will be the focus of our paper. Gestational Diabetes can temporarily occur during pregnancy [2, 3].

Some people are at greater risk of developing diabetes. For example, the probability of having Diabetes increases if a person is aged 40 or more, or is a member of a high risk group like Aboriginals or African-Americans [2, 3].

Diabetes, if treated inadequately or managed poorly, can result in serious complications such as heart disease and stroke, kidney disease, eye complications (which may lead to blindness), neuropathy and nerve damages, foot problems, or skin complications [3]. However, Diabetes can be managed if one pays attention to a number of critical items, such as food intake, exercise, and medicine consumption [2]. Regularly scheduled medical tests, such as blood glucose tests and blood pressure measurements, are also critical for management of Diabetes.

1.2 Mobile Software Agents

Mobile Agents are autonomous software programs that may start running on a host computer, stop what they are doing, move to another host computer and start up from where they left off. Mobile Agent programs run with the aid of another program called an Agent Platform that must be installed and be running on a host computer before the mobile agent can run. The Agent Platform provides the mobile agent with services for mobility, messaging, resource access, and security. Mobile Agents have four key attributes: Autonomy, Re/Pro-Activity, Mobility, and Group Action or Derived Behaviour.

In this paper, the agent technology is based on the TEEMA platform. TEEMA stands for TRLabs Execution Environment for Mobile Agents [4, 5]. This platform has been developed jointly by TRLabs Regina and the University of Regina, and it is a relatively standard Agent Execution Environment (AEE) or agent platform. TEEMA provides addressing and naming, messaging, mobility, security, logging, and a procedure for the addition of services, like many AEEs. TEEMA and the Patient Agent in this work are programmed with Java. While TEEMA is used in this work because of its familiarity to the authors, no unique or special features are required of the AEE, and any other AEE with these same basic functions could be used in this work.

1.3 Agent Simulation

Agent systems can be used to model or simulate complex interacting systems. They can be more effective than conventional simulation tools because individual components can be mapped into individual software agents taking advantage of the concept of agent autonomy to completely describe the component.

We believe that this Patient Agent can further be employed as a component in a larger model of the Canadian Health Care system. The Patient Agent can cooperate with other mobile software agents which represent other important components in the Canadian Health Care system. These components interact with each other and simulate people's actions and behaviours. Thus, an agent system model can, in principle, be used to study macroscopic system behaviour of the Canadian Health Care system.

1.4 Physiological Data Systems

Physiological data collection and presentation systems are commonly used in telemedicine and other areas of health care. We believe that these systems can benefit from being exercised by simulation data. For example glucose measurements may be submitted into a physiological data collection and presentation system, designed for Diabetes patients, to exercise the presentation capabilities of the system.

2 Methods

The structure of the Patient Agent through its inputs and outputs is illustrated in Figure 1.

In the design of our Patient Agent, the critical items like Food, Exercise and Medications (Insulin intakes) are considered as Input Variables, as well as a number of risk factors like Age, Gender, and Ethnicity. The Outputs are the major medical test results for diabetes.

A simple example is to examine the effect of changes in the Inputs and observe the effects on the Outputs; such as modifying the amount of patient's food intake and observing the effects on his/her blood glucose level; or including an exercise program in his/her lifestyle and noticing the results.

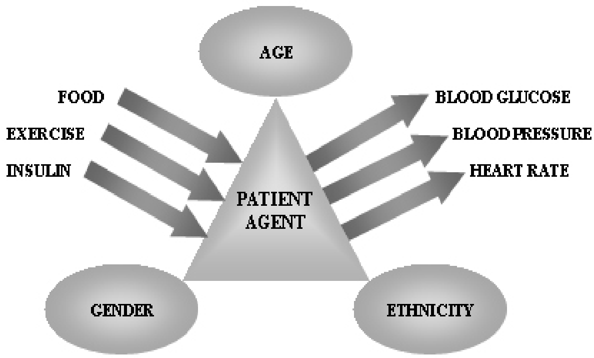


Fig. 1. Generalized diagram of Inputs and Outputs for the Patient Agent

For a typical diabetic patient, there can be two types of Input Variables:

The first type of variables (considered as Risk Factors for diabetes) are Age, Ethnicity, and Gender. Depending on their value, the patient is more or less likely to be influenced by diabetes [2]. People in different age groups have different risks of developing diabetes. People aged 40 or over are more prone to be diabetic. In addition, diabetes is more common in some ethnic groups, such as Aboriginals, or African-Americans. In addition, typically more men are affected by diabetes than women [6].

The other type of variables will directly have impacts on the medical test results. Examples of this type of variable are Food (particularly Carbohydrates), Medications (Insulin intakes), and Exercise.

A number of models exist for diabetes [7, 8] and in order to transform the Input parameters into Outputs, using agent simulation, the following equation has been applied [9]:

$$x(t) = \frac{F}{\omega} e^{\frac{-\beta t}{2}} \sin \omega t \tag{1}$$

In this equation, the Output Variable x(t) represents the blood glucose level, F is the food intake, β shows the effects of exercise and medications, and ω is the frequency of the system [9]. In order to have a measure of the Food, its Glycemic Index is calculated. Glycemic Index is a numerical index that classifies carbohydrates depending on their rate of glycemic response, or their conversion to glucose within the human body [10]. A moderate daily exercise program is included in our Patient Agent model.

The model variables have been shown to be independent of each other [9] and we propose that the relationship may be derived from future comparisons of the model to existing physiological data. In other words F is not simply Glycemic Load but is instead some combination of GL together with functional relationships to ω and β.

The Output Variable Blood Pressure is another important variable to be monitored.

In this work we obtained a typical blood pressure trace for normal and diabetic patient. These traces can be modified based on the patient agents' life style and activities [11, 12].

The Output Variable Heart Rate is modeled an empirical form similar to the Output Variable Blood Pressure [13, 14].

In this work we present results over the short term and the long term. The short term results are based on the effects of the Input Variables on the Output Variables using the above-mentioned formula. The long term results for the Output Variables represent the typical deterioration of these variables with the age of the patient.

3 Results

We now illustrate the typical Input and Output curves for a normal and diabetic patient. The input and output curves are archetypal.

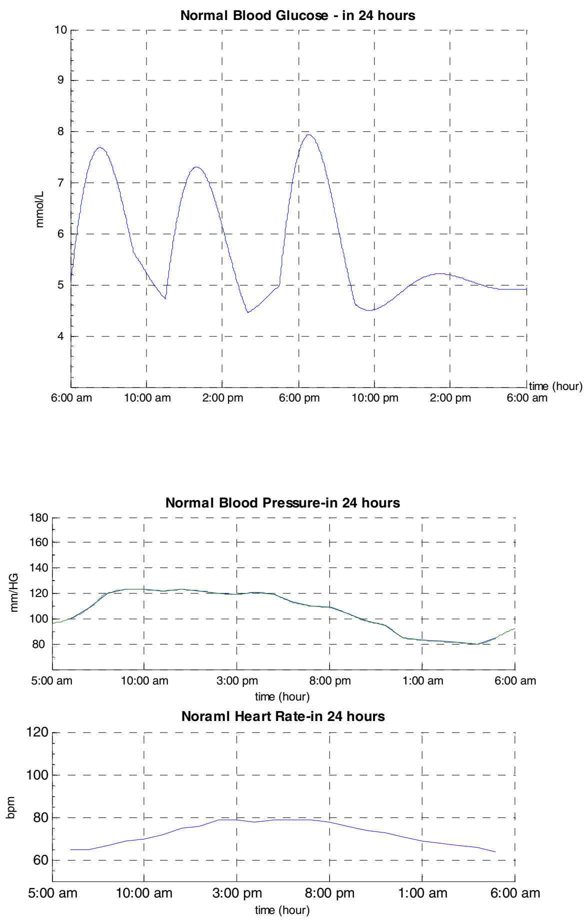


Fig. 2. Output parameters of Normal Patient over Short Term

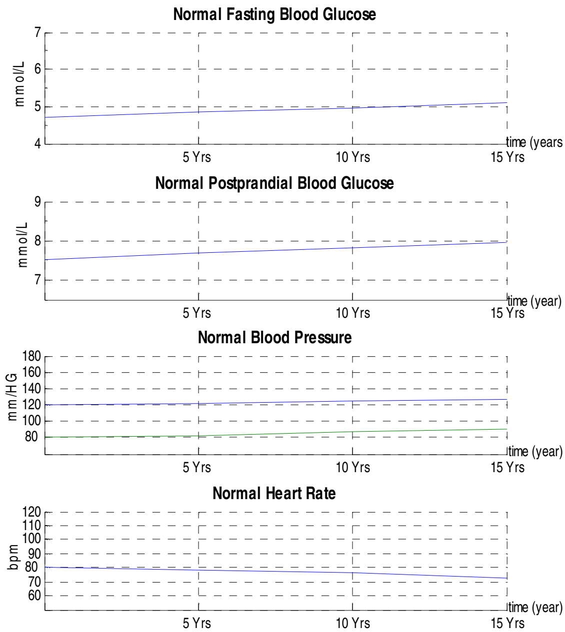


Fig. 3. Output Parameters of Normal Patient over Long Term

4 Discussion

In the Results section, some typical input and output results for the Patient Agent have been presented. The configuration of the input curves is specified by the system designer and may be tailored to different types of patients using an input form as shown in Figure 6. The output curves on the other hand are either generated by a functional relationship between the Input Variables and the Output Variables, or are based on empirical data typical for normal and diabetic patients.

The time evolution of the Postprandial blood glucose shown in Figure 4 is limited to a 24 hour period after consumption of the patient's breakfast, and is based on the Input Variables. This is achieved through a summation of the applied equation's outputs over the time period and adjusted by a baseline value. Exercise levels are assumed to be a regulatory variable rather than a time dependent variable, as shown in the input curves.

Blood pressure and heart rate curves are empirical measured data typical of normal and diabetic patients, and are also not tightly coupled to the activity in the other Input Variables.

As the results over the long term period demonstrate, there is a substantial change in the blood glucose, blood pressure and heart rate levels for the diabetic patient. However, there is not a big change in such levels for a normal person.

As observed in the short term diagrams, Preprandial blood glucose level, or the blood glucose level before eating, is higher for the diabetic patient. Similarly, Postprandial blood glucose level, or the blood glucose after eating, is higher for the diabetic patient. Also, the Recovery Period, which is the time period after eating in which the blood glucose rises, reaches a peak, and then dips to the levels of before eating, is longer for a diabetic patient.

Furthermore, the differences between the levels of blood pressure and heart rate in a diabetic patient and a normal person are noticeable.

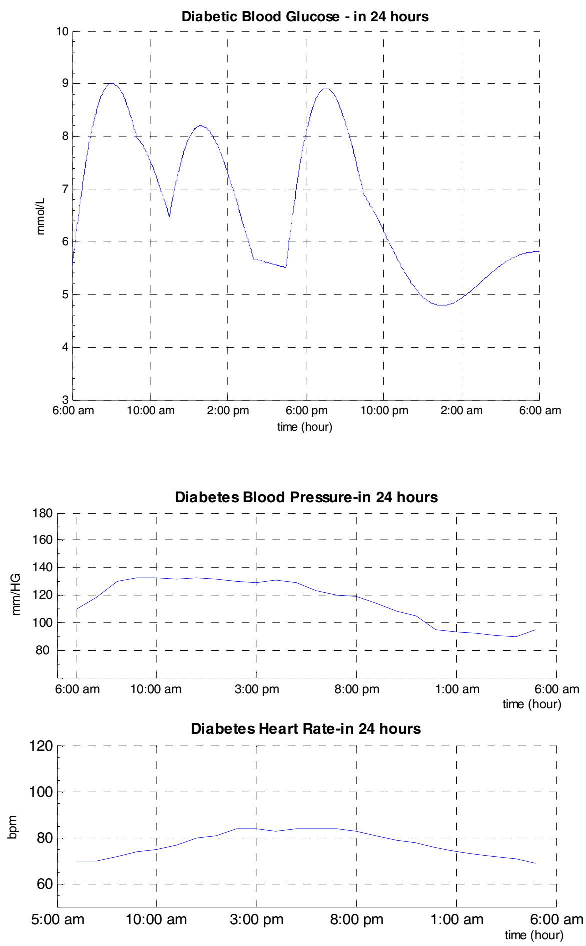


Fig. 4. Output Parameters of Diabetic Patient over short Term

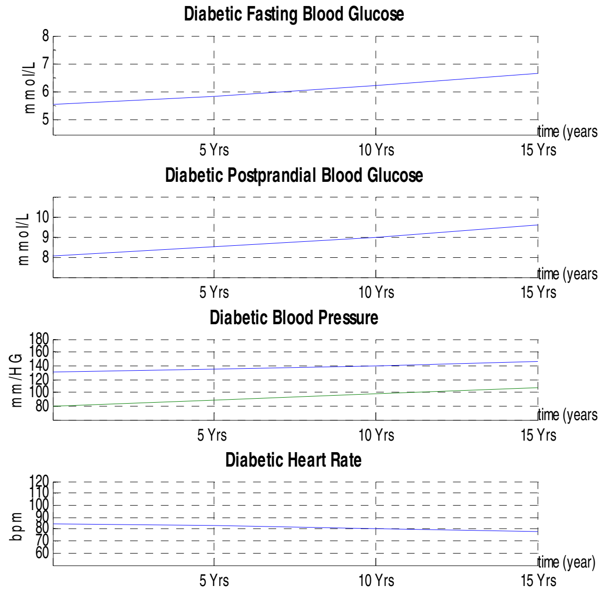


Fig. 5. Output Parameters of Diabetic Patient over Long Term

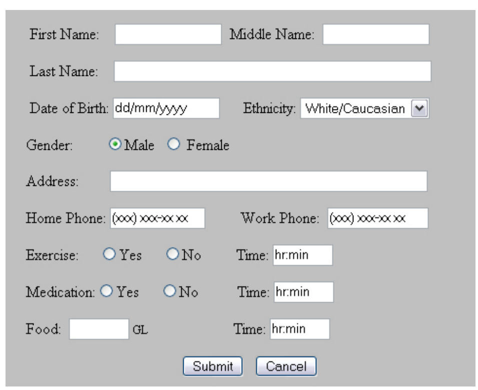


Fig. 6. Input Form to Gather Patient Input Variables

5 Conclusion

We have proposed a novel and dynamic model of illness using Mobile Software Agents. In this model, we specifically concentrated on Diabetes Mellitus. The major influential parameters on diabetes have been considered as Input Variables, while medical tests results are used as the Output Variables. The continuous output signals generated in this model can have profound impact on predicting outcomes for patients with diabetes, based on their behaviour and lifestyle. This model can help patients explore their prognosis if they are not meticulous in controlling their blood sugar and insulin levels. Additionally this model can exercise physiological data collection and presentation systems.

6 Future Work

The model of the Patient Agent can be used by patients to investigate the effects of their activities and life styles on the critical parameters which measure their medical condition. The Patient Agent Model has to be further developed in order to more fully connect the model variables to the input variables. Specifically, model variables are dependant on each other [9] therefore relationships between input variables and model variables need to be explored.

Interfaces between the Patient Agent Model and existing physiological data systems may be developed as a potential commercialisation of this work.

The patient agent model can also be used as a component in a complex model of the Canadian Health Care System. The system model would be populated with a statistically distribution of patient agents with similar but not identical curves. The distributions of the patient agent population would be taken for the actual population of patients visiting the health care system. By sampling various inputs and outputs, Patient Agents can simulate different illnesses. The Physicians, Nurses, and other specialized agents will also have to be developed. This configuration of agents can interact with the patient agents. The specific manner in which they interact will characterize a particular scheduling system. These components interact with each other and engage in a manor analogous the operation of the health care system.

References

1. Moreno, A., Isern, D., Sanchez, D.: Provision of Agent-Based Healthcare Services. In: Quaglini, S., Barahona, P., Andreassen, S. (eds.) AIME 2001. LNCS (LNAI), vol. 2101, Springer, Heidelberg (2001)
2. Canadian Diabetes Association website, As seen on December 12 (2006)
3. American Diabetes Association website, As seen on December 12 (2006)
4. Liu, Q.A.: Master Degree Thesis: A Mobile Agent System for Distributed Mammography Image Retrieval, University of Regina, Regina Saskatchewan (March 2002)
5. Gibbs, C.: TEEMA Reference Guide, Version 1.0, TRLabs Regina, Saskatchewan 2000-11-01
6. CDC: Centers for Disease Control and Prevention website As seen on December 11 (2006), http://www.cdc.gov/diabetes/statistics/prev/national
7. Boutayeb, A., Chetouani, A.: A Critical Review of Mathematical Models and Data used in Diabetology., BioMedical Engineering Online 2006, 5:43 (June 2006)
8. Makroglou, A., Li, J., Kuang, Y.: Mathematical Models and Software Tools for the Glucose-Insulin Regulatory System and Diabetes: An Overview. Applied Numerical Mathematics 56, 559–573 (2005)
9. Wu, H.-I.: A Case Study of Type 2 Diabetes Self-Management., BioMedical Engineering Online 2005, 4:4 (January 2005)
10. As seen on December 12 (2006), http://www.nutritiondata.com/glycemic-index.html

11. Canadian Coalition for High Blood Pressure Prevention and Control website, As seen on December 12 (2006)
12. As seen on December 12 (2006), http://www.healthsystem.virginia.edu/uvahealth/adult_diabetes/hbp.cfm
13. Massin, M.M., Maeyns, K., Withofs, N., Ravet, F., Gérard, P.: Circadian Rhythm of Heart Rate and Heart Rate Variability. Archives of Disease in Childhood 83, 179–182 (2000)
14. As seen on December 12 (2006), http://www.webmd.com/hw/heart_disease/hw233473.asp

A Multi-agent System for Computer Network Security Monitoring

Agnieszka Prusiewicz

Institute of Information Science & Engineering, Wrocław University of Technology, Poland
agnieszka.prusiewicz@pwr.wroc.pl

Abstract. Due to the complexity of the problem of the network security diagnosis it is necessary to apply effective mechanisms allowing coping with not only knowledge enormity, but also incompleteness and inconsistency. In this connection a proper solution is to apply the multi-agent approach in the network security systems. In this paper the multi-agent system for computer network security monitoring is presented. The overall architecture including agents' knowledge representation, knowledge management and behavior specification is proposed.

Keywords: Intrusion detection, multi-agent approach, DDos attack.

1 Introduction

Network monitoring consists in permanent monitoring of the states of network components for avoiding the breaches of the security policy. The general aim of network security solutions is to provide the stability of the monitored systems. The states of the network components collected from the distributed resources are the basis for the processes responsible for the network security evaluation. A lot of attention has been paid to the problem of the network security maintenance [3, 4, 5, 12]. A comprehensive survey of an anomaly detection system is presented in [10]. A comparison of different approaches to intrusion detection systems is given in [2]. Undoubtedly, the network security systems have to deal with a lot of problems. The most important problem for designers of the network security systems is to find a difference between normal user behaviour and potential attacker [6]. Besides, most of the commercial intrusion detection systems recognise only previously known patterns of attacks. To solve this problem the advanced methods of knowledge processing must be applied. Due to the complexity of the problem of the network security diagnosis it is necessary to apply effective mechanisms allowing coping with not only knowledge enormity, but also incompleteness and inconsistency.

In this connection a proper solution is to apply the multi-agent approach in the network security systems. In works [1, 11] the general description of the security system based on the multi-agent paradigm is presented.

In this work we propose the detailed characteristic of the intrusion detection system based on the multi-agent approach. In work [6] a framework of an original proposal of the intrusion detection system based on the multi-agent approach was presented. In

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 842–849, 2008.

particular, the architecture of such a system and the task of agents were specified. Proposed ideas were further developed and in work [8] the problem of anomalies detection on the basis of the nodes traffic analysis was discussed. However in this paper we proposed agents' internal organization specification with emphasis on the monitoring agents' knowledge representation. Also the procedure of the sources of some attack detection is given. This paper is organised as follows. In Section 2 the overall characteristic of the intrusion detection system is described and the anatomy of some type of attack is explained. In Section 3 the internal agents' organization is proposed. In particular the structure of agents' knowledge is given. Next, in Section 4 the problem of intrusion detection is discussed. In particular the idea of the algorithm for the sources of DDoS attack detection is given. Finally some conclusions are drawn.

2 A Concept of the Intrusion Detection System

In this work distributed denial of service (DDoS) attack is taken into account. Two types of security policy violation with reference to the DDoS attack are considered in this paper: detection of the sources of the attack and detection of the attack propagation scenarios.

It is assumed that network traffic should be analyzed in order to detect anomalies. The main purpose is to detect the attack before "bombarding" chosen goal. In case of the DDoS attack two types of anomalies are observed: anomalies in traffic characteristics and anomalies in communication schemes. For realizing this task standard traffic analysis methods for traffic measurement and also graph methods to identify abnormal communication patterns must be used, see [8] for details. In the DDoS attack it is assumed that *Attacker* is the source of attack and initializes the overall attack in a computer network by putting a maliciously software in chosen computers (nodes). These nodes become *Masters*. In the next step each *Master* starts to infect other nodes that become *Demons*. Next *Demons* infect other nodes and those next ones etc. and in this way spreading stage is expanded. When sufficient number of nodes is infected or at the fixed time point final attack or on earlier chosen computer (attack goal) is carried out. Due to the numerous numbers of requests sent simultaneously from different *Demons* at the same time interval to the *attack goal* all his resources are allocated and blocked [7].

There are two types of agents in our monitoring system: monitoring agents and managing agents. Monitoring agents observe the nodes, capture the observations and draw conclusions that are necessary to evaluate current states of the system security. Monitoring agents obey managing agents that act as system coordinators.

Each monitoring agent subordinates only to one managing agent. Monitoring agent analyzes captured observations in order to detect security policy violation. All the agents communicate with each other using language of communication based on KQML performatives.

A basic goal of this paper is to give a model of agents' knowledge organisation and agents' behaviour characteristic.

3 Agents' Internal Organisation

As it was described in Section 1 two types of agents are considered: monitoring agents and managing agents. Each monitoring agent is equipped with sensors that are used to observe the states of nodes with reference to some properties. The results of observations are stored in monitoring agent's private *database*. This *database* consists of two components: 1) *a set of observations* and 2) *communication data*. *Set of observations* consists of the results of observations with the reference to the states of the nodes belonging to the monitoring region and *communication data* consists of the time points in which the communication between nodes took place. The monitoring agents are also equipped with *the anomaly detection module* that is supplied by the observations from *the set of observations*. The third module closed in internal agents' knowledge structure is *ontology module* that consists of the attack patterns. All the agents are able to communicate with each other to exchange knowledge about the states of nodes, attack patterns and also in order to send requests in case of managing agents using *communication language* compatible with ACL specification [9].

3.1 Agents Knowledge Representation and Behaviour Characteristic

In this section a detailed characteristic of the monitoring agents internal organization and knowledge representation is presented. Let us denote:

- V as the set of the nodes and $V=\{v_1,v_2,\ldots,v_Q\},\ Q\in N$.
- M as the set of the monitoring agents: $MA=\{MA^1,MA^2,\ldots,MA^S\},\ S\in N$.
- ZA as the set of the managing agents: $ZA=\{ZA^1,ZA^2,\ldots,ZA^G\},\ G\in N$.
- MR as the set of the monitoring regions: $MR=\{MR^k : MR^k \subseteq V\}$.
- P as the set of the observed parameters: $P=\{P_1,P_2,\ldots,P_M\},\ M\in N$.

Monitoring agents observe the states of the nodes with the reference to the values of the parameters from the set P in their monitoring regions. Each monitoring agent $MA^k \in MA$ is responsible for one monitoring region $MR^w \in MR$. Managing agents may modify the size of the monitoring regions by adding or removing nodes from the set V. It is assumed that the monitoring regions may overlap. Overlapping the monitoring regions increases the security of the monitoring system. Since a managing agent receives a *warn* message about the anomaly state in a monitored node then he might ask other agents about their opinions. In consequence, the managing agent receives general opinions about a given node from different sources.

Monitoring agents monitor the states of the nodes and capture the results of observations in their private set of observations.

Definition 1. A single observation of agent MA^k is stored as a tuple:

$$O^k\left(v_i,(P_j,x),t_n\right) \in DB^k \tag{1}$$

where $v_i \in MR^k$, $P_j \in P$, $t_n \in T$ and T is the universe of the timestamps and DB^k denotes the database of the agent MA^k.

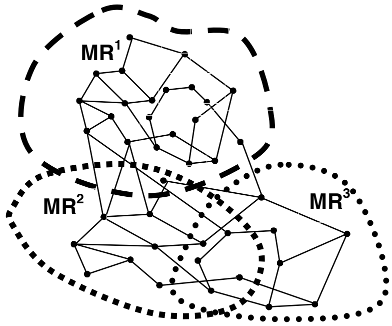


Fig. 1. Monitoring regions of three agents: MA^1, MA^2, MA^3

Such observation refers to the situation that at the timestamp t_n the agent MA^k has observed in the node v_i the value of the parameter P_j equals x.

Additionally, each agent MA^k creates a communication matrix CM^k in which he remembers the timestamps of observed communication acts between nodes.

Definition 2. A *communication matrix* is defined as follows:

$$CM^k = [a_{il}]_{QxQ} \tag{2}$$

where $Q \in N$ and a_{il} contains the timestamps of communication acts between the nodes v_i and v_l.

	v_1	v_2	v_5	v_{12}	v_{13}	v_{17}	v_{31}
v_1							
v_2					a_{213}	a_{217}	
v_5							
v_{12}							
v_{13}							
v_{17}			a_{175}				
v_{31}							

(t_2, t_5, t_9, t_{23}, t_{28}, t_{34})

Fig. 2. En example of communication matrix: the node v_2 communicated with the node v_{12} at the timestamps: t_2, t_5, t_9, t_{23}, t_{28}, t_{34}

Each agent MA^k is equipped with *the anomaly detection module* AD^k (Fig. 3) for observations' analyzing. These observations are stored in DB^k. The aim of *anomaly* *detection module* is detecting anomaly states in the nodes belonging to the agent's MA^k monitoring region with reference to the observed properties.

Definition 3. An *anomaly detection* module of agent MA^k is defined as a triple:

$$AD^k = \left\langle DB^k[t_{b'}, t_{e'}], AT^k, AC^k \right\rangle \quad (3)$$

where:

- $DB^k[t_{b'}, t_{e'}]$ is a subset of DB^k restricted to time interval $[t_{b'}, t_{e'}]$ and is defined as:

  $$DB^k[t_{b'}, t_{e'}] = \left\{ O^k\left(v_i, \left(P_j, x\right), t_n\right) : P_i \in P, v_i \in MR^k; t_n \in [t_{b'}, t_{e'}] \right\} \quad (4)$$

  For a chosen time interval $[t_b, t_e]$ each tuple $O^k\left(v_i, \left(P_j, x\right), t_n\right) \in DB^k$ that fulfills the following condition: $t_n \in [t_{b'}, t_{e'}]$ is sent into *the anomaly detection module* AD^k.
- AT^k is a table of discovered anomalies defined as:

  $$AT^k = \left\{ A^k\left(v_i, P_j, \alpha, [t_b, t_e]\right) : P_i \in P, v_i \in MR^k; [t_b, t_e] \subseteq [t_{b'}, t_{e'}] \right\} \quad (5)$$

  A single anomaly $A^k\left(v_i, P_j, \alpha, [t_b, t_e]\right) \in AT^k$ is interpreted as the agent MA^k believes at the level α that in the time interval $[t_b, t_e]$ in the node v_i an anomaly connected with the property P_i occurred.
- AC^k is an *anomaly archive* consisted of discovered anomalies that are sent from AT^k after attack detection. However only anomalies connected with discovered attack are sent to the archive AC^k. In this way the time interval of discovered anomalies is not a criterion of sending given tuples $A^k\left(v_i, P_j, \alpha, [t_b, t_e]\right)$ into *anomaly archive*.

Anomaly table AT^k is created for each monitoring region independently. On agents' autonomy assumption it is quite possible that one agent may discover anomaly state of the node v_i with reference to some property's value, while other agent having the same node in its monitoring area will not observed any anomaly with reference to this property's value.

The second type of agents, considered in the network monitoring system, are managing agents. Managing agents are responsible for coordination and management of monitoring agents. They do not observe the states of nodes as monitoring agents do, but collect data received from monitoring agents that are sent from *anomaly modules*. Each managing agent is equipped with: *data storage module*, *data analysis module*, *attack detection module*, *queue task module* and *communication module*. In *data storage module* two types of data are stored: 1) data received from the monitoring agents using the messages with *inform* performative and 2) knowledge about the pattern of

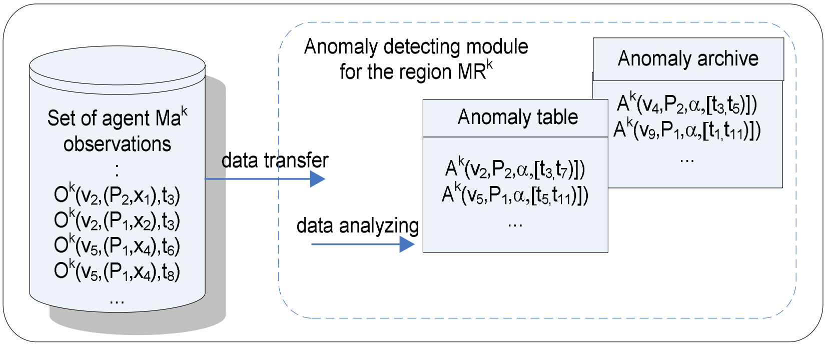


Fig. 3. Anomaly detection module for the region MR^k monitored by the agent MA^k

attacks. In *the data analysis module* conflicts of inconsistent knowledge incoming from distributed resources are solved applying consensus methods. *Attack detection module* is responsible for the attack detection. Managing agents collect results of anomaly detection modules, embodied in monitoring agents, remove existing inconsistencies and create patterns of attacks. In *the attack detection module* there are embodied: the algorithm for determining of the sources attack and the algorithm for determining of the attack propagation scenarios. *The Queue task module* consists of the tasks that must be executed or delegated by managing agent. For example if one managing agent receives *warn* message from monitoring agent that in some node anomaly state has been discovered then managing agent sends a query to the other agents that are also monitoring this node, asking about anomaly occurrence. In case of disconfirmation of the anomaly occurrence the managing agent puts in the *queue task module* the task of verification the state of this node in some time.

4 A Detection of the Security Breach

In the multi-agent system for the computer network monitoring in order to detect a security breaches two approaches are proposed: the procedure for determining of the sources attack and the procedure for determining of the attack scenarios propagation during spreading stage. The aim of applying the procedure for determining of the sources attack is to detect the sources of attacks in the network system. However the aim of the procedure for determining of the attack propagation scenario is to detect not only the sources of attack but also how the attacks propagate in the network system. This procedure consists of two steps. In the first step managing agent requests monitoring agents to determine the local graphs of the attack propagation in their monitoring regions. In the second step the managing agent applies the algorithm for the local graphs integration in order to obtain the consistent scenario of the attack propagation.

In this paper the procedure for determining of the sources attack is given. The goal is to determine the node $v_i^* \in V$ that is an *Attacker*. This procedure consists of two main steps. In the first step the set of *Masters* is detected and in the second one on the base of *Masters'* input traffic analysis the *Attacker* is trailed. The overall procedure for determining of the sources attack is as follows:

Step 1. Determine the set *H* of *Masters*.
Step 2. If the cardinality of the set *H* is higher than threshold value τ then go to Step 3 else go to Step 1.
Step 3. Apply the *procedure of the Attacker detection*.

In the next section the Step 1 of the procedure for determining of the sources attack is described. The aim of the procedure of the Attacker detection is to trail the *Attacker*. For realizing this task, the input Masters' traffic must be analyzed to detect the node that sent maliciously software.

4.1 The Set of the Masters Determining

Let us assume that the set of managing agents is one-element and we detect the anomaly with reference to one monitored property.

The Algorithm for determining of the Set of the Masters
The algorithm for determining of the set of the *Masters* is divided into two parts: *Preliminary part* and *Principal part*.

Preliminary Part
Step 1. Monitoring agent detects the anomaly in the node v*.
Step 2. Monitoring agent sends a w*arn message* to the managing agent.
Step 3. Managing agent asks other monitoring agents about their opinions with the reference to the state of the node v* .
Step 4. Managing agent obtains the consensus opinion about the anomaly state in the node v*.
Step 5. If anomaly exists then managing agent initializes *Principal part* of the algorithm else sends into the *queue task module* the task of the node state verification in some time.

Principal Part
Step 1. Managing agent determines the set *K* of the nodes that communicated with the node *v** in a given time interval.
Step 2. From among the nodes obtained in the Step 2 managing agent determines the set *K'* of infected nodes.
Step 3. If there are no infected nodes i.e. the set *K'* is empty then go to the Step 5 else go to the Step 4.
Step 4. From the set *K'* managing agent selects the node *v'* that communicated with the node *v** at earliest and applies for the node *v'* the *Principal Part*. Go to the Step 1.
Step 5. The node *v** is a *Master*. Include *v** into the set *H*. Stop.

5 Conclusions

In this paper the problem of automatic intrusion detection was discussed. In particular the multi-agent system for network security monitoring was presented. The agent's knowledge structure was described and finally the algorithm for determining of the sources attack was drawn. As a future work presented algorithm must be developed to work out the complete solution. Also the algorithm for *the Attacker* determining of must be formulated.

References

1. Balasubramaniyan, J.S., Garcia-Fernandez, J.O., Isacoff, D., Spafford, E., Zamboni, D.: An Architecture for Intrusion Detection Using Autonomous Agents. In: Proceedings of the 14th Annual Computer Security Applications Conference (1998)
2. Biermann, E., Cloete, E., Venter, L.M.: A comparison of Intrusion Detection systems. Computers and Security 20(8), 676–683 (2001)
3. Bejtlich, R.: Tao of Network Security Monitoring, The: Beyond Intrusion Detection. Addison-Wesley, Reading (2004)
4. Dasgupta, D.: Immunity-Based Intrusion Detection System: A General Framework. In: Proceedings of the 22nd National Information Systems Security Conference, USA (1999)
5. Denning, D.E., Edwards, D.L., Jagannathan, R., Lunt, T.F., Neumann, P.G.: A prototype IDES: A real-time intrusiondetection expert system. Technical report, Computer Science Laboratory, SRI International, Menlo Park (1987)
6. Kolaczek, G., Pieczynska, A., Juszczyszyn, K., Grzech, A., Katarzyniak, R., Nguyen, N.T.: A mobile agent approach to intrusion detection in network systems, Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3682, pp. 514–519. Springer, Heidelberg (2005)
7. Mirkovic, J., Dietrich, S., Dittrich, D., Reiher, P.: Internet Denial of Service: Attack and De fense Mechanisms. Prentice-Hall, Englewood Cliffs (2005)
8. Nguyen, N.T., Juszczyszyn, K., Kolaczek, G., Grzech, A., Pieczynska, A., Katarzyniak, R.: Agent-based approach for distributed intrusion detection system design. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)
9. FIPA Specifications: `http://www.fipa.org/`
10. Patcha, A., Park, J.-M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51(12), 3448–3470 (2007)
11. Spafford, E., Zamboni, D.: Intrusion detection using autonomous agents, Computer Networks. The International Journal of Computer and Telecommunications Networking 34(4), 547–570 (2000)
12. Wilson, E.: Network Monitoring and Analysis: A Protocol Approach to Troubleshooting. Prentice-Hall, Englewood Cliffs (1999)

Vulnerabilities in a Remote Agent Authentication Scheme Using Smart Cards*

Youngsook Lee[1], Junghyun Nam[2], and Dongho Won**,[1]

[1] Department of Computer Engineering, Sungkyunkwan University, Korea
{yslee,dhwon}@security.re.kr
[2] Department of Computer Science, Konkuk University, Korea
jhnam@kku.ac.kr

Abstract. Agent technology is emerging as a new software paradigm in the areas of distributed computing. The use of multiple agents is a common technique in agent-based systems. In distributed agent systems, for secure communication, the communicating agents should authenticate each other by using authentication protocols. A remote agent authentication scheme is a two-party protocol whereby an authentication server in a distributed system confirms the identity of a remote individual logging on to the server over an untrusted, open network. This paper discusses the security of Yoon et al.'s remote agent authentication scheme making use of smart cards. Yoon et al.'s scheme was proposed to solve the security problem with Hwang et al.'s authentication scheme and was claimed to provide mutual authentication between the server and the remote agent. But, unlike the claim, in Yoon et al.'s scheme, if an attacker steals some agent's smart card and extracts the information stored in the smart card, he/she can violate the authentication goal of the scheme without knowing the agent's password. We show this by mounting two attacks, a agent impersonation attack and a sever impersonation attack, on Yoon et al's scheme. In addition, in Yoon et al.'s scheme, if an attacker steals some agent's smart card and extracts the information stored in the smart card and reads U_i's login massage, he/she can violate its fundamental goal of a password security. We show this by mounting a dictionary attack on Yoon et al.'s scheme and also figure out what has gone wrong with the scheme.

Keywords: distributed system, authentication scheme, smart card, impersonation attack.

1 Introduction

Agent technology is a new generation of intelligent software designed to manage complex, real-time systems. Unlike other software, agents can be programmed

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement).
** Corresponding author.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 850–857, 2008.

to operate proactively to achieve predefined goals. The use of multiple agents is a common technique in agent-based systems. In distributed agent systems, it is often required for two agents to communicate securely over a public network. For secure communication, the communicating agents should authenticate each other by using authentication protocols. In a typical password-based authentication scheme [12,4,18,8,17,5,20,13] using smart cards, remote agents are authenticated using their smart card as an identification token; the smart card takes as input a password from a agent, recovers a unique identifier from the agent-given password, creates a login message using the identifier, and then sends the login message to the server who then checks the validity of the login request before allowing access to any services or resources. This way, the administrative overhead of the server is greatly reduced and the remote agent is allowed to remember only his password to log on. Besides just creating and sending login messages, smart cards support mutual authentication where a challenge-response interaction between the card and the server takes place to verify each other's identity. Mutual authentication is a critical requirement in most real-world applications where one's private information should not be released to anyone until mutual confidence is established [1].

The experience has shown that the design of secure authentication schemes is not an easy task to do, especially in the presence of an active adversary; there is a long history of schemes for this domain being proposed and subsequently broken by some attacks (e.g., [6,2,3,14,7,20,19,11,16]). Therefore, authentication schemes must be subjected to the strictest scrutiny possible before they can be deployed into an untrusted, open network, which might be controlled by an adversary.

To analyze the security of remote agent authentication schemes using smart cards, we need to consider the capabilities of the adversary. First, we assume that the adversary has complete control of every aspect of all communications between the server and the remote agent. That is, he/she may read, modify, insert, delete, replay and delay any messages in the communication channel. Second, he/she may try to steal a agent's smart card and extract the information in the smart card by monitoring the power consumption of the smart card [10,15]. Third, he/she may try to find out a agent's password. Clearly, if both (1) the argent's smart card was stolen and (2) the agent's password was exposed, then there is no way to prevent the adversary from impersonating the agent. However, a remote agent authentication scheme should be secure if only one of (1) and (2) is the case. So the best we can do is to guarantee the security of the scheme when either the agent's smart card or its password is stolen, but not both.

In 2002, Hwang et al. [9] proposed a simple remote agent authentication scheme. Hwang et al.'s scheme exhibits various merits: (1) it does not require the server to maintain a password table for verifying the legitimacy of login agents; (2) it allows agents to choose and change their passwords according to their liking and hence gives more agent convenience; and (3) it is extremely efficient in terms of the computational cost since the protocol participants perform only a few hash function operations. However, Yoon et al. [20] have pointed out that

Hwang et al.'s scheme is vulnerable to a denial of service attack, in which the server rejects any login request regardless of whether the login request is valid or not. To fix this security problem, Yoon et al. have presented a modified scheme of Hwang et al.'s scheme, and have claimed, among others, that their modified scheme achieves mutual authentication between the server and remote agents. However, in Yoon et al.'s scheme, if an attacker steals some agent's smart card and extracts the information stored in the smart card, he/she can violate the authentication goal of the scheme without knowing the agent's password. We show this by mounting two attacks, a agent impersonation attack and a server impersonation attack, on Yoon et al.'s scheme. The agent impersonation attack and the server impersonation attack are given to infringe respectively agent-to-server and server-to-agent authentications of Yoon et al.'s scheme. Besides reporting two attacks on Yoon et al.'s scheme, in Yoon et al.'s scheme, if an attacker steals some agent's smart card and extracts the information stored in the smart card and reads U_i's login massage, he/she can violate its fundamental goal of a password security. We show this by mounting a dictionary attack on Yoon et al.'s scheme and also figure out what has gone wrong with the scheme.

The remainder of this paper is organized as follows. We begin by reviewing Yoon et al.'s remote agent authentication scheme in Section 2. Then in Section 3, we present security weaknesses in Yoon et al.'s authentication scheme. Finally, we conclude this work in Section 4.

2 Review of Yoon et al.'s Authentication Scheme

Yoon et al. [20] have recently presented an improved version of Hwang et al.'s [9] scheme. Besides preventing the denial of service attack, Yoon et al.'s scheme intends to provide mutual authentication between the remote server and the agent. We begin by describing the top level structure of the scheme. The scheme consists of two phases: registration phase and authentication phase. The registration phase is performed only once per agent when a new agent registers itself with the remote server. The authentication phase is carried out whenever a agent wants to gain access to the server.

Before the registration phase is performed for the first time, the server S decides on the following system parameters: a one-way hash function h, a cryptographic key x, and a timestamp T_{TSA}, where T_{TSA} is generated by a trusted time stamping authority. The key x and the timestamp T_{TSA} are kept secret by the server. A high level depiction of the scheme is given in Fig. 1, where dashed lines indicate a secure channel, and a more detailed description follows:

Registration Phase. The registration of a new agent U_i to the server S proceeds as follows:

Step 1. A agent U_i, who wants to register with the server S, chooses its password PW_i at will and submits a registration request $\langle ID_i, PW_i \rangle$ to the remote server S via a secure channel.

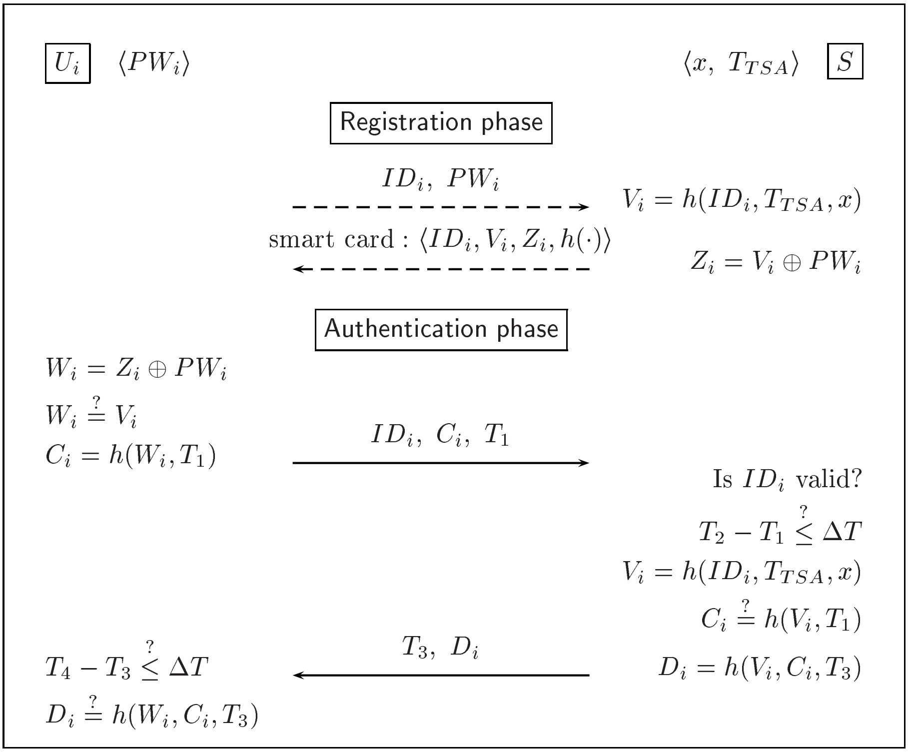


Fig. 1. Yoon et al.'s authentication scheme

Step 2. When the server S receives the request, it first computes

$$V_i = h(ID_i, T_{TSA}, x) \quad \text{and} \quad Z_i = V_i \oplus PW_i$$

and then issues a smart card containing $\langle ID_i, V_i, Z_i, h(\cdot) \rangle$ to the agent U_i.

Authentication Phase. This phase constitutes the core of the scheme, and is performed whenever some agent U_i wants to log on to the server S. U_i initiates this phase by inserting its smart card into a card reader and then entering its identity ID_i and password PW_i. Given the agent input, the smart card and the server perform the following steps:

Step 1. Using the agent-given PW_i, the smart card computes $W_i = Z_i \oplus PW_i$. Then the smart card checks that W_i is equal to the value V_i stored in its smartcard. If they are equal, the smart card proceeds to the next step. Otherwise, the smart card aborts the authentication phase.

Step 2. The smart card generates the current timestamp T_1 and computes $C_i = h(W_i, T_1)$. Then the smart card sends the login request message $\langle ID_i, C_i, T_1 \rangle$ to the server S.

Step 3. When the login request arrives, S first acquires the current timestamp T_2 and computes $V_i = h(ID_i, T_{TSA}, x)$. Then S verifies that: (1) ID_i is valid, (2) $T_2 - T_1 \leq \Delta T$, where ΔT is the maximum allowed time interval for transmission delay, and (3) C_i equals $h(V_i, T_1)$. If all of these conditions hold, S generates a new timestamp T_3, computes $D_i = h(V_i, C_i, T_3)$, and sends the response message $\langle T_3, D_i \rangle$ to the smart card. Otherwise, S rejects it and aborts the protocol.

Step 4. After receiving $\langle T_3, D_i \rangle$, the smart card obtains a new timestamp T_4 and checks that: (1) $T_4 - T_3 \leq \Delta T$ and (2) D_i is equal to $h(W_i, C_i, T_3)$. If both of these conditions hold, the smart card believes the responding party as the authentic server. Otherwise, the smart card aborts its login attempt.

3 Attacks on Yoon et al.'s Authentication Scheme

3.1 Impersonation Attacks

As stated in the Introduction, it is desirable that any attacker who does not know U_i's password cannot impersonate U_i to S and impersonate S to U_i even if the attacker has gained access to U_i's smart card and to the information inside it. But, Yoon et al.'s scheme does not satisfy this desideratum. To show this, we present two impersonation attacks, a agent impersonation attack and a server impersonation attack. In describing the attacks, we assume that the attacker has stolen U_i's smart card and extracted the secret values stored in it by monitoring its power consumption [10,15].

Impersonating U_i to S. First, we present a agent impersonation attack where an attacker U_a is able to impersonate a legitimate agent U_i to the remote server S without knowing U_i's password. The corresponding attack scenario is depicted in Fig. 2, and is described in detail as follows:

1. The attacker U_a generates the current timestamp T_1 and computes $C_i' = h(V_i, T_1)$ by using the value V_i extracted from U_i's smart card.
2. Then U_a sends the forged login request message $\langle ID_i, C_i', T_1 \rangle$ to the server S.
3. After receiving $\langle ID_i, C_i', T_1 \rangle$, the server S proceeds to verify the authenticity of the login request. S first acquires the current timestamp T_2, then computes $V_i = h(ID_i, T_{TSA}, x)$, and finally checks that: (1) ID_i is valid, (2) $T_2 - T_1 \leq \Delta T$, and (3) C_i' equals $h(V_i, T_1)$. Since all of these conditions hold, S will welcome U_a's visit to the system without noticing any discrepancy.
4. Thus, the server S proceeds as specified in the scheme, generating the timestamp T_3, computing $D_i = h(V_i, C_i', T_3)$, and then sending $\langle T_3, D_i \rangle$ to U_a.

Impersonating S to U_i. Now, we present a server impersonation attack where an attacker U_a can easily impersonate the remote server S to a legitimate agent U_i. As depicted in Fig. 3, the attack works as follows:

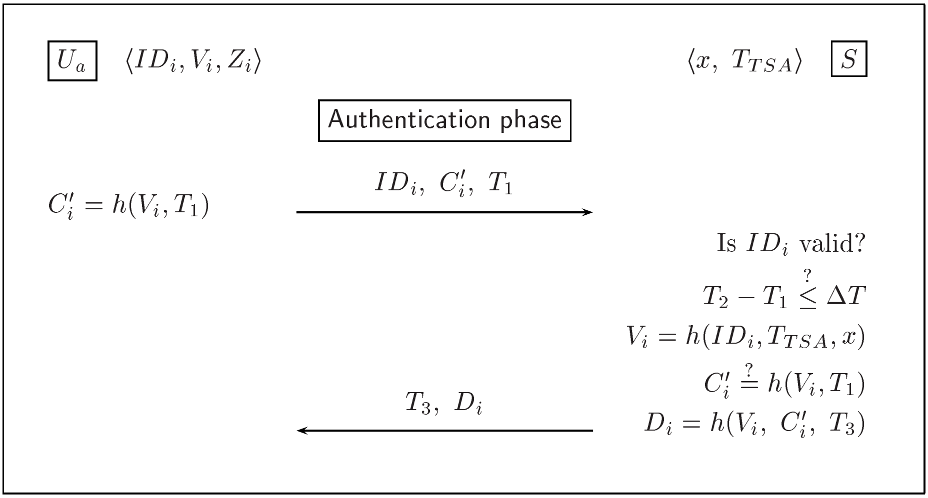


Fig. 2. A agent impersonation attack on Yoon et al.'s scheme

1. As usual, the authentication phase begins when agent U_i sends the login request message $\langle ID_i, C_i, T_1 \rangle$ to the server S.
2. But, the attacker U_a posing as S intercepts this login request and sends immediately back to U_i a forged response message as follows: U_a first generates the current timestamp T_3, computes D_i=$h(V_i, C_i, T_3)$ by using the value V_i extracted from U_i's smart card, and sends $\langle T_3, D_i \rangle$ in response to U_i's login request.
3. The forged response $\langle T_3, D_i \rangle$ will pass the verification test by U_i since the timestamp T_3 is valid and D_i is equal to $h(W_i, C_i, T_3)$. Hence, U_i believes U_a as the authentic server.

The vulnerability of Yoon et al.'s scheme to the impersonation attacks is due to the following fact: to forge a valid login request message $\langle ID_i, C_i, T_1 \rangle$ or a valid response message $\langle D_i, T_3 \rangle$, it suffices to obtain the information stored in U_i's smart card. More concretely, the problem with the scheme is that anyone who knows the value of V_i (=W_i) is able to compute $C_i = h(W_i, T_1)$ (even without knowing U_i's password PW_i) or able to compute $D_i = h(V_i, C_i, T_3)$ (even without knowing the server's secret value x and the timestamap T_{TSA}).

3.2 Password Guessing Attack

We showed that Yoon et al.'s remote agent authentication scheme is vulnerable to a password guessing attack in which an attacker U_a can easily find out the password of the agent U_i by eavesdropping a login request message of it. In this attack, it is assume that the attacker U_a has obtained the secret values extracted from U_i's smart card. The attack scenario is described as follows:

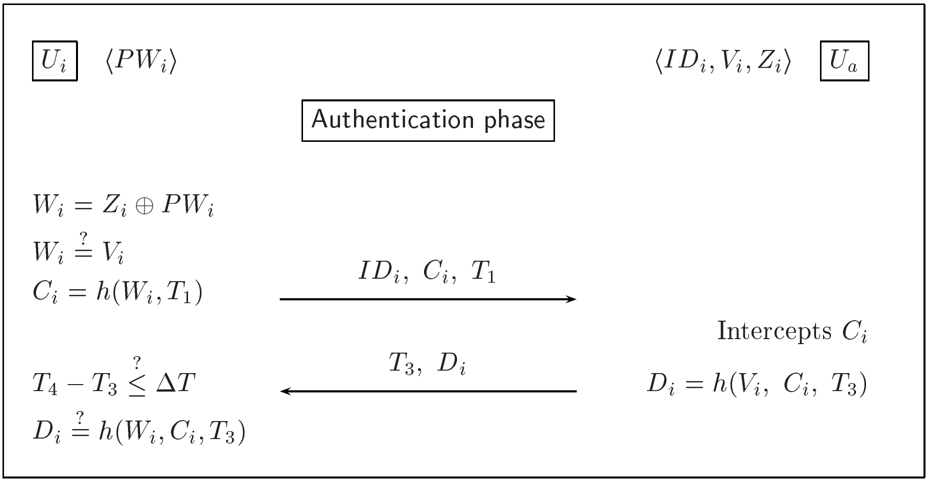


Fig. 3. A server impersonation attack on Yoon et al.'s scheme

1. The agent U_i wants to log on to the remote server and thus it sends the login message $\langle ID_i, C_i, T_1 \rangle$ to the server S.
2. But, at this moment, U_a reads this login request message and thereby obtains C_i and T_1 contained in the login request.
3. Finally, U_a who has extracted Z_i from U_i's smart card, can find out PW_i by employing a dictionary attack, in which each guess PW_i' for PW_i can be verified by computing $W_i' = Z_i \oplus PW_i'$ and $C_i' = h(W_i', T_1)$, and by checking the equality $C_i' = C_i$.

In Yoon et al.'s scheme, the value Z_i contained in the agent U_i's smart card is used in computing $C_i : C_i = h(W_i, T_1)$, where $W_i = Z_i \oplus PW_i$. Based on this observation, the attacker succeeds in finding the correct password of U_i by comparing C_i obtained from agent U_i's login request with C_i' obtained using the guessed password PW_i'.

4 Conclusion

We have analyzed the security of the smart card based agent authentication scheme proposed by Yoon et al. [20]. Our security analysis uncovered that Yoon et al.'s scheme does not achieve its main security goal of authenticating between a remote individual and the server. The failure of Yoon et al.'s scheme to achieve authentication has been made clear through two attacks, a server impersonation attack and a agent impersonation attack, on the scheme. The server impersonation attack and the agent impersonation attack have been considered to infringe respectively server-to-agent and agent-to-server authentications of the scheme. In addition, Yoon et al.'s scheme is vulnerable to a password guessing attack.

We show this by mounting a a dictionary attack on Yoon et al.'s scheme. Besides reporting the security problems, we showed what really is causing the problem.

References

1. Anti-Phishing Working Group, http://www.antiphishing.org
2. Bird, R., Gopal, I., Herzberg, A., Janson, P.A., Kutten, S., Molva, R., Yung, M.: Systematic design of a family of attack-resistant authentication protocols. IEEE Journal on Selected Areas in Communications 11(5), 679–693 (1993)
3. Carlsen, U.: Cryptographic protocol flaws: know your enemy. In: Proceedings of the 7th IEEE Computer Security Foundations Workshop, pp. 192–200 (1994)
4. Chang, C.-C., Wu, T.-C.: Remote password authentication with smart cards. IEE Proceedings E - Computers and Digital Techniques 138(3), 165–168 (1991)
5. Chien, H.-Y., Jan, J.-K., Tseng, Y.-M.: An efficient and practical solution to remote authentication: smart card. Computers & Security 21(4), 372–375 (2002)
6. Diffie, W., van Oorschot, P.C., Wiener, M.J.: Authentication and authenticated key exchange. Designs, Codes and Cryptography 2(2), 107–125 (1992)
7. Hsu, C.-L.: Security of Chien et al.'s remote user authentication scheme using smart cards. Computer Standards and Interfaces 26(3), 167–169 (2004)
8. Hwang, M.-S., Li, L.-H.: A new remote user authentication scheme using smart cards. IEEE Trans. on Consumer Electronics 46(1), 28–30 (2000)
9. Hwang, M.-S., Li, L.-H., Tang, Y.-L.: A simple remote user authentication. Mathematical and Computer Modelling 36, 103–107 (2002)
10. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M.J. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999)
11. Ku, W.-C., Chang, S.-T., Chiang, M.-H.: Weaknesses of a remote user authentication scheme using smart cards for multi-server architecture. IEICE Trans. on Commmunications E88-B(8), 3451–3454 (2005)
12. Lamport, L.: Password authentication with insecure communication. Communications of the ACM 24(11), 770–772 (1981)
13. Lee, Y., Nam, J., Won, D.: Security enhancement of a remote user authentication scheme using smart cards. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4277, pp. 508–516. Springer, Heidelberg (2006)
14. Lowe, G.: An attack on the Needham-Schroeder public-key authentication protocol. Information Processing Letters 56(3), 131–133 (1995)
15. Messerges, T.-S., Dabbish, E.-A., Sloan, R.-H.: Examining smart card security under the threat of power analysis attacks. IEEE Trans. on Computers 51(5), 541–552 (2002)
16. Nam, J., Kim, S., Park, S., Won, D.: Security analysis of a nonce-based user authentication scheme using smart cards. IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences E90-A(1), 299–302 (2007)
17. Sun, H.-M.: An efficient remote user authentication scheme using smart cards. IEEE Trans. on Consumer Electronics 46(4), 958–961 (2000)
18. Yang, W.-H., Shieh, S.-P.: Password authentication schemes with smart card. Computers & Security 18(8), 727–733 (1999)
19. Yoon, E.-J., Kim, W.-H., Yoo, K.-Y.: Security enhancement for password authentication schemes with smart cards. In: Katsikas, S.K., López, J., Pernul, G. (eds.) TrustBus 2005. LNCS, vol. 3592, pp. 90–99. Springer, Heidelberg (2005)
20. Yoon, E.-J., Ryu, E.-K., Yoo, K.-Y.: An improvement of Hwang-Lee-Tang's simple remote user authentication scheme. Computers & Security 24(1), 50–56 (2005)

A Security Enhancement of the E0 Cipher in Bluetooth System

HyeongRag Kim[1], HoonJae Lee[2], and SangJae Moon[3]

[1] Dept. of Computer Application, Pohang College, Pohang, Korea
[2] School of Computer Engineering, Dongseo University, Busan, Korea
[3] School of Electrical Eng. &Computer Science, Kyungpook Nat'l Univ., Daegu, Korea
{Hyeong-Rag Kim, hrkim@pohang.ac.kr}

Abstract. Summation generator was analyzed by Dawson's divide-and-conquer- attack[1] and showed us the weakness by Golic's correlation attack[3] and Meier's Fast correlation attack[4]. In this paper, We propose a Circular-Clock-Controlled Summation Generator(CCC-SG), an enhanced E0 algorithm for GSM system, and analyze the security aspects and the performances.

Keywords: Cryptosystem, CCC-SG , Stream cipher.

1 Introduction

Bluetooth is the main technology which was developed by the Swedish Ericson in 1998 , and it uses E0 algorithm for the user's information security. In this moment it uses the summation generator with 4 pieces of linear feedback shift register(LFSR) for the generation of key streams. LFSR is fit for hardware and software, which is used by the stream ciphers so much that support fast encryption and decryption. Furthermore the sequence generated by LFSR with primitive polynomial has the long-term period and statistical merit. But LFSR is easy to be predicted (cryptanalysis) because of its linear property from the output sequence, and in the case when the feedback polynomial is known, the whole period of key stream will be the length of "L" of LFSR from the continual L unit of sequence. Meanwhile in case of not being known, it will be 2L unit[1]. They generally use LFSR for the element of key summation generator for not only evading the weakness of linear property but also using the statistical merit and enlarge the nonlinear property of stream cipher with use of nonlinear double function used with such conjugation function as filter function and with the use of the irregular clock control LFSR.

Summation generator was proposed by Rueppel[2] in 1985 as the key stream generator for stream cipher. The summation generator is using the binary LFSRs of " r " each with constant clock and $\lceil \log_2^r \rceil$ bit memory, output is gained with the input summation over the integers. LSB(least significant bit) of summation produces key stream and the rest bits which are carry bits are put into the memory. The carry sequence is to be used as the input of combining function for the production of next bits.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 858–867, 2008.

Summation generator was analyzed by Dawson's divide-and-conquer-attack[1] and showed us the weakness by Golic's correlation attack[3] and Meier's Fast correlation attack[4].

In this paper, we propose a Circular-Clock-Controlled Summation Generator(CCC-SG), an enhanced E0 algorithm[5] for GSM system, and analyze the security aspects and the performances.

2 A proposal of the CCC-SG generator

2.1 E0 Encryption Algorithm

The user is under protection of the encryption of packet payload encryption in bluetooth technology, which is the communication provider between individuals in a short distance. In this moment, the data confidential service is given by a stream cipher called E0[5]. The block diagram of E0 is shown in Fig. 1.

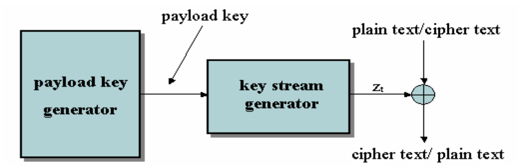


Fig. 1. Stream Ciphering for Bluetooth with E0

The E0 stream cipher is consists of three parts. The first part executes the initial stage, the second part produces key stream bits and the third part executes encryption and decryption process. After payload key generator combines the input bits in a suitable procedure, they are transferred to each of the four LFSRs which are used in key stream generation. Key stream generator uses summation stream cipher generator proposed by Massey and Rueppel [1,2].

2.2 In this clause, we propose a Circular-Clock-Controlled Summation Generator (CCC-SG) , an improved version of the irregular-clock-controlled type for enhancing a nonlinearity of the original summation generator in the bluetooth encryption algorithm.

2.2.1 A Proposal of the CCC-SG Generator

The E0 algorithm is the summation generator based on each of the four LFSRs as depicted in Fig. 2. In this case, x_t^1, x_t^2, x_t^3 and x_t^4 are the output of LFSR1, LFSR2, LFSR3 and LFSR4 respectively and SG box output z_t is s_t^0, where s_t^0 is least significant bit of $s_t = (s_t^2, s_t^1, s_t^0)$. In Fig. 3, Carry c_t^0 is decided by the output of C box that receives s_t as input.

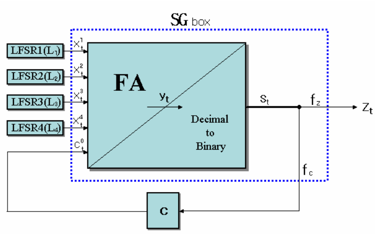


Fig. 2. E0 Encryption Engine

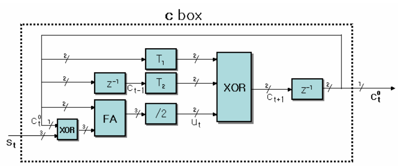


Fig. 3. C Box

In Fig. 2, when x_t^i is marked as the symbol of t^{th} of $LFSR_i$, the value of y_t will be induced from each of the four inputs $x_t^1,\ldots, x_t^4$ as bellow.

$$y_t = \sum_{i=1}^{4} x_t^i \tag{1}$$

The value of y_t can get the value of 0,1,2,3,4 or 5. Now the output of summation generator is to get the below formula.

$$z_t = f_z(x_t^1, x_t^2, x_t^3, x_t^4, c_t^0) = x_t^1 \oplus x_t^2 \oplus x_t^3 \oplus x_t^4 \oplus c_t^0 \in \{0,1\} \tag{2}$$

$$u_{t+1} = (u_{t+1}^1, u_{t+1}^0) = \left\lfloor \frac{s_t + c_t}{2} \right\rfloor \in \{0,1,2,3\} \tag{3}$$

$$c_{t+1} = f_c(x_t^1, x_t^2, x_t^3, x_t^4, c_t^0) = (c_{t+1}^1, c_{t+1}^0) = u_{t+1} \oplus T_1[c_t] \oplus T_2[c_{t-1}] \tag{4}$$

In above formula (3), $T_1[\cdot], T_2[\cdot]$ are two different linear bijections on the GF(4). Supposing that GF(4) is produced by the minimum polynomial of $x^2 + x + 1$ and α at GF(4) is put as zero of this polynomial, the mapping of T_1 and T_2 shall be defined as below formula (5) and formula (6).

$$T_1 = GF(4) \to GF(4) \qquad x \mapsto x \tag{5}$$

$$T_2 = GF(4) \to GF(4) \qquad x \mapsto (\alpha + 1)x \tag{6}$$

As summarized at Table 1, elements of GF(4) can be used as binary vector.

Table 1. The mapping T1 and T2

x	$T_1[x]$	$T_2[x]$
00	00	00
01	01	11
10	10	01
11	11	10

As the mapping is a linear, LSFR can be realized as formula (7) and formula (8) with the use of XOR gate.

$$T_1 : (x_1, x_0) \mapsto (x_1, x_0) \tag{7}$$

$$T_2 : (x_1, x_0) \mapsto (x_0, x_1 \oplus x_0) \tag{8}$$

CCC-SG is the summation generator added with circular-clock-control structure, which is the same as Fig. 4. In the figure, key stream generator consists of four LFSRs and output bits of LFSR are to be input to the combining function of f_z and carrying function of f_c in order to produce the next memory status and key stream bits. The irregular clock is supplied to the LFSR, the four irregular clocks to LFSRs are gained from the nonlinear filter function(g_a, g_b, g_c or g_d) produced by other LFSRs. The status of each LFSR is defined by the memory status of each LFSR and the output z_t at time point "t" is produced by f_z. The clock control function of g_a, g_b, g_c and g_d are gained by the present status of LFSR and LFSR produces output Carry and Key Stream after clock control at random. In addition, the number of irregular clocking in LFSR is decided by according to the contents of two specified taps from different LFSR at random. In the first stage of CCC-SG, the internal status are filled with key (k) and the initial vector (i, initial key), and the process of key expansion for filling up the internal status is required additionally because the length of the internal status is longer than the key.

2.2.2 A Proposal of the CCC-SG Generator

In this clause, we deal with key stream generator and clock control of CCC-SG.

2.2.2.1 Key stream generation. CCC-SG has both circular-clock-controlled LFSRs and Carry bit, and the length of LFSRs are $L_1 = 117$, $L_2 = 127$, $L_3 = 131$, $L_4 = 137$. All memory bits supplied to CCC-SG with the internal status of 512-bit and the internal status is filled with both 256-bit key and the initial vector (iv) of 256-bit. The output key sequences of CCC-SG are produced with summation of the

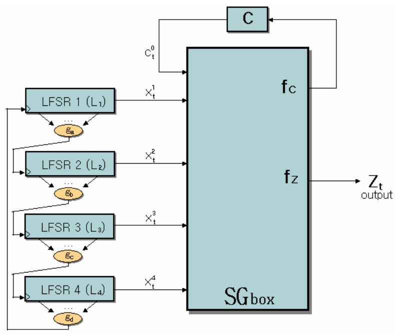


Fig. 4. Circular-Clock-Controlled Summation Generator

sequences of LFSR and Carry. L_1, L_2, L_3 and L_4 of feedback polynomial are chosen as the below primitive polynomial of $p_1(x)$, $p_2(x)$, $p_3(x)$ and $p_4(x)$ by each and it is not allowed to be the first stage as all-zero-state of all bits of LFSR.

$$p_1(x) = x^{117} \oplus x^{101} \oplus x^8 \oplus x^7 \oplus x^6 \oplus x^5 \oplus x^4 \oplus x^3 \oplus x^2 \oplus x \oplus 1 \quad (9)$$

$$p_2(x) = x^{127} \oplus x^{110} \oplus x^{11} \oplus x^9 \oplus x^8 \oplus x^7 \oplus x^6 \oplus x^5 \oplus x^4 \oplus x^3 \oplus x^2 \oplus x \oplus 1 \quad (10)$$

$$p_3(x) = x^{131} \oplus x^{71} \oplus x^{10} \oplus x^7 \oplus x^6 \oplus x^5 \oplus x^4 \oplus x^3 \oplus x^2 \oplus x \oplus 1 \quad (11)$$

$$p_4(x) = x^{137} \oplus x^{136} \oplus x^{14} \oplus x^9 \oplus x^8 \oplus x^7 \oplus x^6 \oplus x^5 \oplus x^4 \oplus x^3 \oplus x^2 \oplus x \oplus 1 \quad (12)$$

The output key stream bit z_t and Carry bit c_t^0 take the same shape (Formula (2) ~ (4)] in the E0 structure, but the output sequence and the security level shall be improved much better.

2.2.2.2 Clock Control. In CCC-SG, each of the LFSR control the clocks of the different LFSRs at random to produce the irregular clocks to each register. After taking the random value range between 1 ~ 4 from the tap of L_1, they control the clock of L_2, control L_3 from the value of two tap in L_2, control L_4 from the value of two tap in L_3 and control L_1 from the value of two tap in L_4. The control function g_a, g_b, g_c and g_d are defined as below.

$$g_a(L_a) = 2L_{a42}(t) + L_{a85}(t) + 1 \quad (13)$$

$$g_b(L_b) = 2L_{b43}(t) + L_{b86}(t) + 1 \quad (14)$$

$$g_c(L_c) = 2L_{c44}(t) + L_{c87}(t) + 1 \quad (15)$$

$$g_d(L_d) = 2L_{d45}(t) + L_{d88}(t) + 1 \quad (16)$$

This device is applied to CCC-SG, and in the key stream generator, output of the i^{th} time of LFSR based on "n" is used to clock the $i+1^{th}$ time of LFSR. As a result, CCC-SG is possible for expanding by the $LFSR_n$.

3 Analysis

In this clause, we show the properties of keystream of CCC-SG based on the experimental results, and the security properties of CCC-SG against the known attack.

3.1 The Property of Keystream

The three basic requirements for the PN binary sequences are a long-term period, a high linear complexity and a good statistical property, and when we use a encrypted long message, the long-term period protects the re-use of the same key and the high linear complexity makes it be endurable against the attack which was utilizing Berlekamp-Massey algorithm[6]. Finally the better statistical property makes it endurable against the attack utilizing the weakness of the key stream biased to any direction between " 0 " and " 1 ".

We call each of the binary m-sequence as relative prime that are $\{x_t^1\}$, $\{x_t^2\}$, $\{x_t^3\}$ and $\{x_t^4\}$ as well as the minimum primitive polynomial of L_1 , L_2 , L_3 and L_4 .

In this time, when the sequence of $\{x_t^1\}$, $\{x_t^2\}$, $\{x_t^3\}$ and $\{x_t^4\}$ are added on the real number, the sequence of the real number summation of $\{z_t\}$ is satisfied in formula(17) at very close to the period of linear complexity LC value[5].

$$LC(z_t) \le (2^{L1}-1)(2^{L2}-1)(2^{L3}-1)(2^{L4}-1) \quad (17)$$

The result of simulation in Table 2 shows that Formula(17) represents upper limits very suitable instead of the proof of Formula (17).

Table 2. The small-scale simulation giving evidence that the bound (17) is very tight

N	L_1	L_2	L_3	L_4	T	LC
7	2	3	5	7	82,767	82,676
19	3	4	5	7	413,385	41,3384
21	2	3	5	11	1,332,597	1,332,593
512	117	127	131	137	$(2^{117}-1)(2^{127}-1)(2^{131}-1)(2^{137}-1)$	

In the table 2, column of " N " represents the number of M-sequences being added on the real number. The column of L_1 , L_2 , L_3 and L_4 is a degree of the minimum polynomial in the m-sequences to be added. T column represent the summation sequence period, and the LC column shows the value of the minimum linear complexity acquired from all possible combination of the different primitive polynomial uttered a degree. In the results of small-sized simulation of table 1 about short-stage, the value of the linear complexity LC shows it very close to the period of T, that is similar to the original summation generator. If expand N value to 512, LC value will be very close to the value of the period. The cryptographic strength of CCC-SG is 2^{256} and the basic property of keystream is safe against various attacks because of both big linear complexity and the long-term period.

3.2 Attack Analysis

In this clause we explain the divide-and-conquer-attack, fast correlation attack and Time/Memory/Data Tradeoff attack of CCC-SG.

3.2.1 Divide-and-Conquer-Attack

The known divide-and-conquer-attack algorithm is possible only when we know the correct clock[1]. But the applied circular-clock-control structure in CCC-SG is secure in this attack because both of the value are between 1 ~ 4 continuously by each clock and correct clock information not being known to the cipher attack.

3.2.2 Fast correlation Attack

CCC-SG has a correlation property with memory bit c^0_{t+1} (here, $c_{t+1}=(c^1_{t+1}, c^0_{t+1})$) and key output stream bit z_t as like Table 3. That is $p(c^0_{t+1}=z_t)=0.516$. It is $z_t=x^1_t \oplus x^2_t \oplus x^3_t \oplus x^4_t \oplus c^0_t$ and also it is $z_{t+1} = x^1_{t+1} \oplus x^2_{t+1} \oplus x^3_{t+1} \oplus x^4_{t+1} \oplus c^0_{t+1}$. As the $c^0_{t+1} = z_t \oplus 1$ keeps it the probability of 0.488, $z_{t+1}=x^1_{t+1} \oplus x^2_{t+1} \oplus x^3_{t+1} \oplus x^4_{t+1} \oplus z_t \oplus 1$ keeps the probability of 0.488.

Supposing z_t' as binary time-lag bit of key stream, it will be $z_t'=z_{t+1} \oplus z_t$ at the temporary point of time, and at this moment it becomes as $z_t' = x^1_{t+1} \oplus x^2_{t+1} \oplus x^3_{t+1} \oplus x^4_{t+1} \oplus z_t \oplus 1$. The binary time-lag stream z_t' of CCC-SG will be summation of both the outputs of the four LFSRs and binary noise e_t as in Fig. 5, and the noise probability of this model is 0.012 . This is possible for correlation property attack[3] .

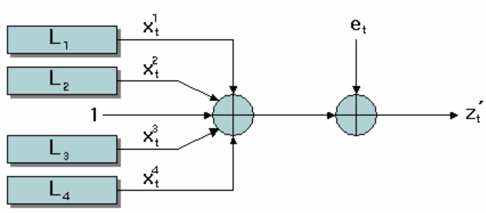


Fig. 5. Fast correlation property Attack of Summation Generator

Table 3. Correlation property of c_{t+1} and z_t of CCC-SG

x_t^1	x_t^2	x_t^3	x_t^4	c_t^0	$c_t=(c_t^1,c_t^0)$	c_{t+1}^0	$c_{t+1}=(c_{t+1}^1,c_{t+1}^0)$	Z_t
0	0	0	0	0	00 10	0 0	00 10	0
0	0	0	0	1	01 11	0 0	10 00	1
0	0	0	1	0	00 10	0 1	00 01	1
0	0	0	1	1	01 11	0 1	10 11	0
0	0	1	0	0	00 10	0 0	00 10	1
0	0	1	0	1	01 11	0 1	10 11	0
0	0	1	1	0	00 10	1 1	01 01	0
0	0	1	1	1	01 11	1 1	11 11	1
0	1	0	0	0	00 10	1 1	01 01	1
0	1	0	0	1	01 11	0 1	00 11	0
0	1	0	1	0	00 10	1 1	01 01	0
0	1	0	1	1	01 11	1 1	11 11	1
0	1	1	0	0	00 10	1 1	01 01	0
0	1	1	0	1	01 11	1 1	11 11	1
0	1	1	1	0	00 10	1 1	01 01	1
0	1	1	1	1	01 11	1 0	11 10	0
1	0	0	0	0	00 10	0 1	00 01	1
1	0	0	0	1	01 11	0 1	10 11	0
1	0	0	1	0	00 10	1 1	01 01	0
1	0	0	1	1	01 11	1 1	11 11	1
1	0	1	0	0	00 10	1 1	01 01	0
1	0	1	0	1	01 11	1 1	11 11	1
1	0	1	1	0	00 10	0 0	10 00	1
1	0	1	1	1	01 11	1 0	11 10	0
1	1	0	0	0	00 10	1 1	01 01	0
1	1	0	0	1	01 11	1 1	11 11	1
1	1	0	1	0	00 10	1 0	01 00	1
1	1	0	1	1	01 11	1 0	11 10	0
1	1	1	0	0	00 10	1 1	01 01	1
1	1	1	0	1	01 11	1 0	11 10	0
1	1	1	1	0	00 10	0 0	10 00	0
1	1	1	1	1	01 11	0 0	00 10	1

The fast correlation attack algorithm for summation generator as the noise model based on the sequence of LFSR can be thought. We make it possible to execute the error correcting process for recompose the LFSR sequence base on the binary time-lag sequence observed and based on a parity check. The fast correlation attack algorithm[3] is shown to you as below.

① Calculate the binary time-lag sequence from the key stream observed.
② Calculate the parity check value for the each binary time-lag stream of z_t, t=1,2,3,...,t
③ Calculate the probability of error p_t by using the parity check values for each z_t.
④ If we see it as $p_t > 0.5$, we are to set it up $z_t' = z_t' \oplus 1$ and $p_j = 1 - p_j$ under the condition of t =1......k .
⑤ Repeat it until all the parity checks are satisfied.

When x is put as the total degree for the each LFSR feedback polynomial, the complexity for fast correlation attack and the requirement quantity of key sequence are as $o(2^{x/4})$ [3]. In the meantime, the fast correlation attack is applied as a model of repeated error correction algorithm for the key stream observed. But the repeated error correction algorithm requires the regular clocking of LFSR. There is also fast correlation attack that is not based on the repeated error correction algorithm, for example, the fast correlation attack of Jonsson and Johansson against LILI-128[7]. The method of attack is to predict the status of LILI-128 clock control register, and in that time it executes the fast correlation attack against irregular output stream on the data register in production. As the key stream generator of CCC-SG series is based on the mutual irregular clock of LFSR, not easy to separate the clock and data in production, the fast correlation attacks[7] mentioned are not to be applied.

3.2.3 Time/Memory/Data Tradeoff Attack

The purpose of Time/Memory/Data Tradeoff Attack[8] is to find out internal states within given time and the attack is dealt with two stages. During the first stage, cryptanalyzer draws a look-up table at the output key stream and related prefix of feasible internal states. In real attack stage, they try to find out the similar internal states by having a part of bit in the key stream known to the public through the check on the look-up table. S, M, T, P and D indicate the each inner status space size, the memory capacity (of the same size of binary work as $\log_2 S$), the calculating time (about the look-up table), the calculating time in advance (about the look-up table) and the length of data(without the renewal of key, that is, the length of data known to the the public.). Time/Memory/Data Tradeoff Attack[8] satisfies those of $T.M = s$, $P = M$ and $D = T$. Concerning CCC-SG with the internal states of 512 bit, T or M is shown up bigger than 2^{256} and this is more difficult than the total attack of key.

4 Conclusion

In this paper, we proposed CCC-SG to enhance the nonlinear property from key stream generator of the existing bluetooth encryption algorithm. The cryptographic strength of CCC-SG is 2^{256} and basic property of key stream against the various attacks is resulted as safe because of both big linear complexity and the long-term period, and we could know it that CCC-SG satisfied with the basic random test items. Therefore, we expect CCC-SG to enhance the security level of bluetooth technology occupying the standard of short distance wireless communication technology.

Acknowledgement

This research was supported by University IT Research Center Project of Korea, and by the Program for Training of Graduate Students in Regional Innovation.

References

[1] Dawson, E.: Cryptanalysis of Summation generator. In: Zheng, Y., Seberry, J. (eds.) AUSCRYPT 1992. LNCS, vol. 718, pp. 209–215. Springer, Heidelberg (1993)
[2] Rueppel, R.: Correlation Immunity and the Summation Generator. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 260–272. Springer, Heidelberg (1986)
[3] Golic, J., Salmasizadeh, M., Dawson, E.: Fast Correlation Attacks on the summation Generator. Journal of cryptology 13(2), 245–262 (2000)
[4] Meier, W., Staffelbach, O.: Correlation Properties of combiners with Memory in Stream Ciphers. In: Damgård, I.B. (ed.) EUROCRYPT 1990. LNCS, vol. 473, pp. 204–213. Springer, Heidelberg (1991)
[5] Specification on the Bluetooth System, version 1.1 February 22 (2001)
[6] Massey, J.: Shift-Register Synthesis and BCH Decoding. IEEE Transactions on Information Theory IT15(1), 122–127 (1969)
[7] Jonsson, F., Johansson, T.: A fast correlation attack on LILI-128. Information Processing Letters 80(3), 122–127 (2003)
[8] Babbage, S.: Improved exhaustive search attacks on stream cipher. In: European Convention on Security and Detection, IEEE Conference Publication, vol. 408, pp. 161–166 (1995)

Study of Optimal Traffic Information Using Agents Techniques

You-Sik Hong[1], CheonShik Kim[2], and Geuk Lee[3]

[1] Dept. of Computer Science, Sangji Univ., Usan-dong, Wonju-si, Korea
yshong@sangji.ac.kr
[2] Major in Digital Media Engineering Anyang Univ., Anyang-si, Gyeonggi-do, Korea
kimcsik@emapl.com
[3] Dept. of Information & Communication, Han-nam University, Daejon, Korea
leegeuk@hannam.ac.kr

Abstract. Recently in Korea, DMB TPEG service showing the optimal road while escaping the jammed way is popular. Such latest service also provides the traffic information after average 5 minutes of the recent traffic situation. Thus it cannot well reflect the present traffic information. Further, this service generates the problem that if it is away from the Metropolitan area, it cannot receive the traffic information to 100%. This thesis likes to provide the real-time traffic information as utilizing the intelligent agent to improve such problem and suggest the system which calculates optimal car speed in the real time as considering the conditions of intersection and of weather. The result of trial examination of system suggested in the thesis has proved that the system to forecast the traffic accident as utilizing the intelligent agent exactly calculates the traffic limit speed.

Keywords: Optimal car speed, Prevention of traffic accident.

1 Introduction

Most of drivers in Korea carefully listen to the traffic information on Radio broadcast to select the short way as grasping the traffic state to the destination. However, the traffic information on Radio concentrates in the information of downtown and main roads where the traffic jams are often occurred. It is difficult to sufficiently obtain the traffic information which drivers want. The alternative to settle such problems is customer real-time traffic information. Not only MBC IDIO providing the traffic information as utilizing FM Radio frequency but also 'Nate Driver' of SK telecom based on mobile telecommunication networks, etc are the representative example. Recently even TPEG of traffic information service based on DMB has been appeared to solve the concerns of drivers [1-3]. KBS, MBC, SBS and YTN, etc as DMB operators are providing TPEG service competitively. The broadcasting companies are unfolding the active business to compensate the limit of free DMB business with the chargeable TPEG as an added service. The remarkable number of DMB terminals recently offered in the market provides the real-time traffic information. But the customers who bought TPEG terminals while being exited with the dream that they

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 868–877, 2008.

could be free from the traffic jam might be disappointed around one time. There was often the case that although it is at the jammed section, it is openly signed as the green color and although it is at the smooth flowing section, it is shown as the red color. This thesis wants to settle such problems as utilizing Fuzzy Rule to resolve such problems [4-8]. Also improving the accuracy of traffic information is one of the purposes. The present traffic information of the Metropolitan area is emitted every 5 minutes as calculating the time that the probe cars prepared with sensor or GPS move the specific place of point. Although Lotis Co., Ltd. And SK Energy as the traffic information operators operate 15,000 and 11,000 probe cars respectively, the number of cars is too small to collect the road information of whole Metropolitan areas. Although the same traffic information is presented, it also is the problem that the results are of various kinds according to the peculiar map of companies and the algorithm of route recommendation [10-13]. The standards signing the road states such as the smooth, the jammed and the delayed, etc is different among companies. If we pass by the place of point although it is the section signed as the jammed, the traffic state is often smooth at the section [14-16]. There are many cases that the estrangement between the traffic information measuring the average speed of the specific section and the customer's perception of the state of the specific place of point is high.

This thesis thinks that the problem of such traffic information is due to the poor ness of the exact forecasting capability about the traffic states. Thus, we would like to settle the suggested problems as utilizing the Fuzzy algorithm and the intelligent agent techniques to forecast the more exact traffic states. The structure of this thesis is as the followings. Firstly, at the Chapter 2, the problems of existing traffic information are reviewed. At the Chapter 3, as utilizing the Fuzzy Rule, the intelligent car system automatically decreasing optimal car speeds to be suitable to the disaster broadcasting and the road environment, in the case that the traffic accidents are occurred or the obstacles are suddenly appeared while driving cars, is reviewed. At the Chapter 4, the simulation calculating optimal traffic speed as utilizing the intelligent technique is reviewed. And at the Chapter 5, the strong points and the weak points of traffic information system utilizing the intelligent agent are explained.

2 Problems of Existing Real-Time Traffic Information System

The largest lacking of drivers so far is the point that the various traffic information or instruments cannot provide 100% information that drivers seriously need. Although Radio broadcasting provides traffic broadcasting all day long, practically it does not provide the traffic information about where drivers pass by right now. Although the main roads are broadcasted repetitively in the several times a day, practically the traffic information drivers need is not very often like 'Murphy's law'. Although there are many drivers installing the navigation to resolve such weak points and easily find out the unfamiliar place to be there, navigation shows only the present location of car on the installed map data and does not show whether it is the jammed road or not.

Although if the services of mobile telecommunication companies are utilized, the real-time traffic information is available sometimes, it would be easy to have an accident if you continuously look at the small LCD of mobile phone while driving.

The service to resolve this kind of weak points in a time is just TPEG. TPEG (Transport Protocol Expert Group) is referred, easily speaking, as the real-time traffic information service based on navigation. As adding the real-time traffic information to ready-stored map information like existing navigation, it is the service making it possible to know the present road states or the speed information, etc from time to time. TPEG service is provided in the way how DMB operators has received the traffic information sent by Contents Provider (CP) and then shots it to user's terminal through broadcasting network. Despite of such strong points, there are the problems as the below.

(1) 5 minute time gap: Traffic status of Seoul down town in the rush hour used to be changed from the smooth traffic status into the jammed by 2~3 minutes. Under such traffic condition, most of traffic information is renewed every 10 minutes. As the results, traffic information confirmed on navigation screen may be the information prior to maximum 10 minutes and from that time gap, the traffic information shown on navigation may differ from the actual status. This thesis wants to provide optimal traffic information as utilizing the intelligent agent technique.

(2) Speed-limit of vehicle: The existing method has the danger that the traffic accidents are occurred because speed limit on signal board is fixed at the same 80Km/h even when it rains or snows and drivers cannot adjust to be suitable to the characteristic of road. This thesis assumes that as utilizing the intelligent agent technique, sensor for temperature and humidity on the roads perceives weather condition on roads and suggests the system to change speed limit into over 1/3 while adjusting the condition of raining or snowing.

(3) Showing the different traffic information from actual flowing status: Beacons as receivers are mostly installed at the intersection in case of downtown and Probes of transmitters are installed at the vehicles which are always being driven like taxies. It is previously reviewed that Information Collection Center grasps traffic information of section between A and B as collecting the times and the speeds which the car with probe passes by Beacon A of intersection and Beacon A of the next intersection. By the way, at the time being when the traffic is very small at the concerned region, if the empty taxies with probes are slowly driven to find out passengers on the not-blocked road, absurdly the concerned section is judged to be jammed in i-dio.

3 Proposed System of the Fuzzy Knowledge Rule Base

Generally, the fuzzy theory is used in the case that it is difficult to interpret in the quantitative method because the characteristic of system is complex or the information is qualitative and inaccurate. For example, in the case that there is the knowledge of expert about the system, but the mathematical analysis is difficult, fuzzy theory is used. Fuzzy logic makes input date into being fuzzy, derives the result through fuzzy inference and then shows in crisp value the result as making not-fuzzy. Fuzzy Rule is generally able to be showed in the form of IF-THEN and fuzzy inference is a series of process to infer the new relation or fact from some given rule.

If the system with input variable as 2 and output variable as 1 is composed as utilizing max-min inference, min operation of Mamdani may be described like the formula (1).

- Input : x is A' AND y is B'
- R^1 : IF x is A_1 AND y is B_1, THEN z is C_1
- OR R^2 : IF x is A_2 AND y is B_2, THEN z is C_2
- ....
- OR R^n : IF x is A_n AND y is B_n, THEN z is C_n
- Conclusion : z is C

$$R_C : \mu_C(z) = \bigvee_{i=1}^{n} (\alpha_i \wedge \mu_{Ci}(z)) \equiv \bigvee_{i=1}^{n} [\mu_{Ai}(x_0) \wedge \mu_{Bi}(y_0)] \wedge \mu_{ci}(z) \quad . \qquad (1)$$

Herein A', B' : Fuzzy set against input

A_i, B_i, C_i : Fuzzy set against x, y and z as variables of fuzzy rule
C : Fuzzy set against result
R_C : min operation rule of Mamdani
α_i : Firing strength against the i^{th} rule
$\mu_C, \mu_C, \mu_A, \mu_{B_i}$: Belonging function value against C, C_i, A_i, B_i of fuzzy set
x_0, y_0 : Actual input value of system
$\vee, \wedge$: OR (max operation), AND (min operation)
$\vee_{i=1}^{n}$: Composition of OR,
R^i : Number of fuzzy rule

And formula (1) can be expressed like the <Fig. 2>, <Fig. 2> is graphing the process to calculate the belonging function μ_C of concluding part as composing the values after calculating the belonging function value $\mu_{A_i}(x_0), \mu_B(y_0)$ against fuzzy set A_i, B_i of input value x_0, y_0 and then calculating $\mu_{Ci}(z)$ against fuzzy set C_i by min operation. A defuzzification is to convert the fuzzy quantity expressed as the belonging function into crisp value and although there is several methods, the centroid of gravity method is generally used.

$$z = \frac{\sum_{j=1}^{n} \mu_C(z_j) z_j}{\sum_{j=1}^{n} \mu_C(z_j)} \qquad (2)$$

Herein n : Quantization level of total output of power

z_j : Output of power according to quantization level j
$\mu_C(z_j)$: Belonging function value of z_j against fuzzy set C of concluding part

<Table 1> shows variable values for generating fuzzy set and the fuzzy sets in triangle and trapezoid were generated as using these values. And traffic safety speeds are classified into 3 grades such as 0-3(E grade), 4-6(D grade) and 7-9(C grade) based on the standard of 100 points. The below "If Then" rules show the whole rules of fuzzy system to obtain learning data.

It is very important that a disease is occurred at human body and to calculate optimal traffic safety speed according to the road condition. Says, as human takes health examination, the smoked intersection, the frozen section and the sharp curved road should be speed-down over 30-50% comparing to the regular speed. As utilizing agent technique, this thesis wants to settle these problems as utilizing fuzzy rule because the traffic accident may be easily occurred from a little bit over speed of driver at the intersection condition, the sharply curved section and the section where traffic sign cycle is not suitable to the actual.

Table 1. Data for determining safety speed as utilizing fuzzy variable

Grade	VS	S	M	L	VL
Sensor car speed(Km/H)	10-29 km/h	Below 30 km/h	30-50 km/h	50-80 km/h	Over 80 km/h
Number of intersection lane	2 lanes	3 lanes	4 lanes	5 lanes	6 lanes
Length of intersection	Below 30m	Below 50m	Below 70m	Below 100m	Over 100m
Rainfall (cm)	Below 5	Below 10	10~20	20~40	Over 40
Fog (%)	Below 5	Below 10	10~20	20~70	Over 70
Humidity (%)	Below 10	Below 20	20~40	40~70	Over 70
Snow (cm)	Below 5	Below 10	15~20	15~30	Over 30

Fig. 1 shows the chart of collection and analysis of traffic data. This knows traffic flowing of my destination as utilizing fuzzy algorithm after collecting traffic data and then provides the information to driver through DMB TPEG.

Fig. 2 is the system advising the danger to driver as utilizing the speed of driver's car and the state information of destination. Especially, it helps driver drive safely as providing the information on the heavy rain and the heavy snow of destination.

Fig. 3 shows 0.8-1.0 in the case of showing the jammed section with sensor data car speed of 21-30Km installed at 4 places of point like a, b, c and d and 0.4-0.7 in the case of 11-30Km. Lastly, in the case that car driving distance is under 10 Km as the delayed section, 0.1-0.3 is appeared. Herein ------- appears the speed forecasting judgment algorithm while considering the road conditions. Herein the digits indicated on the linking line means traffic accident and stoppage time of taxi.

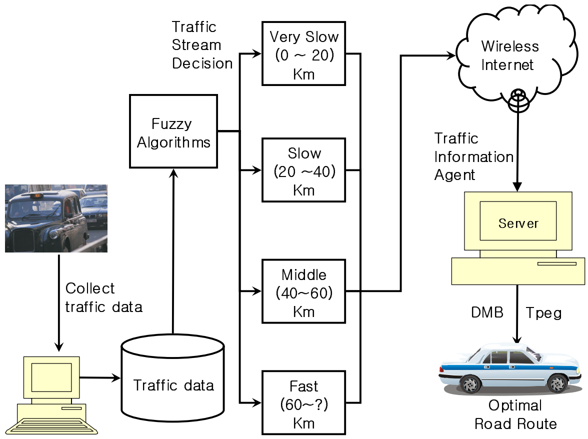


Fig. 1. Chart of collection and analysis of traffic data

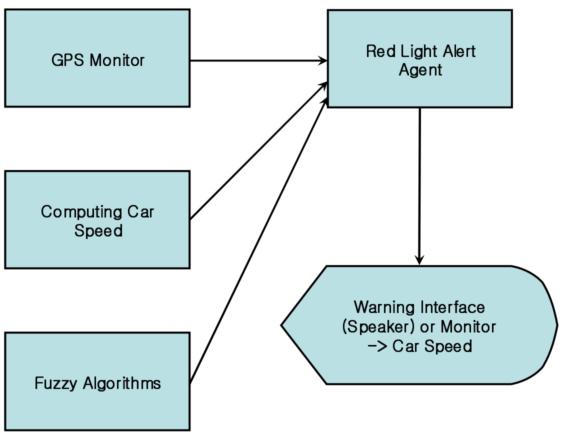


Fig. 2. Alert System: Architecture

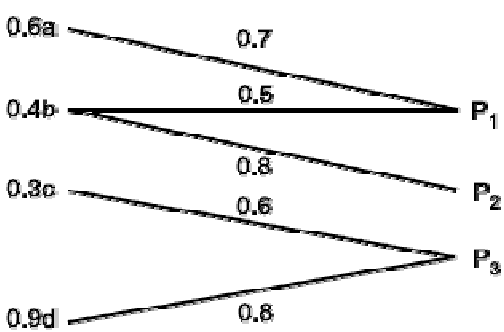


Fig. 3. Safety Speed Judgment Utilizing Fuzzy Rule

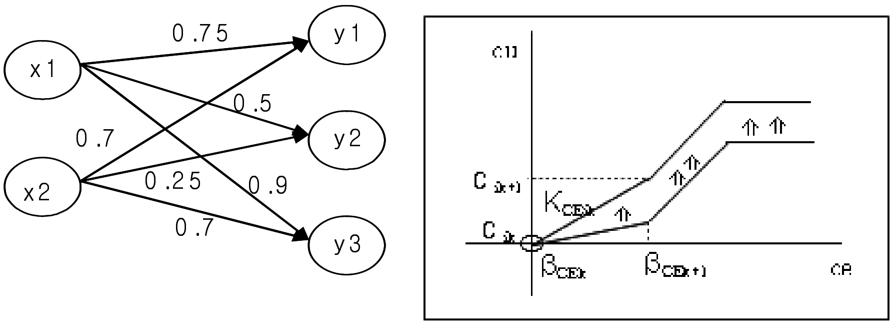


Fig. 4. Adaptation-type Fuzzy Rule considering intersection condition

Fig. 4 shows adaptation-type Fuzzy Rule considering weather condition, passage vehicle and intersection condition in the delayed and the jammed section.

(RULE 1) IF DPSV IS PB
AND USPC IS PB
THEN OPRG IS BIG
(RULE 2) IF DPSV IS PB
AND USPC IS NS
THEN OPRG IS MEDIUM
(RULE 3) IF DPSV IS NS
AND USPC IS NS
THEN OPRG IS SMALL

Herein,
DPSV: Sensor Data Condition (Car Speed)
USPC: Road Condition (Jammed, Delayed, Accident)
- Change Error (CE)
OPRG: Car Safety Speed considering road condition (10degrees)

Table 2. Fuzzy Rule for compensation input *v*

x_e \ y_e	**NB**	**NM**	**NS**	**ZO**	**PS**	**PM**	**PB**
NB	NB	NM	NM	NS	NS	NS	NS
NM				NM	NM		
NS			NM	NS			
ZO		NS	NS	ZO			
PS			ZO	PS	PM		
PM			PS	PS	PS		
PB	PS	PS	PS	PS	PM	PM	PB

4 Simulation and Analysis

This thesis assumes 80km/hr as safety speed on 4 lanes road and calculated optimal safety speed as considering the shape of intersection(uphill and downhill), the number of lanes(2 lanes, 4 lanes, 6 lanes and 8 lanes) and the weather conditions. As appeared on Table 1, the trial examination shows to reduce the safety speed by over 20% in the case of rain or snow as automatically detecting road states with humidity sensor and temperature sensor. Especially, the case of heavy rain or the process making it possible to prevent traffic accident as reducing speed to 40 km/hr as the safety speed by maximum 50% reduction of speed is shown. In addition, the process to reduce speed to safety speed of 56km/hr as reducing speed by 30% when it rains lightly, despite of the same rain is shown. The existing safety speed sign is 80 km/hr either in rain or snow but this thesis performed trial examination to calculate optimal safety speed as detecting weather condition.

Table 3. The Specification of WMR

Symbol	Numerical Value	Unit	Description
l_a	4-10	No	Number of Road Lane
l_b	20-500	***M***	Length of Road
H	0-1000	***Humidity %***	Section it is raining or snowing
F	0-100	***Fog %***	Foggy Section

To prevent traffic accident, the trial examination for optimal car speed reduction, the road states and the traffic flowing was performed. As appeared on Table 4, the section where traffic accident was occurred shows that 35 km/hr reduced by over 55% is calculate as the safety speed and the beginner driver does not regard road sign as the safety speed, but reduces speed in the real-time to prevent traffic accident. This thesis performed trial examination about the process to reduce safety speed as perceiving each road sates and traffic accident section with the standard of 4 lanes and 80km/hour. Table 4 shows the result of trial examination calculating optimal road sign speed as considering the weather conditions(sun, snow and rain), the shape of intersection(uphill and downhill) and the traffic volume(number of intersection lane).

Table 4. Reduced Car Speed Rate

Road Weather Conditions	Speed Reduction	Flow Rate Reduction
Dry	0%	0%
Rain	10%	6%
Wet & Snowing	13%	11%
Wet & Slushy	25%	18%
Wheel Path Slush	30%	18%
Snowy & Sticking	36%	20%

5 Conclusions

This thesis used Fuzzy algorithm to improve the existing system that provides real-time traffic information. The method utilizing probe car which is used for collecting traffic data at the present is still utilized despite of many weak points. This thesis has proved through experiment that traffic jammed section and smooth flowing section is apparently known as improving the problems of traffic data collecting method using probe car through fuzzy algorithms. And the existing method has the danger that traffic accident may be occurred because limit speed on sign board is fixed at the same 80Km/hr even when it snows or rains and then drivers cannot adjust in accordance with the characteristic of road. This thesis assumed that temperature and humidity sensor on road perceived weather conditions and performed trial examination about automatic speed limit system that changed speed limit into over 1/3 as adjusting speed limit to the weather conditions of rain or snow. The suggested intelligent agent technique in this thesis presented Fuzzy algorithm controller to determine traffic safety speed which is most suitable to road conditions and weather conditions. Input variable of fuzzy controller is speed error generated every sampling time. Output value of fuzzy controller is optimal speed considering weather conditions. I think it is very helpful to decrease traffic accidents if we perform the trial examination of traffic disaster broadcasting and automatic reducing speed equipment which automatically reduces speed as considering road conditions and weather conditions in the case that although driver is given the traffic states information , he does not reduce speed.

References

1. Lee, J.-H., Ham, Y.-K., Lee, S.-I.: Latest Trend in DMB Service in Korea. http://www.itfind.co.kr
2. ERTICO, DAB-based Multimedia ITS Applications, ERTICO stratety for implementation (October 1998)
3. Young-Han, Kim: Trend of terrestrial DMB service, University of Seoul (2003)
4. Youn-Hyung, Kim: Time-series analysis, academy (1990)
5. Ozawa, K., Niimura, T.: Fuzzy Time Series Model of Electric Power Consumption. IEEE Canadian Conference on electrical and Computer Engineering, 1195–1198 (1999)
6. Nie, J.: Nonlinear Time-Series Forecasting: A Fuzzy Neural Approach. Neuro computing 16, 63–76 (1997)
7. Lee, K.-H., Gil-Rok, O.: Fuzzy theory and application I. II. Hong-Lung Press (1991)
8. Chae, S., Young-Sok, O.: Fuzzy theory and control. Chong-Moon Press (1997)
9. Kim, I., Lee, S.-R.: A Fuzzy Time Series Prediction Method based on Consecutive Values. IEEE International Fuzzy Systems 2, 703–707 (1999)
10. Looney, C.G.: Pattern Recognition Using Neural Networks. Oxford University Press, Oxford (1997)
11. Moon, H.-D.: Telematics device is new technology incoming and market is opposition. ITFIND Mailzine 155, 9–17 (2004)
12. Bennet, J.A., Park, G.J.: Automotive Occupant DynamicsOptimization. DE- vol 32- 1 Advances in Design Automation - vol.1 ASME (1991)

13. Hou, J., Tomas, J.: Optimization of Driver - Side Airbag and Restraint Sy stem by Occupant Dynamics Simulation, SAE 952703 (1995)
14. Niii, N.: Structural Crashworthiness Optimization of Large Vehicles using Detailed models. In: EuroPAM 1998 (1998)
15. MBC research technology, Transport Protocol Experts Group(TPEG) Specifications (August 2003)
16. MBC technology research, TPEG history and trend in korea (August 2003)
17. Keti, Trend of DMB technology and TPEG (August 2003)

A Recommendation Using Item Quality*

Sung-hoon Cho[1], Moo-hun Lee[1], Bong-Hoi Kim[2], and Eui-in Choi[1]

[1] Dept. of Computer Engineering, Hannam University
Daedeok-gu, Daejeon, 306-791, Korea
{shcho, mhlee, eichoi}@dblab.hannam.ac.kr
[2] UBNC Co., Ltd.,
Yuseong-gu, Daejeon, 305-732, Korea
kbh@ubnc.net

Abstract. Today there are a lot of recommender systems operating on the web. These systems use content-based filtering or collaborative filtering or hybrid approach that was studied before. These techniques operate recommendation by using features of user and item, similarity of users, and items. Even though there is a consideration of attributes of items and users, but there is not much consideration of the quality of items. This is why item quality is not easy to be measured. This paper computes item quality, suggests it to apply to the recommender system, and presents it by analyzing the influence.

1 Introduction

A recommendation system judges what a customer asks, looks for items according to that, and operates the role that provides the items that optimize the customer's request. A research to operate this work exactly and effectively has been doing since the mid of 1990's. A technology research to apply to the recommendation system started from a collaborative filtering[1, 2, 3]. This collaborative filtering finds the neighbor that has a similar preference and is used to predict the preference of item which evaluated the preference of customers. This technique is rather based on the user.

There is another technique called a content-based filtering which is used in the recommendation system with collaborative filtering. This technique focuses on finding out the items including textual information on the based of user's preference and recommending.

What these techniques should consider to recommend are user's features, features of item, the similarity of users, rating of item by user. However, these hardly consider the quality of the item itself. User's rating can be confusing with quality. User's rating is not the same as quality because user's rating means the preference of the item. If quality of item itself can be offered as a standard of recommendation, it can come closely to the users' requests.

* This research was financially supported by the Ministry of Commerce, Industry and Energy(MOCIE) and Korea Industrial Technology Foundation(KOTEF) through the Human Resource Training Project for Regional Innovation.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 878–886, 2008.

The biggest reason why item quality could not be applied to established techniques, because it asks for the user's feedback to measure the quality. Actually, it is not possible to ask every user who uses the recommendation system for an estimation value. Even if that kind of system could be operated, users would give up using the recommendation system in order to avoid the trouble.

Among the problems of information retrieval field, there is a rich-get-richer problem. This problem is that the item with higher rank will be offered as a searching result when searching for information, while a new item is rarely included. This also can be happened in the recommendation system. If we apply this problem to each filtering technique, the result in the collaborative filtering will be followed by item rank according to users' similarity. But in the case that a new item is added, the recommendation rank for it will be very low, so the item is hard to be a recommendation candidate. The content-based filtering returns the optimum item as a recommendation candidate for the user's preference. Then, content-based filtering is compared to textual information of the item, so it should have less problem of rich-get-richer, but it is difficult to satisfy user's requests by only rating of new item. After all, it has a problem that the precision in recommendation is low.

Therefore, in this paper, an item quality technique is suggested as a technique that well satisfies rich-get-richer and user's requests and also a measurement method for item quality is presented without explicit user feedback. This paper consists; in chapter 2, there is a brief about established recommendation techniques, in chapter 3, it suggests the concept of item quality, in chapter 4, it talks about the technique that measures the item quality, in chapter 5, it evaluates the effect that item quality suggested affects the recommendation, and finally it meets to conclusion in chapter 6.

2 Related Works

2.1 Content-Based Recommendation

Many current content-based systems focus on recommending items containing textual information, such as documents and Web sites URLs[4]. For generating recommendation about textual information, Term Frequency/Inverse Document Frequency(TF-IDF) measurement was proposed to estimate the weight of keywords[5]. It is based on the frequency of keywords. TF-IDF measure is defined as follows: Assume that N is the total number of documents that can be recommended to users and that keyword k_j appears in n_i of them. Moreover, assume that $f_{j,j}$ is the number of times keyword k_j appears in document d_j. Then, $TF_{i,j}$, the term frequency (or normalized frequency) of keyword k_i in document d_j, is defined as

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}$$

where the maximum is computed over the frequencies $f_{z,i}$ of all keywords k_z that appear in the document d_j. However, keywords that appear in many documents are not useful in distinguishing between a relevant document and a nonrelevant one. Therefore, the measure of inverse document frequency IDF_i is often used in combination with simple term frequency $TF_{i,j}$.

$$IDF_i = \log \frac{N}{n_i}$$

However Content-based techniques are limited by the features that are explicitly associated with the objects that these systems recommend[4]. Therefore, in order to have a sufficient set of features, the content must either be in a form that can be parsed automatically by a computer or the features should be assigned to items manually.

2.2 Collaborative Recommendation

Collaborative filtering systems try to predict the utility of items for a particular user based on the items previously rated by other users[3]. The utility u(c,s) of item s for user c is estimated based on the utilities u(c_j,s) assigned to item s by those users($c_j \in$ C) who are "similar" to user c. For example, in a movie recommendation application, in order to recommend movies to user c, the collaborative recommender system tries to find other users that have similar tastes in movies. Then, only the movies that are most liked by the "peers" of user c would be recommended[6, 7].

Collaborative recommendations can be grouped into two general classes: memory-based and model-based. Memory-based algorithms essentially are heuristics that make rating predictions based on the entire collection of previously rated items by the users. That is, the value of the unknown rating $r_{c,s}$ for user c and item s is usually computed as an aggregate of the ratings of some other users for the same item s.

Various approaches have been used to compute the similarity *sim(c, c')* between users in collaborative recommender systems. In most of these approaches, the similarity between two users is based on their ratings of items that both users have rated. The two most popular approaches are correlation and cosine-based. To present them, let $S_{x,y}$ be the set of all items co-rated by both users *x* and *y*.

$S_{x,y}$ is used mainly as an intermediate result for calculating the "nearest neighbors" of user x and is often computed in a straightforward manner, i.e., by computing the intersection of sets S_x and S_y. In the correlation-based approach, the Pearson correlation coefficient is used to measure the similarity:

$$sim(x, y) = \frac{\sum_{s \in Sxy} (r_{x,s} - \overline{r_x})(r_{y,s} - \overline{r_y})}{\sqrt{\sum_{s \in Sxy} (r_{x,s} - \overline{r_x})^2 (r_{y,s} - \overline{r_y})^2}}$$

In the cosine-based approach, the two users x and y are treated as two vectors in m-dimensional space, where m=| $S_{x,y}$ |[7]. Then, the similarity between two vectors can be measured by computing the cosine of the angle between them.

$$sim(x, y) = \frac{\sum_{s \in Sxy} r_{x,s} r_{y,s}}{\sqrt{\sum_{s \in Sxy} r^2_{x,s}} \sqrt{\sum_{s \in Sxy} r^2_{y,s}}}$$

The mean squared difference measure can also measure the similarities between users [3]. Different recommender systems may take different approaches in order to implement the user similarity calculations and rating estimations as efficiently as possible. If the recommendations are needed, the ratings are calculated efficiently through the pre-computed similarities.

Model-based algorithms use the collection of ratings to learn a model, which is then used to make rating predictions[3, 7, 8, 9]. For example, [7] proposes a probabilistic approach to collaborative filtering, where the unknown ratings are calculated as and it is assumed that rating values are integers between 0 and n and the probability expression is the probability that user *c* will give a particular rating to item s given that user's ratings of the previously rated items.

$$r_{c,s} = \sum_{i=0}^{n} i \times \Pr(r_{c,s} = i \mid r_{c,s'}, s' \in S_c)$$

However collaborative filtering approaches have several problems that is associated with new user, new item and sparsity[5].

3 The Concept of Item Quality

In this chapter, the definition of item quality and related concepts to derive item quality are discussed.

3.1 Popularity

Popularity is defined and the definition is discussed in this section.

The recommendation makes candidates through collaborative filtering or content-based filtering or hybrid approach which takes advantages of two techniques. A user chooses items from the recommendation candidates as many as he/she wants to. Let's suppose that when there is the same user's queries, if item a, b, c are given as candidates and the ration of item chosen by users is $a > b > c$, normally we can say that *a* is the most popular item. The more choosing, the more popular the item has. So popularity can be defined as definition 1.

Definition 1. *Popularity*
Popularity is the proportion of $P(i, t)$, that the users' proportion who like item *i* for a given time(t).

For instance, if there are a million of recommendations operated and item *i* is chosen by 100,000 people among them, its popularity is 0.1. Popularity can be confused with item quality. Item quality popularity is the ration that a user could like the item and item quality popularity is the ration of users who like the item.
In this paper, I will evaluate item quality by using this popularity.

3.2 User Awareness

In this section, user awareness is defined and the definition is discussed.

Generally users look for popular items and choose. Popular items could be chosen because they are hot or the have a good quality or meet people's taste. Like that, popular items make a lot of users choose them. So it is clear that popular items can be spread widely and fast to among users. By this fact, user awareness can be defined as definition 2.

Definition 2. *User Awareness*
User awareness is the proportion of *A(i, t)*, that the users' proportion who are aware of item *i* for a given time *t*.

User awareness is that how many people know the item. For example, if 100,000 of users out of a million are recommended for item i, user awareness of the item, *A(i, t)* is 0.1. Here user awareness does not mind whether users like the item or not. It means only the ration of recognition of the item.

The precise meaning of user awareness mentioned in this paper is user awareness increasing per unit time. It is about how much user awareness has been known for a unit time rather than accumulated awareness of the item.

3.3 Item Quality and Requirements

In the legacy techniques, a group of features of item is used to operate recommendation as an evaluation factor. I early said that quality of the item is not considered in general. But, I also said that item quality would be used in this paper. So, I would like to say what item quality is in this section.

The main idea of item quality is based on that (1) if quality of item is high, there will be a lot of users to use it, and (2) that item will be recognized among users rapidly. Therefore, popularity of item will increase more rapidly.
Hence, item quality can be defined as definition 3.

Definition 3. *Item Quality*
Item quality is a probability, *Q(i)*, which a user might like the item *i* which is given.

As mentioned item quality, there is a relation among quality, popularity and user awareness as below.

$$P(i,t) = Q(i) \cdot A(i,t) \tag{1}$$

The relation above shows that it will be the popularity of item if *A(i, t)* that recognizes time t and p multiplies item quality *Q(i)*. The relation above makes clear that the item will be preferred among every user and item quality is a probability of the item preference of a user. Through each concept which has this relation, item quality will be computed.

However, it is impossible to get item quality through explicit user feedback. So, we need another way to get item quality.

We use the fact that when item quality gets high, a probability to choose the item will get high too. Let me turn over the idea. If the item is chosen frequently by users, it can be said that the quality of the item is high. Like this, achievement of item quality can be gotten by implicit feedback which is a user's choice of item. And it enables to evaluate how many people like the item. Also if an item with high item quality is added to the recommendation system, its popularity will be increased fast compared to other items. Because of that most users who are recommended the item will like it. Therefore, as we observe the increase of popularity or time derivative, we will enable to evaluate quality of item. You should bear in mind that time derivative of popularity could be computed easily in relative. For instance, you can compute how many times the choice of the item can be changed for a given time if the chosen times of recommended item through the recommendation system could be evaluated by time derivative.

As suggested in definition 3, you should know user awareness apart from popularity in order to compute item quality. For user awareness, you need not know how many users prefer the item or how many users do not prefer the item. Simply, you need to know how many users know the item.

Through (1), the relation between item quality and user awareness can be shown as follows.

$$A(i,t) = \frac{P(i,t)}{Q(i)} \tag{2}$$

The equation above does not mean that user awareness is in inverse proportion to item quality, but a probability that how much each user knows item i can be computed with popularity and item quality. However, we in this paper have to find item quality and apply it to the recommendation system, so it is impossible to get user awareness with item quality in real. We would like to talk about how to get user awareness later.

Another fact to know through (1) and (2) is that item quality is a relative value, not an absolute value. Therefore, if popularity or user awareness happens to change, item quality can be changed relative to this.

The change of popularity and user awareness has something in common with fashion. In real world, a popular item has high popularity and user awareness. While an item which is out of date or not in fashion has low popularity and user awareness.

Due to these points, item quality reacts fashion sensibly and can be applied to item recommendation.

In order to evaluate item quality, you should first know how to get popularity and user awareness. So we would like to talk about how to compute popularity and user awareness in the next chapter.

4 Item Quality Measurement

In the chapter 3, we defined item quality and related concepts and discussed relations among them. Here we would like to talk about the computation of popularity and user awareness which are needed to get item quality and explain how to get item quality.

4.1 Measuring Popularity

Popularity in the recommendation system is based on user's selection of items as mentioned in 3.1. The more selected the item was, the higher popularity it has. Therefore, popularity can be got rather easily.

Popularity in 3.1 is *R(i, t)*, the recommended numbers for item *i* for time t, and 100,000 is *S(i, t),* the number of users, who selected (or preferred) *i*, among them. So popularity shows the preference of the item how much or how long users like it. If I make it as an equation, it will be as follows.

$$P(i,t) = \frac{S(i,t)}{R(i,t)} \tag{3}$$

In order to get *S(i, t)*, you have to monitor user's behavior in the recommendation system. So, you need to have another module which can measure item choosing behavior among one's behavior.

4.2 Measuring User Awareness

Let me talk about how to get *A(i, t)* in this section.

It is not easy to know how many users recognize an item in a certain time in the recommendation system. As it is hard for user feedback to be applied to item quality, it cannot apply user feedback to user awareness. Also you cannot use a way that measures user awareness with the purchase of the item. Because when an item can be purchased, the item has been already known and that's why there is a trade.

Then how can we know user awareness from the recommendation system? I would like to use a recommendation list of item in order to solve this problem. The item which has been recommended could have more possibility to be recognized by users than one which has not been. Also the higher rank of item is could be recognized better than the lower rank of one.

This kind of behavior pattern of users is directly related to the rank of item recommended. So, it should be employed as an element of awareness analysis. It was mentioned earlier that user awareness shows how widely well-known the item is. Hence you should know user awareness of by the time the recommendation has done in order to get it. The equation for it is as follows.

$$A(i,t) = \frac{CP_{i,t} \cdot RW_{i,t}}{RC_t} \tag{4}$$

RC_t is the count of recommendation processes operated so far, CP_i is the number of i included to the recommendation system, $RW_{i,t}$ is the rank ration of i in the recommendation list of whole users that was recommended for t and t-1.

A(i, t) gotten from the equation above is a measure that how much the item from recommendations till the time t has been known for compared to other items.

Once previous user awareness has been gotten, for the recommendations made after *t*, we apply (3) to the recommendations and evaluate user awareness. The equation to get user awareness after t is as follows.

$$A(i,T) = \frac{\int_t^T CP_{i,T} dt \cdot \int_t^T RW_{i,T} dt}{RC_T} \tag{5}$$

If *A(i, T)* divides into *A(i, t)* of before t, increasing ration of user awareness will be gotten.

4.3 Measuring Item Quality

Quality is the most important attribute among information of item. When an item is shown, the quality must be considered to make a decision whether the item could be chosen or not.

Early in the discussion of item quality and requirement in 3.3, we made sure for the relation of (1). Through this relation, we can draw an equation to get item quality. The deriving equation is as follows.

$$Q(i) = \frac{P(i,t)}{A(i,t)} \tag{6}$$

The equation (6) shows that the item quality is computed by popularity and user awareness which increased for time t. The item quality gotten from this equation will be a probability that a user might like the item when he/she is aware of it. We drew the equations to get popularity and user awareness in 4.1 and 4.2. Item quality can be gotten by applying these equations to (6).

5 The Evaluation of Effect of Item Quality

We would like to present some various examples that item quality may affect real recommendation. In the table below, *count* is the total number of recommendations for a certain time, *cont* is the number of item included in recommendations, *sel* is the number that users choose *i* among recommendation, *pop* is the popularity of *i*, *i-rank* is the relative recommendation rank ration of *i*, *UA* is user awareness, *IQ* is item quality. Let's suppose users 10,000.

Table 1 shows that if there is an increase of the user choice of i, popularity gets high, and item quality also increase up to that.

Table 1. Change of Item Quality according to Increase of Popularity

count	*cont*	*sel*	*pop*	*i-rank*	*UA*	*IQ*
1000	10	1	0.1	0.1	0.001	0.01
1000	10	2	0.2	0.1	0.001	0.02
1000	10	3	0.3	0.1	0.001	0.03
1000	10	4	0.4	0.1	0.001	0.04
1000	10	5	0.5	0.1	0.001	0.05

Table 2 shows that when there are frequent recommendations of i, user awareness increases, while entire item quality gets low as popularity lowers.

Table 2. Change of Item Quality according to Recommendation Ratio

count	*cont*	*sel*	*pop*	*i-rank*	*UA*	*IQ*
1000	10	1	0.1	0.1	0.001	0.01
1000	12	1	0.0833	0.1	0.0012	0.0069
1000	14	1	0.0714	0.1	0.0014	0.0051
1000	16	1	0.0625	0.1	0.0016	0.0039
1000	18	1	0.0556	0.1	0.0018	0.0031

What we can know from two tables above is that if we apply user awareness, popularity, and the number of recommendations to content-based filtering, it will act as a weighted value on the item.

6 Conclusions

In this paper we have proposed an application of item quality as a weighted value to when there happens to rank in the content-based filtering. This concept can apply the features that item quality changes according to relative popularity and user awareness of the item, so it makes a recommendation which suits for fashion possible. There is also an advantage that a probability in satisfaction could get higher when a user chooses the item because relative quality of the item is applied to the recommendation.

The experiment of this work presents that item quality can affect the generation of the recommended candidates. And we would like to work that apply the proposed recommendation for personalized recommendation.

References

1. Resnick, P., Iakovou, N., Sushak, M., Bergstrom, P., Riedl, J.: GroupLens: An open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 Computer Supported Cooperative Work Conference (1994)
2. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Application of Dimensionality Reduction in Recommendation System-A Case Study, ACM WebKDD 2000 Web Mining for E-Commerce Workshop (2000), http://robotics.stanford.edu/~ronnyk/WEBKDD2000/papers/
3. Shardanand, U., Maes, P.: Social Information Filtering: Algorithms for Automating Word of Mouth. In: Proceeding of CHI 1995 Conference on Human Factors in Computing Systems (1995)
4. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. Knowledge and Data Engineering, IEEE Transactions on 17(6), 734–749 (2005)
5. Salton, G.: Automatic Text Processing. Addison-Wesley, Reading (1989)
6. Nakamura, A., Abe, N.: Collaborative Filtering Using Weighted Majority Prediction Algorithms. In: Proc. 15th Int'l Conf.Machine Learning (1998)
7. Delgado, J., Ishii, N.: Memory-Based Weighted-Majority Prediction for Recommender Systems. In: Proc. ACM SIGIR 1999 Workshop Recommender Systems: Algorithms and Evaluation (1999)
8. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval J 4(2), 133–151 (2001)
9. Getoor, L., Sahami, M.: Using Probabilistic Relational Models for Collaborative Filtering. In: Masand, B., Spiliopoulou, M. (eds.) WebKDD 1999. LNCS (LNAI), vol. 1836, Springer, Heidelberg (2000)

Secure Storage and Communication in J2ME Based Lightweight Multi-Agent Systems

Syed Muhammad Ali Shah[1], Naseer Gul[1],
Hafiz Farooq Ahmad[2], and Rami Bahsoon[3]

[1] International Islamic University, Sector H-10 Islamabad, Pakistan
[2] NUST Institute of Information Technology, 166-A, Street 9, Chaklala Scheme 3, Rawalpindi, Pakistan
[3] School of Computer Science, University of Birmingham, Edgbaston, Birmingham B15 2TT England, UK
alishah_ph@yahoo.com, gul_192@hotmail.com,
hfarooqahmad@yahoo.com, r.bahsoon@cs.bham.ac.uk

Abstract. Securing data and applications is critical in the wireless era. The successive increase in wireless communication opens vulnerabilities to confidential information. This paper aims at investigating the issues related to security when the Java 2 Micro Edition (J2ME) based multi-agent system applications are run on different handheld devices. Agent's information must be protected on the local device as well as on other devices during communication. J2ME MIDP security model does not provide secure data storage and communication mechanism for applications running on different mobile devices. This paper proposes a mechanism for secure storage of agent's information and also securing communication between agents in J2ME based lightweight multi-agent system. We present a new solution to this problem by providing a security framework that can be used to provide agent's secure storage and communication on the application layer by using the existing cryptographic techniques. Storing the encrypted data on local device helps securing the agent's information. In this approach for secure agent's messaging the encrypted data is communicated with the help of session keys exchange, so no need to encrypt the data at runtime for communication as only the keys are encrypted. This results in the better security level and minimal performance overhead. The solution has been evaluated on the SAGE-Lite (lightweight multi-agent system) in order to cope with the security drawbacks related to agent's secure storage and communication.

Keywords: Lightweight multi-agents system, J2ME storage, agent security, agent communication, SAGE-Lite.

1 Introduction

Designing and implementing secure mobile applications of multi-agent systems is becoming a concern as the commercial usage of m-commerce is growing. It is essential to apply security on the multi-agent systems because of their huge predicted

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 887–896, 2008.

commercial usage. Java 2 Micro Edition (J2ME) [1] provides an environment for developing mobile applications, where portability of Java code is possible. One of J2ME features is that it provides built in caching mechanisms for locally storing data on a mobile device. Another feature of J2ME is that it can allow the client to set its security policies and encryption mechanisms based on the client's preference, content, and the sensitivity of data [2].

Applications built on top of J2ME are often vulnerable to security attacks. Such is a drawback in the environment, as described in the J2ME Mobile Information Device Profile 2.0 security model [3]. Moreover, strengthening security is often a concern in mobile environments for example; the e-commerce or m-commerce applications of multi-agent systems often necessitate special security requirements due to the nature of the transactions carried on such an environment. In such an environment, clients communicate by exchanging information, where the generated data are stored to databases. Securing the data locally on the device as well as during the communication with other devices using Bluetooth or WiFi (Wireless Fidelity) are the major problems faced by the J2ME developers.

The Mobile Information Device Profile profiles (MIDP) add APIs (Application Programming Interface) for user interaction, network connectivity, and persistent storage. Two profiles have been developed for J2ME: MIDP 1.0 and 2.0 [1]. J2ME MIDP 1.0 and 2.0 does not provide the secure storage as the data is stored in the database without encryption [3]. Hence, confidential information can be vulnerable to any attack.

In order to overcome the above mentioned problems, this paper provides a solution that makes agent's information and communication with other agents secure on one or more platforms. This can be achieved by using the existing java components with minimal affect on the system's performance. A lightweight multi-agents system SAGE-Lite (Scalable fault tolerant Agent Grooming Environment - Lite) [4] is used to illustrate the implementation of the proposed solution.

Rest of the paper is organized as follows: section 2 describes the motivation of work. Section 3 describes the architecture of the SAGE-Lite. Section 4 gives the proposed architecture of Secure SAGE-Lite. Section 5 explains the evaluation criteria. Section 6 includes the conclusion and Section 7 hints on the future work.

2 Motivation

The advancement in wireless technology has driven the trend towards ubiquitous computing (i.e. anywhere anytime), where modern mobile devices tend to enable such a computing model. Mobile devices are generally small and are constrained with low computing power, small memory and limited user interface. Java 2 Micro Edition (J2ME) is a popular technology for developing the applications for mobile devices, providing the benefits of Java technology suiting the requirements of small devices like constrained resources including small memory, display size, processing power, network bandwidth and battery life [1].

Security is an important concern in the ubiquitous computing. This is because of the ad hoc nature of communication in mobile setting. Intruders can join the mobile network much easier than a fixed one. Realizing security in the applications that are developed for small devices poses some constraints on the security architecture. Jonathan Knudsen [5] provides an overview of secure system design and describes how applying cryptography helps to design of secure applications.

Developing secure J2ME applications is difficult due to limitations of the underlying J2ME security model (and sometimes due to faults in manufacturer design). Portability concerns are common as the J2ME application need to run on "lightweight" devices, with different characteristics, and portability requirements. Security related functionality is often affected with the portability by J2ME applications. Several security frameworks have been proposed: Mobile Information Device Profile (MIDP 2.0) security framework [6] includes several mechanisms to ensure security of the applications and communication channels by using application signing and HTTPS connections.

Kayssi et al. [2] uses pure Java components for providing end-to-end client authentication and data confidentiality between Wireless J2ME based clients and J2EE based servers by using the Advanced Encryption Standard (AES) algorithm Rijndael. In the current challenges in securing networked J2ME applications [3], we can find that security architecture crumbles when we try to install the same application with security related components on different handheld devices, e.g. Public-Key Infrastructure (PKI)-based signature verification, an optional feature in the MIDP 2.0 specification, allows devices to verify a signed J2ME application's source and integrity. Signed applications are generally desirable for m-commerce and mobile government services, but when users install a signed application on devices that do not verify signatures, part of the security architecture disintegrates.

Existing techniques for securing the applications have also their own limitations. For example, the use of traditional public key encryption at the application layer has proved to be unsuitable for the case of lightweight devices [7], as a significant amount of processing power and resources are required. Barka et al. [8] present a lightweight application level security solution for handheld devices in wireless LAN. They have compared the performance impact of two private key security algorithms RC4 & AES on small devices and it was determined that both RC4 and AES performed similarly in different experimental cases. On the other hand public key algorithms were found to be more computational expensive. SSL/TLS on transport layer can provide End-to-End security, but are considered to be too heavy for the wireless environment.

Another used technique is application signing [3]. Signing is used to obtain the authenticity and integrity and widely used in m-commerce and m-government services. The limitation of using application signing, however, is that when users try to install these applications on the devices that don't support signature verification, security architecture could collapse e.g. this was tested on a newly released Samsung mobile device, which refused to install J2ME applications signed with the private key, belonging to a code signing certificate obtained from Verisign. However, the Nokia devices validated the certificate and applications without problems. These resource constrained devices limit the application developer to think individually for a small range of devices. There is pressing need to provide a solution, which can be generalized to the vast range of devices.

We propose security in lightweight multi-agent systems based on J2ME that tends to build on previous work while addressing much of the described limitations. Our proposed design aim at providing a highly secure environment for agent's storage and communication in the handheld devices, that is simple to deploy without degrading the system's performance.

We use the bouncy castle's java implementation [9] of the Advanced Encryption Standard (AES) Rijndael algorithm [10] for encrypting/decrypting the data. The advantage of using Private Key algorithms is that they are faster and less computationally expensive than public key algorithms making them more suitable for wireless environment. AES is an encryption algorithm for securing sensitive data and communication. It is a symmetric 128 bit block data encryption technique. The algorithm consists of four stages that make up a round, which is iterated 10 times for a 128 bit length key, 12 times for 192 bit key, and 14 times for a 256 bit key. The first stage "Sub Bytes" transformation is a non-linear byte substitution for each byte of the block. The second stage "Shift Rows" transformation cyclically shifts (permits) the bytes within the block. The third stage "Mix Columns" transformation groups four bytes together forming four turn polynomials and multiples the polynomials with a fixed polynomial mod (X^4+1). The fourth stage "Add Round Key" transformation adds the round key with the block of data.

Due to the speed and memory limitations of the J2ME devices we used the block size of 16 bytes processed with 128-bit keys for encryption / decryption purpose, as this combination best suite the J2ME MIDP applications [2]. Although storing the data encrypted requires more memory space and computing power, but it is assumed that the upcoming handheld devices will have sufficient memory and computing power to be suitable for this purpose.

The requirements for the development of Secure SAGE-Lite are:

a. Providing the secure storage mechanism for the agents residing on the platform.
b. Providing the secure communication between agents on different handheld devices using Bluetooth.

3 Architecture of SAGE-Lite

Scalable fault tolerant Grooming Environment (SAGE) is an open source FIPA (Foundation for Intelligent Physical Agents) compliant 2nd generation multi agent system [11]. Foundation for Intelligent Physical Agents or simply FIPA is one of the standard governing bodies [12], which provides an abstract architecture for lightweight multi-agent system developers to follow. SAGE is a distributed decentralized architecture. Due to the decentralization factor system is more much scalable.

SAGE-Lite [4] is the evolution of SAGE (Scalable Fault-Tolerant Grooming Environment) agent platform from which it inherits the lightweight behavior. SAGE-Lite is a lightweight context aware multi-agent system, which senses the capabilities of the lightweight devices and reacts accordingly. It enhances the capabilities of existing system SAGE and makes it compatible with a vast variety of currently available lightweight devices through its intelligent agents. This framework allows implementing agent-based applications like business applications or e-commerce applications on these resource constrained devices.

SAGE-Lite architecture contains three major system modules that are:

- Light-weight Agent Management System (AMS-Lite)
    - Light-weight Directory Facilitator (DF-Lite)
- Light-weight Message Transport service (MTS-Lite)
- Light-weight Agent Communication Language (ACL-Lite)

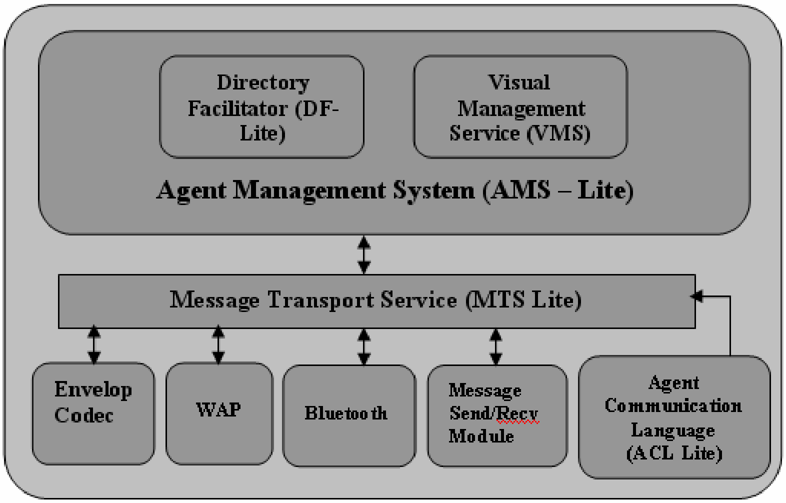


Fig. 1. System Architecture of SAGE-Lite [4]

Agent Management System (AMS-Lite) is a compulsory component of any Agent Platform. This component is responsible for all activities regarding control of Agent Platform. Each Agent Platform has only one AMS and soon after the creation of agent, the agent should be registered with AMS.

Directory Facilitator (DF-Lite) is an optional component of Agent Platform. It provides yellow page services to other agents. All of the services offered by different agents are registered with DF-Lite. Any agent can ask about required service from DF-Lite. In SAGE-Lite, DF-Lite works as a service rather than an agent due to the memory, power and processing constraints in handheld devices.

All agents on the platform communicate with each other through Agent Communication Language (ACL). Visual Management System (VMS) Lite provides the Graphical User Interface to the system. Message Transport Service (MTS) is used for communication through WAP and Bluetooth. Envelop Codec performs the encoding of envelop in Bit Efficient encoding representation.

4 Architecture of Secure SAGE-Lite

The problems faced in implementing the security for lightweight multi-agents system were related to the secure storage of agents and services along with the secure communication between the agents on same and different platforms. Our proposed architecture is designed in a way that implements the security in application with minimal affect on the performance of the system.

We have used two mechanisms to secure the key for encryption / decryption on the client: firstly stored the Keys class in the JAR file and secondly, obfuscated the code to prevent disassembly of code. The obfuscator also reduces the size of MIDLET suite and renames some of the Bouncy Castle classes used in the system. The use of obfuscator is very important regarding MIDLET size, as limited memory is assigned to the java applications on small devices.

Figure 2 shows the high-level architecture of Secure SAGE-Lite with the security implemented that allows the agents to communicate securely with other agents on the same and different platforms.

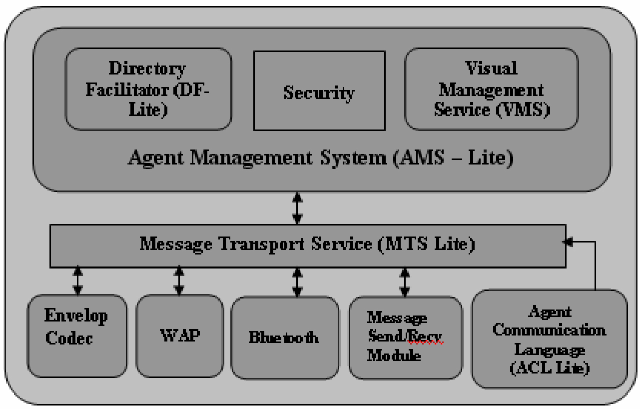


Fig. 2. Architecture of Secure SAGE – Lite

4.1 Secure Agent's Storage

In order to store the platform agent's information securely on the local device, we used the mechanism of encryption / decryption, which is not present in J2ME MIDP2.0. We have used the bouncy castle's java implementation of the AES Rijndael Algorithm with the block size of 16 bytes processed with 128-bit keys for encryption / decryption [9]. The reason for choosing this combination for encryption was due to the speed and memory limitations of small devices. This combination proves to be the best on MIDP applications. The data is stored encrypted in the Record Management System (RMS) of J2ME with minimal overhead, thus saving the loss of storage due to encryption. It is assumed that the minimal computation and memory overhead of storing data encrypted will be overcome by the upcoming handheld devices having much higher memory and computation power.

The key to encrypt and decrypt the data is generated on the basis of fixed device IP (8 bytes) IMEI number and user pin code (8 bytes). All the agents and services information is encrypted / decrypted on the basis of the local key generated. In order to reduce the overhead of the encryption / decryption process we have used this process only when the information is to be inserted / updated into the database. Searching is the operation that is performed more frequently as compared to insert / update. So when the agent platform is booted up then the agent information is decrypted along with its services and the searching operation does not require decryption process to retrieve the information, hence improving the system performance.

Figure 3 shows the security module incorporated in the AMS.

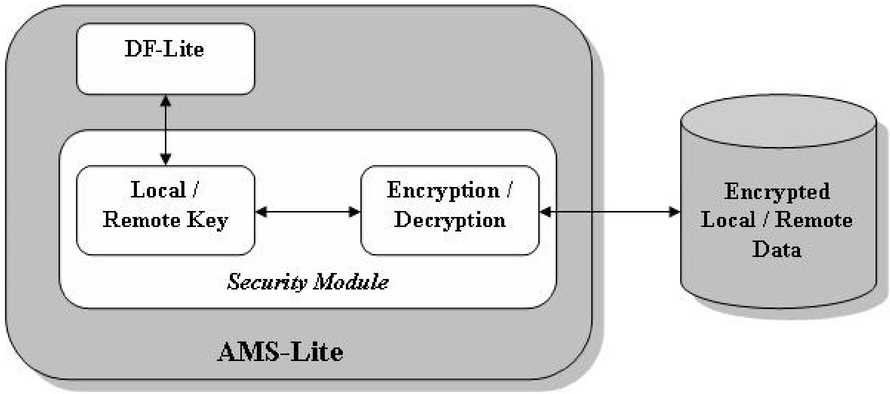


Fig. 3. Secure storage in the record store of J2ME

4.2 Secure Agents Communication

When the agents send messages to other agents following properties should be met:

i) **Confidentiality:** ensuring that access to certain messages is only possible for the intended parties. So the communicated information between agents is not accessible by unauthorized parities.

ii) **Authentication of Origin:** is the assurance that communication originates from its claimant.

iii) **Non Repudiation:** the ability to prove a user performed a certain action, even if that user denies it.

Figure 4 proposes the secure communication through Bluetooth between two wireless mobile clients having SAGE-Lite on the devices.

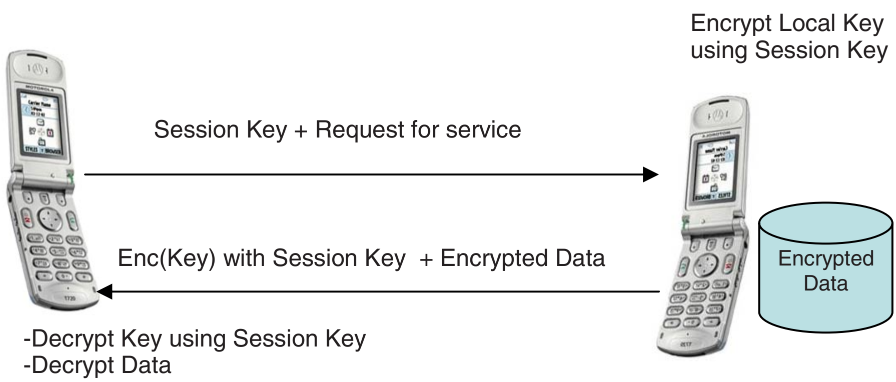


Fig. 4. Secure Communication between agents through Bluetooth

When an agent requests the services from agents residing on another platform, then the services are searched and granted to the requestor. To secure the communication between two devices we have used the following steps:

1) A session will be created between the requestor and provider device with a session key, which will be a 128 bit random number generated by the requestor.
2) The Session key will be transmitted to the provider along with the service request.
3) The provider will encrypt its local key on the basis of Session key sent by the requestor.
4) The provider will send the encrypted local key and encrypted data / service to the requestor.
5) The requestor will then decrypt the remote data by the key sent by provider.

The advantage of using the session ensures the blocking of eavesdroppers to hack the data during the communication. If the data is intercepted during communication then it can not be decrypted as the actual key will not be available to the eavesdropper. Another advantage of this technique is related to the performance of the system. Instead of encrypting the data on runtime (which is very time consuming) we are encrypting only the key and the data is already encrypted.

5 Evaluation

This section establishes the criteria by which we intend to judge the success of our proposed work. The results were tested on Nokia N70 device, which supports J2ME applications with Bluetooth API. We have considered the following scenarios for evaluation purpose:

i) Securing the agent's information on local device helps preventing the unauthorized access to agent information. If the agent information is moved or copied to another device then agent information is still unavailable to the unauthorized users, hence fulfilling the property of confidentiality.
ii) By using the session keys for the exchange of data ensures the authentication of origin. Agent to agent communication on different platforms is made secure on the basis of encrypting key, based on Session keys.
iii) Availability and performance of the system is achieved by communicating the encrypted information between agents with the transformation of the keys on Session Keys basis, thus making it possible to communicate securely with minimal affect on the system's performance as only keys are encrypted during communication.

We used FExplorer, an application that can run on the Nokia devices and can read the information of record stores of MIDlets present in the mobile device. The record stores are accessible in the (.db) file format and can be copied to another device using Bluetooth. The entire agent's information was stored in the record store file on device 'A' that was copied through Bluetooth to another device 'B', but on device 'B' the agent information could not be retrieved as it was encrypted and the same application running on the device 'B' could not decrypt the agent's information, as it was encrypted on device 'A' fixed IP.

6 Conclusion

Multi agents systems on the resource constrained devices require security due to their large deployment in the business and e-commerce sector. We have implemented the security in the SAGE-Lite (lightweight multi-agent system) that focuses on the secure storage and communication of the agents residing on the same or different platforms. The solution provided makes the secure storage and communication with minimal affect on the system's performance. The solution has been implemented using Bluetooth technology; however, it can be generalized to any wireless environment like Wifi, WiMax etc.

7 Future Work

Future work can be related to secure communication of SAGE-Lite with the SAGE through WAP. All the agents created on the clients should be registered with SAGE. Before the communication of the agents; authentication and authorization should be made through the server SAGE via WAP and further communication between clients through Bluetooth.

References

[1] Sun's Java 2 Micro Edition Platform, http://java.sun.com/javame/index.jsp
[2] Itani, W., et al.: J2ME End to End Security for M-Commerce, University of Beirut. Published in WCNC, IEEE (2003)
[3] Klingsheim, A.N., et al.: Challenges in Securing Networked J2ME Applications, Published by the IEEE Computer Society Computing Practices (2007)
[4] Farooq, H., et.,, al,: Persistent Architecture for Context Aware Lightweight Multi Agent System. In: The Fifth International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS 2006), Hakodate, Japan, May 8-12 (2006)
[5] MIDP Security: Design Concerns & Cryptography, http://developers.sun.com/techtopics/mobility/midp/articles/security1/index.html
[6] Mobile Information Device Profile (MIDP 2.0) security framework. http://developers.sun.com/techtopics/mobility/midp/articles/midp20/
[7] Harbitter, A., et al.: The performance of Public key-enabled Kerberos authentication in mobile computing. In: Proceedings of the 8th ACM conference on Computer and Communications Security, 2001.Conference on Communications and mobile computing, pp. 1295–1300 (2001)
[8] Barka, E.S., et al.: End-to-End security solutions for WLAN: A performance analysis for the underlying encryption algorithms in the lightweight devices. In: International Conference on Communications and Mobile Computing (2006)
[9] Bouncy Castle Cryptography APIs, http://www.bouncycastle.org

[10] Daemen, J., et al.: Rijndael, the Advanced Encryption Standard. Published in Journal of Network and Computer Applications (March 2001)
[11] Ali, A., et al: Scalable fault tolerant Agent Grooming Environment – SAGE, NUST Institute of Information Technology (NIIT) National University of Sciences and Technology (NUST) Rawalpindi, Pakistan
[12] Foundation for Intelligent Physical Agents, http://www.fipa.org

A New Key Agreement Protocol Based on Chaotic Maps

Eun-Jun Yoon[1] and Kee-Young Yoo[2,⋆]

[1] Faculty of Computer Information, Daegu Polytechnic College,
42 Jinri-2gil (Manchon 3dong San395), Suseong-Gu, Daegu 706-711, South Korea
ejyoon@tpic.ac.kr

[2] Department of Computer Engineering, Kyungpook National University,
1370 Sankyuk-Dong, Buk-Gu, Daegu 702-701, South Korea
yook@knu.ac.kr

Abstract. In 2007, Xiao et al.'s proposed a novel key agreement protocol based on chaotic maps. Recently, Han et al., however, pointed out that Xiao et al.'s protocol is not secure against their developed new attacks which can hinder the user and the server from accomplishing a shared secret session key. In addition Han et al.'s attack, the current paper demonstrates that Xiao et al.'s protocol is vulnerable to off-line password guessing attack. Furthermore, the current paper presents a new key agreement protocol based on chaotic maps in order to isolate such security problems. In contrast with Xiao et al.'s protocol, the proposed protocol is more secure and can be reduced the numbers of communication rounds.

Keywords: Security, Key agreement, Chaotic maps, Protocol, Session key.

1 Introduction

Recently, chaotic systems [1,2,3,4,5] have been used to design secure communication protocols. There are two main methods in the design of chaotic secure communication protocols: analog [1,4,6] and discrete digital [2].

- In the analog method, chaos synchronization is achieved using chaotic circuits, where the communication parties take advantage of the chaotic synchronization to implement secure communication.
- In the discrete digital method, chaotic systems are used to generate chaotic ciphers for secure communication.

Key agreement protocol [7] is widely used to derive a shared secret session key by two or more parties as a function of information contributed by, or associated with, each of these, but no party can predetermine the resulting value. In 2003,

⋆ Corresponding author: Kee-Young Yoo (yook@knu.ac.kr) Tel.: +82-53-950-5553; Fax: +82-53-957-4846.

N.T. Nguyen et al. (Eds.): KES-AMSTA 2008, LNAI 4953, pp. 897–906, 2008.

Kocarev et al. [8] proposed a public key encryption protocol using chaotic maps. In 2005, Bergamo et al. [9], however, pointed out that Kocarev et al.'s protocol is insecure to their developed attack. That is, an adversary can retrieve the private information from the communication parties. In 2007, Xiao et al. [10] proposed a novel key agreement protocol based on chaotic maps in order to isolate Bergamo et al.'s attack. Recently, Han et al. [11], however, pointed out that Xiao et al.'s protocol is still insecure against their developed new attacks which can hinder the user and the server from accomplishing a shared secret session key.

Beside Han et al.'s attack, Xiao et al.'s protocol is susceptible to off-line password guessing attack [12,13]. Xiao et al.'s protocol uses user's password to prove the security. Password based key agreement protocols, however, are vulnerable to password guessing attacks [12,13] since users usually choose easy-to-remember passwords. Unlike typical private keys, the password has limited entropy, and is constrained by the memory of the user. For example, one alphanumerical character has 6 bits of entropy. Therefore, the goal of the attacker, which is to obtain a legitimate communication parties' password, could can be achieved within a reasonable time. Thus, the password guessing attacks on password based key agreement protocols should be considered realistic.

Accordingly, the current paper demonstrates the vulnerability of Xiao et al.'s key agreement protocol to the off-line password guessing attack, and then proposes a new key agreement protocol based on chaotic maps in order to overcome such security problems. In the proposed protocol, the numbers of communication rounds are reduced that can be executed in seven messages and three rounds, respectively. As a result, the proposed key agreement protocol is more efficient and secure than previously presented key agreement protocols based on chaotic maps.

This paper is organized as follows: In Section 2, we briefly review Xiao et al.'s key agreement protocol based on chaotic maps. An outline of the off-line password guessing attack on Xiao et al.'s protocol is proposed in Section 3. In Section 4, we present an improvement of Xiao et al.'s protocol and then analyze the security and efficiency of our proposed protocol in Section 5. Finally, our conclusions are given in Section 6.

2 Review of Xiao et al.'s Key Agreement Protocol

This section briefly reviews of Xiao et al.'s key agreement protocol based on chaotic maps [10].

2.1 Chebyshev Map

Xiao et al.'s protocol is based on the following Chebyshev map [8,14], which has semi-group property. Chebyshev polynomial of degree p is defined as

$$T_p(x) = \cos(p \times \arccos x)$$

Its domain is the interval $x \in [-1, 1]$. It can be recurrently defined as

$$T_{p+1}(x) = 2xT_p(x) - T_{p-1}(x)$$

Chebyshev polynomials have the following two important properties [14]:

(1) A chaotic map: If the degree $p > 1$, $T_p : [-1, 1] \mapsto [-1, 1]$ is a prototype of a chaotic map. This map has a unique absolutely continuous invariant measure:

$$\mu(x)\mathrm{d}x = \mathrm{d}x/(\pi\sqrt{1-x^2})$$

with positive Lyapunov exponent $\lambda = \ln p$.

(2) The semi-group property:

$$T_p(T_q(x)) = T_q(T_p(x)) = T_{pq}(x)$$

2.2 Hash Function Based on Chaotic Maps

Xiao et al.'s protocol is based on the following chaotic one-way hash function [10]. One dimension piecewise linear chaotic system was defined as

$$X(t+1) = F_p(X(t)) = \begin{cases} X(t)/P & \text{if } 0 \leqslant X(t) < P, \\ (X(t)-P)/(0.5P) & \text{if } P \leqslant X(t) < 0.5, \\ (1-X(t)-P)/(0.5-P) & \text{if } 0.5 \leqslant X(t) < 1-P, \\ (1-X(t))/P & \text{if } 1-P \leqslant X(t) \leqslant 1, \end{cases}$$

where $X \in [0, 1], P \in (0, 0.5)$. Let $X_i (0 \leqslant i \leqslant 3N)$ is the *chaining variable*. At the start of the hashing, the chaining variable has an *initial value* X_0 that is specified as part of the hash algorithm, where X_0 is chosen from $(0, 1)$. Let H_0 is the *encryption key* for the pending message M. Given a pending message M, H_0 is a constant which is chosen from $(0, 1)$. The 3-unit iterations, 1st to Nth, $(N + 1)$th to $2N$th, $(2N + 1)$th to $3N$th, ensure that each bit of the final hash value will be related to all the bits of messages. The following is a brief referring to how to generate the hash value:

(1) Translates the pending message to the corresponding ASCII numbers, then maps these ASCII numbers by means of linear transform into an array C, whose elements are numbers in $[0, 1]$ and whose length N is the number of characters in the message.

(2) Iterates the follows processes:
   - (i) 1st: $P_1 = (C_1 + H_0)/4 \in [0.05), X_1 = F_{P_1}(X_0) \in [0, 1]$;
   - (ii) 2nd to Nth: $P_i = (C_i + X_{i-1})/4 \in [0.05), X_i = F_{P_i}(X_{i-1}) \in [0, 1]$;
   - (iii) $(N+1)$th: $P_{N+1} = (C_N + X_N)/4 \in [0.05), X_{N+1} = F_{P_{N+1}}(X_N) \in [0, 1]$;
   - (iv) $(N+2)$th to $2N$th: $P_i = (C_{2N-i+1} + X_{i-1})/4 \in [0.05), X_i = F_{P_i}(X_{i-1}) \in [0, 1]$;
   - (v) $(2N + 1)$th: $P_{2N+1} = (C_1 + H_0)/4 \in [0.05), X_{2N+1} = F_{P_{2N+1}}(X_{2N}) \in [0, 1]$;

(vi) $(2N+2)$th to $3N$th: $P_i = (C_{i-2N}+X_{i-1})/4 \in [0.05)$, $X_i = F_{P_i}(X_{i-1}) \in [0, 1]$.

(3) Transforms X_N, X_{2N} , X_{3N} to the corresponding binary format, extracts 40, 40, 48 bits after the decimal point, respectively, and juxtaposes them from left to right to get a 128-bit final hash value.

Shared Information: Chebyshev polynomial, $E(\cdot)$, $D(\cdot)$, $H(\cdot)$
Information held by User A: ID_A, $h_{PW} = H(ID_A, PW)$
Information held by Server B: $h_{PW} = H(ID_A, PW)$

User A		**Server B**
Choose $ra \in [-1, 1]$	$\xrightarrow{ID_A, ra}$	Choose $rb \in [-1, 1]$
	$\xleftarrow{rb}$	
$AU = H(h_{PW}, ra, rb)$	$\xrightarrow{ID_A, AU}$	$AU' = H(h_{PW}, ra, rb)$
		Check $AU \stackrel{?}{=} AU'$
Seed $x = ra$ Choose random r $X = E_{h_{PW}}(T_r(x))$	$\xrightarrow{X}$	Seed $x = ra$ Choose random s $Y = E_{h_{PW}}(T_s(x))$
	$\xleftarrow{Y}$	
$T_r(x) = D_{h_{PW}}(X)$		$T_s(x) = D_{h_{PW}}(Y)$

Shared secret session key: $sk = T_r(T_s(x)) = T_s(T_r(x)) = T_{rs}(x)$

Fig. 1. Xiao et al.'s key agreement protocol based on chaotic maps

2.3 Xiao et al.'s Key Agreement Protocol

Xiao et al.'s key agreement protocol uses the above semi-group property of Chebyshev map and one-way hash function based on Chebyshev chaotic maps. Fig.1 illustrates Xiao et al.'s protocol. Before performing the key agreement protocol, assume that user A and server B secretly share the password hash value $h_{PW} = H(ID_A, PW)$, where PW is A's password and $H(\cdot)$ is a one-way hash function based on Chebyshev chaotic maps. The details of Xiao et al.'s key agreement protocol are as follows:

(1) $A \rightarrow B$: ID_A, ra
User A chooses a random number $ra \in [-1, 1]$, then he/she sends his/her identity number ID_A and ra to server B.

(2) $B \rightarrow A$: rb
Server B chooses a random number $rb \in [-1, 1]$ and sends it back to A.

(3) $A \rightarrow B$: ID_A, AU
User A juxtaposes h_{PW}, ra and rb from left to right as the pending message, and uses the one-way hash function $H(\cdot)$ to compute the authentication value $AU = H(h_{PW}, ra, rb)$. Then he/she sends ID_A and AU to B.

(4) Server B takes out his/her own copies of h_{PW}, ra and rb corresponding to the user identity number ID_A. Then he/she uses the same hash function to compute $AU' = H(h_{PW}, ra, rb)$ similarly.
(5) After receiving AU, server B compares it with the computed AU'. If they are equal, then the identity of A is authenticated.

By using the random number ra chosen in Step (1) as the seed x of the Chebyshev polynomial map, the key agreement operations are performed as follows:

(6) $A \rightarrow B$: X
User A chooses a random integer r, computes $X = E_{h_{PW}}(T_r(x))$, and then sends it to B.
(7) $B \rightarrow A$: Y
Server B chooses a random integer s, computes $Y = E_{h_{PW}}(T_s(x))$, and then sends it to A.
(8) User A and server B decrypt X and Y, respectively, then compute the shared session secret key: $sk = T_r(T_s(x)) = T_s(T_r(x)) = T_{rs}(x)$.

3 Off-Line Password Guessing Attack on Xiao et al.'s Protocol

This section shows that Xiao et al.'s key agreement protocol based on chaotic maps is vulnerable to an off-line password guessing attack [12,13]. Unlike typical private keys, a password has low entropy, and is constrained by the memory of the user. Roughly speaking, the entropy of a user generated password is about 2 bits per character [7]. Therefore, the goal of the attacker, which is to obtain a legitimate communication party's password, can be achieved within a reasonable time. Thus, password guessing attacks on key agreement protocols are a real threat. Let Adv be an active adversary who interposes the communication between A and B. Suppose that Adv has eavesdropped valid messages (ID_A, ra, rb, AU) from an open network. It is easy to obtain the information since the messages are all exposed over the open network. Then, the off-line password guessing attack proceeds as follows:

(1*) To obtain the password PW of user U_A, the adversary Adv makes a guess at the secret password PW^* from dictionary D and computes $h^*_{PW} = H(ID_A, PW^*)$.
(2*) Adv computes $AU^* = H(h_{PW^*}, ra, rb)$ and checks if $AU \stackrel{?}{=} AU^*$. If the computed value is the same as AU, Adv then can guess the user U_A's password PW. Otherwise, Adv repeatedly performs it until $AU \stackrel{?}{=} AU^*$.

The algorithm of an off-line password guessing attack is as follows:
As a result, Xiao et al.'s key agreement protocol is susceptible to an off-line password guessing attack.

Off-line Password Guessing Attack (ID_A, ra, rb, AU)

{

for $i := 0$ **to** $|D|$

{

$PW^* \leftarrow D$;

$h^*_{PW} \leftarrow H(ID_A, PW^*)$;

$AU^* = H(h_{PW^*}, ra, rb)$;

if $AU = AU^*$ **then return** PW^*

}

}

4 Proposed Key Agreement Protocol

This section proposes an enhancement to the Xiao et al.'s protocol [10] that can withstand the security flaws described in Han et al.'s paper [11]. Fig.2 illustrates the proposed protocol. Before performing the key agreement protocol, we also assume that user A and server B secretly share the password hash value $h_{PW} = H(ID_A, PW)$ like Xiao et al.'s. The details of our proposed key agreement protocol are as follows:

(1) $A \rightarrow B$: ID_A, X
User A chooses a random number $ra \in [-1, 1]$ and a random integer r, and then, by using the random number ra as the seed x of the Chebyshev polynomial map, computes $X = E_{h_{PW}}(ra, T_r(x))$. Finally, A sends X with ID_A to server B.

(2) $B \rightarrow A$: ID_B, Y, AU_B
After receiving ID_A and X, server B extracts ra and $T_r(x)$ by decrypting X using the shared password hash value h_{PW}. B chooses a random number $rb \in [-1, 1]$ and a random integer s, and then computes $Y = E_{h_{PW}}(rb, T_s(x))$ and the shared secret session key $sk_B = T_s(T_r(x)) = T_{rs}(x)$. Finally, server B computes the authentication value $AU_B = H(ID_A,\ ra, r, sk_B)$ and sends it with his/her identity number ID_B and Y to A.

(3) $A \rightarrow B$: ID_A, AU_A
After receiving ID_B, Y and AU_B, user A extracts rb and $T_s(x)$ by decrypting Y using the shared password hash value h_{PW}. A computes the shared secret session key $sk_A = T_r(T_s(x)) = T_{rs}(x)$ and $AU'_B = H(ID_A, ra, r, sk_A)$, and then compares it with the received AU_B. If they are equal, then the identity of B is authenticated. Finally, for mutual authentication, A computes $AU_A = H(ID_B, rb, s, sk_A)$ and sends it with ID_A to B.

(4) After receiving ID_A and AU_A, server B computes $AU'_A = H(ID_B, rb, s, sk_B)$ and then compares it with the received AU_A. If they are equal, then the identity of A is authenticated.

After mutual authentication and session key agreement between A and B, $sk_A = T_r(T_s(x)) = T_{rs}(x)$ and $sk_B = T_s(T_r(x)) = T_{rs}(x)$ are used as a shared secret session key, respectively.

Shared Information: Chebyshev polynomial, $E(\cdot)$, $D(\cdot)$, $H(\cdot)$
Information held by User A: ID_A, $h_{PW} = H(ID_A, PW)$
Information held by Server B: ID_B, $h_{PW} = H(ID_A, PW)$

User A		**Server B**
Choose $ra \in [-1, 1]$ Choose random r Seed $x = ra$ $X = E_{h_{PW}}(ra, T_r(x))$	ID_A, X $\longrightarrow$	$ra, T_r(x) = D_{h_{PW}}(X)$ Choose $rb \in [-1, 1]$ Choose random s Seed $x = ra$ $Y = E_{h_{PW}}(rb, T_s(x))$ $sk_B = T_s(T_r(x)) = T_{rs}(x)$ $AU_B = H(ID_A, ra, r, sk_B)$
$rb, T_s(x) = D_{h_{PW}}(Y)$ $sk_A = T_r(T_s(x)) = T_{rs}(x)$ $AU'_B = H(ID_A, ra, r, sk_A)$ Check $AU_B \stackrel{?}{=} AU'_B$ $AU_A = H(ID_B, rb, s, sk_A)$	$\longleftarrow$ ID_B, Y, AU_B	
	ID_A, AU_A $\longrightarrow$	$AU'_A = H(ID_B, rb, s, sk_B)$ Check $AU_A \stackrel{?}{=} AU'_A$

Shared secret session key: $sk_A = sk_B = T_r(T_s(x)) = T_s(T_r(x)) = T_{rs}(x)$

Fig. 2. Proposed key agreement protocol based on chaotic maps

5 Security Analysis

This section provides the proof of correctness of the proposed key agreement protocol. First, we define the security terms [15,16] needed for security analysis of the proposed protocol as follows:

Definition 1. *A weak secret (Password PW) is a value of low entropy $Weak(k)$, which can be guessed in polynomial time.*

Definition 2. *The discrete logarithm problem (DLP) is explained by the following: Given an element α, find the integer r, such that $T_r(x) \equiv \alpha$.*

Definition 3. *The Diffie-Hellman problem (DHP) is explained by the following: Given an elements $T_r(x)$ and $T_s(x)$, find $T_{rs}(x)$.*

Table 1. Comparison of security properties

	Xiao et al.'s protocol [10]	Proposed protocol
Replay attacks	Insecure	Secure
Han et al.'s attacks	Insecure	Secure
Password guessing attacks	Insecure	Secure
Man-in-middle attacks	Insecure	Secure
Mutual authentication	Not provide	Provide
Perfect forward secrecy	Provide	Provide
Communication rounds	5	3

Definition 4. *A secure chaotic one-way hash function $y = H(x)$ is one where given x to compute y is easy and given y to compute x is hard.*

Here, five security properties [15,16]: replay attacks, password guessing attacks, man-in-middle attacks, and perfect forward secrecy, must be considered for the proposed key agreement protocol. Under the above four definitions, the following theorems are used to analyze five security properties in the proposed protocol.

Theorem 1. *Proposed protocol can resist the replay attacks and Han et al.'s developed attacks [11].*

Proof: Suppose an adversary Adv intercepts $X = E_{h_{PW}}(ra, T_r(x))$ from A in Step (1) and uses it to impersonate A by using replay attack. However, Adv cannot compute a correct an authentication value $AU_A = H(ID_B, rb, s, sk_A)$ and deliver it to B unless he/she can correctly guess password PW to obtain ra and $T_r(x)$ from X and guess the right random integer number r from $T_r(x)$, and then Adv must face the discrete logarithm problem (DLP). On the other hand, suppose Adv intercepts Y and $AU_B = H(ID_A, ra, r, sk_B)$ from B in Step (2), and uses it to impersonate B by using replay attack. For the same reason, if Adv cannot gain the correct random integer number rb and $T_s(x)$ from Y, A will find out that $AU'_B = H(ID_A, ra, r, sk_A)$ is not equivalent to AU_B, and then A will not send AU_A back to Adv. Because Han et al.'s developed attacks are also a kind of replay attacks, the proposed protocol can simply detect the attack by verifying AU_B and AU_A in Step (3) and (4), respectively. Therefore, the proposed protocol can resist the replay attack and Han et al.'s developed attacks.

Theorem 2. *Proposed protocol can resist the password guessing attacks.*

Proof: An on-line password guessing attack cannot succeed since B can choose appropriate trail intervals. On the other hand, in an off-line password guessing attack, due to the Definition 1, Adv can try to find out a weak password by repeatedly guessing possible passwords and verifying the correctness of the guesses based on information obtained in an off-line manner. In our proposed protocol, Adv can gain the knowledge of X, Y, AU_B, and AU_A in Steps (1), (2), and (3), respectively. Assume that Adv wants to impersonate A. He/she first guesses password PW^* and then finds ra and $T_r(x)$ by decrypting $X = E_{h^*_{PW}}(ra, T_r(x))$ and rb and $T_s(x)$ by decrypting $Y = E_{h^*_{PW}}(rb, T_s(x))$, where $h^*_{PW} = H(ID_A, PW^*)$. However, Adv has to break the discrete logarithm problem (DLP) and the Diffie-Hellman problem (DHP) [15] to find the session key $sk_A = sk_B$ to verify his/her guess. Thus, Adv cannot gain the shared session key without r and s. Therefore, the proposed protocol can resist the password guessing attacks.

Theorem 3. *Proposed protocol can resist the man-in-middle attacks.*

Proof: A mutual password PW between A and B is used to prevent the man-in-middle attack. The illegal Adv cannot pretend to be A or B to authenticate the other since he/she does not own the mutual password PW. Therefore, the proposed protocol can resist the man-in-middle attacks.

Theorem 4. *Proposed protocol can achieve mutual authentication.*

Proof: In Step (2) of the proposed protocol, A can authenticate B by checking $AU_B \stackrel{?}{=} AU'_B$ because only a valid B can compute AU_B. In Step (4) of the proposed protocol, B also can authenticate A by checking $AU_A \stackrel{?}{=} AU'_A$ because only a valid A can compute AU_A. Therefore, the proposed protocol can achieve mutual authentication.

Theorem 5. *Proposed protocol provides perfect forward secrecy.*

Proof: Perfect forward secrecy means that if long-term private keys of one or more entities are compromised, the secrecy of previous session keys established by honest entities is not affected. If the user's password PW is compromised, it does not allow Adv to determine the session key $sk_A = sk_B = T_{rs}(x)$ for past sessions and decrypt them, since Adv is still faced with the Diffie-Hellman problem (DHP) to compute the session key $sk_A = sk_B = T_{rs}(x)$. Therefore, the proposed protocol satisfies the property of perfect forward secrecy.

The security properties of Xiao et al.'s protocol and of the proposed protocol are summarized in Table 1. In contrast with Xiao et al.'s protocol, the proposed protocol is more secure and efficient.

6 Conclusions

The current paper demonstrated the vulnerability of Xiao et al.'s key agreement protocol based on chaotic maps to the off-line password guessing attack, and then proposed a new key agreement protocol based on chaotic maps in order to overcome such security problems. In the proposed protocol, the numbers of communication rounds are reduced that can be executed in seven messages and three rounds, respectively. As a result, compare with Xiao et al.'s protocol, the proposed protocol has same complexity and is more secure and round efficient.

Acknowledgements

This research was supported by the MIC of Korea, under the ITRC support program supervised by the IITA(IITA-2007-C1090-0701-0026).

References

1. Dachselt, F., Schwarz, W.: Chaos and cryptography. IEEE Trans Circ Syst-I, Fundam Theory 48(12), 1498–1509 (2001)
2. Fridrich, G.: Symmetric ciphers based on two-dimensional chaotic maps. Int. J. Bifur. Chaos. 8, 1259–1284 (1998)
3. Pecorra, T.L., Carroll, T.L.: Driving systems with chaotic signals. Phys. Rev. A 44(4), 2374–2383 (1991)
4. Kocarev, L.: Chaos-based cryptography: a brief overview. IEEE Circ. Syst. Mag. 1, 6–21 (2001)
5. Wong, K.W.: A fast chaotic cryptographic scheme with dynamic look-up table. Phys. Lett. A 298, 238–242 (2002)
6. Grassi, G., Mascolo, S.: IEEE Trans. Circ. Syst. I. Fundam. Theory. Appl. 46, 1135 (1999)
7. Menezes, A., van Oorschot, P., Vanstone, S.: Handbook of applied cryptography. CRC Press, Boca Raton (1997)
8. Kocarev, L., Tasev, Z.: Public-key encryption based on Chebyshev maps. In: Proc IEEE Symp Circ Syst (ISCAS 2003), vol. 3, pp. 28–31 (2003)
9. Bergamo, P., D'Arco, P., Santis, A., Kocarev, L.: Security of public key cryptosystems based on Chebyshev polynomials. IEEE Trans Circ Syst-I 52(7), 1382–1393 (2005)
10. Xiao, D., Liao, X.F., Deng, S.: A novel key agreement protocol based on chaotic maps. Inform. Sci. 177, 1136–1142 (2007)
11. Han, S.: Security of a key agreement protocol based on chaotic maps. Chaos, Solitons & Fractals (in press) doi:10.1016 /j.chaos.2007.01.017
12. Gong, L., Lomas, M.A., Needham, R., Saltzer, J.: Protecting poorly chosen secrets from guessing attacks. IEEE Journal on Selected Areas in Communications 11(5), 648–656 (1993)
13. Ding, Y., Horster, P.: Undetectable on-line password guessing attacks. ACM Operating Systems Review 29(4), 77–86 (1995)
14. Rivlin, T.J.: Chebysheve polynomials. John Wiley and Sons, Inc., New York (1990)
15. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Transaction on Information Theory IT-22(6), 644–654 (1976)
16. Schneier, B.: Applied cryptography protocols. Algorithms and Source Code in C, 2nd edn. John Wiley, Chichester (1995)

Author Index

Ahmad, Hafiz Farooq 887
Ahn, Hyung Jun 594
Ahn, Kwang Seon 182
Amor, M. 505
Anthony, Patricia 381
Ayanegui, Huberto 391
Aye, Nilar 252

Bahsoon, Rami 887
Banaszak, Zbigniew 421
Bandar, Zuhair 172
Ben Hmida, Faten 773
Bocewicz, Grzegorz 421
Boryczka, Mariusz 302
Breslin, John G. 83
Brzykcy, Grażyna 162
Bui, Ngot Phu 242
Bujang, Siti Dianah Abdul 813

Cao, Longbing 333, 361
Chae, Oksam 152
Chang, Pai-Yu 21
Chang, Yuming 723
Chen, Chih-Chun 1
Chen, Pei-Lin 113
Chen, You-Jia 93
Cheng, Chia-Ying 281, 485
Cho, Jae Hyung 614
Cho, Sung-hoon 878
Choi, Eui-in 878
Choi, Hyung Rim 614
Choi, YoungSik 454
Chung, TaeChoong 242
Clack, Christopher D. 1
Crockett, Keeley 172
Czarnowski, Ireneusz 536

Delhom, M. 323
Dewan, M. Ali Akber 152
Doi, Miwako 11

Encheva, Sylvia 32

Fernández-Caballero, Antonio 63, 73
Frackowiak, Grzegorz 271
Fu, Yingfang 584
Fuentes, L. 505
Fukumoto, Taro 351

Ganzha, Maria 271
Gatton, Thomas M. 547
Gawinecki, Maciej 271
Gedeon, Tom 401
Gul, Naseer 887

Haarslev, Volker 526
Håkansson, Anne 122, 261, 411
Han, Yo-Sub 271
Han, Young-Shin 232
Han, Zhangang 604
Hannaford, Michael 132
Hartung, Ronald 261, 411
Hattori, Fumio 252
Hayashi, Hisashi 11
He, Jingsha 584
Hector, Aaron 132
Heh, Tzeng-Yuan 21
Henskens, Frans 132
Heo, Jeong Seok 624
Higgin, Henry 464
Hla, Khin Haymar Saw 454
Ho, Chong Mun 381
Ho, M.K. 822
Hoang, Trong Nghia 743
Hong, You-Sik 868
Hossain, M. Julius 152
Huang, Chenn-Jung 93
Huang, Chung-Yuan 281, 485
Huang, Shell Ying 723

Ishii, Naohiro 474
Ito, Nobuhiro 474
Iwata, Kazunori 474

Jain, Lakhmi C. 341, 495
Jędrzejowicz, Piotr 536
Jung, Jae-Youn 192
Jung, Jae Un 614
Jung, Jason J. 567

Kaneda, Youhei 474
Kang, Eunyoung 733
Kang, Sanggil 212
Karras, D.A. 431
Katarzyniak, Radosław Piotr 341, 693, 763
Kim, Bong-Hoi 878
Kim, CheonShik 868
Kim, ChongGun 192, 203
Kim, Chul-Su 683
Kim, Eungyeong 547
Kim, Hak-Lae 83
Kim, Hanseup 594
Kim, Hong-Gee 83
Kim, HyeongRag 858
Kim, Hyun Soo 614
Kim, Kono 567
Kim, Tae-young 232
Kim, Taekyu 232
Kim, Ungmo 733
Kim, Won Tae 182
Kim, Yung Bok 664, 683
Król, Dariusz 793
Kumar, Harshit 212
Kuribara, Syuusuke 351
Kuwabara, Kazuhiro 252
Kwak, Kyung Sup 464

Lai, Kin Keung 312
Lau, H.C.W. 103
Lee, Eun-Ser 444
Lee, Eunseok 142
Lee, Geuk 868
Lee, Habin 594
Lee, Hojin 567
Lee, HoonJae 858
Lee, Jae-Jo 664
Lee, Jaewan 753
Lee, Jong-Sik 232
Lee, Jun Wook 683
Lee, Malrey 547
Lee, Moo-hun 878
Lee, Sang-Hong 53
Lee, Sang Ho 444
Lee, Seong Joon 182
Lee, SungSoo 203
Lee, Youngsook 850
Lejouad Chaari, Wided 773
Leng, Jinsong 495
Li, Guorui 584
Li, Yueping 713
Lim, Deborah 381
Lim, Joon S. 53
Lin, Chen 634
Lin, Po-Hung 21
Liu, Bo 333
Liu, Guo 604
Liu, James N.K. 822
Liu, ZhiPeng 203
Lorkiewicz, Wojciech 763
Lu, Yunting 713
Luan, Liangyu 584
Luo, Dan 333, 361

Martens, Robert 832
Mateo, Romeo Mark A. 753
Matsuhisa, Takashi 654
Matsui, Toshihiro 292
Matsuo, Hiroshi 292
McLean, David 172
Minami, Toshiro 222
Moon, SangJae 858
Moradian, Esmiralda 122

Nagl, Sylvia B. 1
Nam, Junghyun 850
Nejad, Sara Ghoreishi 832
Ng, Kam-Wing 557
Nguyen, Ngoc Thanh 341, 703
Nishikawa, Asaki 803

O'Shea, James 172
Oh, Hoon 624
Ou, C.R. 574
Ou, Chung-Ming 574
Ozaki, Fumio 11

Papademetriou, R.C. 431
Papazoglou, P.M. 431
Paprzycki, Marcin 271
Paranjape, Raman 832
Park, Jong Sou 454
Park, Myon-Woong 271
Park, Seongrae 567
Park, Sung Joo 594
Park, Sungjoon 212
Park, Young-Tack 242
Popek, Grzegorz 693
Prusiewicz, Agnieszka 842

Ramos, Fernando 391
Rong, Wang 584

Sathyaraj, Beulah M. 495
Sawamura, Hajime 351
Selamat, Ali 813
Shaban-Nejad, Arash 526
Shadabi, Fariba 674
Shah, Syed Muhammad Ali 887
Sharma, Dharmendra 674
Shepherdson, John 594
Shiah, Sheau-Wen 21
Siddiqui, M. Arif 464
Siricharoen, Waralak V. 516
Sohn, Y.T. 271
Sokolova, Marina V. 63
Sun, Chuen-Tsai 281, 485
Sun, Hua 604
Szczepaniak, P.S. 42, 647
Szymczak, Michal 271

Tadeusiewicz, R. 42
Tagina, Moncef 773
Tidemann, Axel 783
Tokura, Seiji 11
Tran, Trong Hieu 703
Truong, Minh Nhat Quang 743
Tumin, Sharil 32

Uehara, Kuniaki 371
Ullah, Sana 464
Urbani, D. 323

Valencia-Jiménez, J.J. 73
Valenzuela, J.A. 505

Wei, Hsiang-Yu 93
Wierzbowska, Izabela 536
Wójcik, Robert 421
Won, Dongho 850
Wong, Sze-Wing 557

Xia, Xingang 604
Xiao, Yanshan 333, 361

Yang, Dian-Xiu 93
Yang, Sung-Kwon 83
Yang, Yi-Ju 93
Yasumura, Yoshiaki 371
Yen, Jen 113
Yoo, Kee-Young 897
Yoon, Eun-Jun 897
Yoon, Hyogun 547
Yoon, Insook 753
Yoshikawa, Shohei 371
Youn, Hee Yong 733
Youn, Hyunsang 142
Yuen, Kevin K.F. 103

Zelmozer, Michał 793
Zhang, Huiliang 723
Zheng, Pei 312
Zheng, Zhigang 361
Zhu, Dingyun 401
Zin, Hyeon-Cheol 192