

Lecture Notes in Artificial Intelligence 4845

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Ning Zhong Jiming Liu Yiyu Yao
Jinglong Wu Shengfu Lu Kuncheng Li (Eds.)

Web Intelligence Meets Brain Informatics

First WICI International Workshop, WImBI 2006
Beijing, China, December 15-16, 2006
Revised Selected and Invited Papers

Series Editors

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

Volume Editors

Ning Zhong (zhong@maebashi-it.ac.jp)
Maebashi Institute of Technology, Japan
International WIC Institute/BJUT, Beijing, China

Jiming Liu (jiming@Comp.HKBU.edu.hk)
Hong Kong Baptist University, China
International WIC Institute/BJUT, Beijing, China

Yiyu Yao (yyao@cs.uregina.ca)
University of Regina, Saskatchewan, Canada
International WIC Institute/BJUT, Beijing, China

Jinglong Wu (wu@kagawa-u.ac.jp)
Kagawa University, Takamatu, Japan
International WIC Institute/BJUT, Beijing, China

Shengfu Lu (lusf@bjut.edu.cn)
International WIC Institute/BJUT, Beijing, China

Kuncheng Li (likuncheng@vip.sina.com)
Xuan Wu Hospital, Capital University of Medical Sciences, Beijing, China

Library of Congress Control Number: 2007939907

CR Subject Classification (1998): I.2, H.5.2, I.2.10, I.4, I.5, J.1, J.3-4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-77027-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-77027-5 Springer Berlin Heidelberg New York

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

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12196999 06/3180 5 4 3 2 1 0

Preface

All the chapters in this volume are invited contributions following a successful international workshop, namely, The First WICI International Workshop on Web Intelligence meets Brain Informatics (WImBI 2006) held in Beijing, China, December 15–16, 2006. The WICI (International WIC Institute, www.iwici.org/) is an international open educational research organization of the Web Intelligence Consortium (WIC, www.wi-consortium.org/). The institute is affiliated with Beijing University of Technology (BJUT).

The workshop explores a new perspective of Web Intelligence (WI) research from the viewpoint of Brain Informatics (BI). BI is a new interdisciplinary field studying human information processing mechanisms from both macro and micro points of view by cooperatively using experimental cognitive neuroscience and advanced WI-centric information technology. The new instrumentation (fMRI etc.) and advanced information technology are causing an impending revolution in Web Intelligence and brain sciences. This revolution is bi-directional: a new understanding and discovery of human intelligence models in brain sciences will yield a new generation of WI research and development; and WI-based technologies will provide a new powerful platform for brain sciences. The synergy between WI with BI will yield profound advances in our analysis and understanding of the natures of data, knowledge, intelligence, and wisdom, as well as their relationship, organization, and creation process. Fundamentals and implementation issues of WI will be studied as a central topic and in a unique way. It will fundamentally change the nature of information technology in general and artificial intelligence in particular, leading towards human-level WI.

In summary, the main features of the WImBI 2006 workshop and the book include:

- This workshop was the first in the field to focus on the interplay between (a) intelligent technologies, especially in the context of WI and (b) studies on human intelligence as explored in neuroscience, cognitive psychology, and brain science instrumentation.
- The participants of this workshop were by invitation only. All the invited attendees are presently world leaders in their respective areas, and can be expected to build a strong synergy and momentum for the “WI meets BI” research in the near future.
- The book, as a volume in the Springer LNCS/LNAI state-of-the-art survey, will be a milestone publication, with research visions and blueprints, for computer scientists and practitioners at large in this exciting interdisciplinary area. All the post-workshop, full-length papers were carefully reviewed and selected for inclusion.

WImBI 2006 had a very exciting program (www.wi-consortium.org/) with a number of features, ranging from technical sessions, keynote/invited talks,

demos/posters, and social programs. Many thanks go to the distinguished keynote speakers, Tomaso Poggio of MIT and Deyi Li of NSFC. We wish to express our gratitude to all members of the Workshop Organizing Committee and the International Advisory Board for their instrumental and unfailing support.

WImBI 2006 could not have taken place without the great team effort of the Local Organizing Committee and the support of Beijing University of Technology. Our special thanks go to Boyuan Fan and Zhenyang Lu (Organizing Chairs), Chunnian Liu, Baocai Yin, and Xunming Ji (Organizing Vice-Chairs) for their enormous efforts in planning and arranging the logistics of the workshop from registration/payment handling, venue preparation, accommodation booking, to banquet/social program organization. We would like to thank Shuai Huang, Jiajin Huang, Jia Hu, and Juzhen Dong, of the conference support team at the International WIC Institute (WICI), the Knowledge Information Systems Laboratory, Maebashi Institute of Technology, and Web Intelligence Laboratory, Inc. for their dedication and hard work.

We are very grateful to the WImBI 2006 corporate sponsors: State Administration of Foreign Experts Affairs, National Natural Science Foundation of China, Web Intelligence Consortium, Beijing University of Technology, Xuanwu Hospital/Capital University of Medical Sciences, Maebashi Institute of Technology, Web Intelligence Laboratory, Inc., and Springer *Lecture Notes in Computer Science* for their generous support. Last but not least, we thank Alfred Hofmann of Springer for his help in coordinating the publication of this special volume in an emerging and interdisciplinary research area.

September 2007

Ning Zhong
Jiming Liu
Yiyu Yao
Jinglong Wu
Shengfu Lu
Kuncheng Li

WImBI 2006 Workshop Organization

Organizing Chairs

Boyuan Fan	Beijing University of Technology, China
Zhenyang Lu	Beijing University of Technology, China

Organizing Vice-Chairs

Chunnian Liu	Beijing University of Technology, China
Baocai Yin	Beijing University of Technology, China
Xunming Ji	Xuanwu Hospital, Capital University of Medical Sciences, China

Program Chairs

Ning Zhong	International WIC Institute/BJUT, Maebashi Institute of Technology, Japan
Jiming Liu	International WIC Institute/BJUT, Hong Kong Baptist University, China
Yiyu Yao	International WIC Institute/BJUT, University of Regina, Canada
Jinglong Wu	International WIC Institute/BJUT, Kagawa University, Japan
Shengfu Lu	International WIC Institute, Beijing University of Technology, China
Kuncheng Li	Xuanwu Hospital, Capital University of Medical Sciences, China

International Advisory Board

Lin Chen	Chinese Academy of Sciences, China
Boi Faltings	Swiss Federal Institute of Technology, Switzerland
Dieter A. Fensel	University of Innsbruck, Austria
Ian Foster	Argonne National Lab/University of Chicago, USA
Pat Langley	Stanford University, USA
Deyi Li	Institute of China Electronic System Engineering, China
Ruqian Lu	Chinese Academy of Sciences, China
Setsuo Ohsuga	University of Tokyo, Japan

VIII Organization

Tomaso Poggio
Changxiang Shen
Benjamin Wah

Massachusetts Institute of Technology, USA
Beijing University of Technology, China
University of Illinois at Urbana-Champaign,
USA

Bo Zhang

Tsinghua University, China

Table of Contents

Introduction

Web Intelligence Meets Brain Informatics	1
<i>Ning Zhong, Jiming Liu, Yiyu Yao, Jinglong Wu, Shengfu Lu, Yulin Qin, Kuncheng Li, and Benjamin Wah</i>	
Neuroscience: New Insights for AI?	32
<i>Tomaso Poggio</i>	
Network Thinking and Network Intelligence	36
<i>Deyi Li, Liping Xiao, Yanni Han, Guisheng Chen, and Kun Liu</i>	

Synergy of Web Intelligence and Brain Informatics

Web Intelligence Meets Brain Informatics at the Language Barrier: A Procrustean Bed?	59
<i>Nick Cercone</i>	
Conversational Informatics Where Web Intelligence Meets Brain Informatics	73
<i>Toyoaki Nishida</i>	
Intelligence for Upgrading Information	97
<i>Setsuo Ohsuga</i>	
Toward Perception Based Computing: A Rough-Granular Perspective	122
<i>Andrzej Jankowski and Andrzej Skowron</i>	
Granular Computing: Modeling Human Thoughts in the Web by Polyhedron	143
<i>Tsau Young (T.Y.) Lin and Mong-Hang Vo</i>	

Cognitive Science, Neuroscience, and Brain Informatics

Biophysical Models of Neural Computation: Max and Tuning Circuits	164
<i>Ulf Knoblich, Jake Bouvrie, and Tomaso Poggio</i>	
Cognitive Architectures and the Challenge of Cognitive Social Simulation	190
<i>Ron Sun</i>	

ACT-R Meets fMRI	205
<i>Yulin Qin, Daniel Bothell, and John R. Anderson</i>	
The Neural Mechanism of Human Numerical Inductive Reasoning Process: A Combined ERP and fMRI Study	223
<i>Peipeng Liang, Ning Zhong, Shengfu Lu, Jiming Liu, Yiyu Yao, Kuncheng Li, and Yanhui Yang</i>	
Central Nervous Processing for Acupuncture at Liv3 with fMRI: A Preliminary Experience	244
<i>Wei Wang, Kuncheng Li, Baoci Shan, Jianyang Xu, Bin Yan, Jing Hao, Yanhui Yang, Ke Li, and Na Lu</i>	
A Role for Signal Propagation Through the Hippocampal CA2 Field in Memory Formation	254
<i>Yuko Sekino and Tomoaki Shirao</i>	
Genetic Granular Cognitive Fuzzy Neural Networks and Human Brains for Pattern Recognition	267
<i>Cui Lin, Jun Li, Natasha Barrett, Yan-Qing Zhang, and David A. Washburn</i>	
Domain-Oriented Data-Driven Data Mining (3DM): Simulation of Human Knowledge Understanding	278
<i>Guoyin Wang</i>	
An Ontology-Based Mining System for Competitive Intelligence in Neuroscience	291
<i>Jiao Li, Minlie Huang, and Xiaoyan Zhu</i>	
Web Intelligence Applications	
Supervised Web Document Classification Using Discrete Transforms, Active Hypercontours and Expert Knowledge	305
<i>P.S. Szczepaniak, A. Tomczyk, and M. Pryczek</i>	
Fuzzy Web Surfer Models: Theory and Experiments	324
<i>Narayan L. Bhamidipati and Sankar K. Pal</i>	
Intuitive Display for Search Engines Toward Fast Detection of Peculiar WWW Pages	341
<i>Einoshin Suzuki, Shin Ando, Masayuki Hirose, and Masatoshi Jumi</i>	
GridMiner: An Advanced Grid-Based Support for Brain Informatics Data Mining Tasks	353
<i>Peter Brezany, Ivan Janciak, Jarmila Brezanyova, and A Min Tjoa</i>	

A Semantically Enabled Service Oriented Architecture	367
<i>Darko Anicic, Michael Brodie, Jos de Bruijn, Dieter Fensel, Thomas Haselwanter, Martin Hepp, Stijn Heymans, Jörg Ho mann, Mick Kerrigan, Jacek Kopecky, Reto Krummenacher, Holger Lausen, Adrian Mocan, James Scicluna, Ioan Toma, and Michal Zaremba</i>	
Spam Filtering and Email-Mediated Applications	382
<i>Wenbin Li, Ning Zhong, Y.Y. Yao, Jiming Liu, and Chunnian Liu</i>	
Ontology Based Web Mining for Information Gathering	406
<i>Yuefeng Li and Ning Zhong</i>	
A Reasonable Rough Approximation for Clustering Web Users	428
<i>Duoqian Miao, Min Chen, Zhihua Wei, and Qiguo Duan</i>	
E-Business Intelligence Via MCMP-Based Data Mining Methods	443
<i>Yi Peng, Yong Shi, Xingsen Li, Zhengxin Chen, and Gang Kou</i>	
Intelligence Metasynthesis in Building Business Intelligence Systems	454
<i>Longbing Cao, Chengqi Zhang, Dan Luo, and Ruwei Dai</i>	
Risk Mining in Medicine: Application of Data Mining to Medical Risk Management	471
<i>Shusaku Tsumoto, Yuko Tsumoto, Kimiko Matsuoka, and Shigeki Yokoyama</i>	
Using Cryptography for Privacy Protection in Data Mining Systems	494
<i>Justin Zhan</i>	
Author Index	515

Web Intelligence Meets Brain Informatics

Ning Zhong^{1,2}, Jiming Liu^{1,3}, Yiyu Yao^{1,4}, Jinglong Wu^{1,5}, Shengfu Lu¹,
Yulin Qin^{1,6}, Kuncheng Li^{1,7}, and Benjamin Wah^{1,8}

¹ The International WIC Institute, Beijing University of Technology, China

² Dept of Life Science and Informatics, Maebashi Institute of Technology, Japan

³ Dept of Computer Science, Hong Kong Baptist University, Hong Kong

⁴ Dept of Computer Science, University of Regina, Canada

⁵ Dept of Intelligent Mechanical Systems, Kagawa University, Japan

⁶ Dept of Psychology, Carnegie Mellon University, USA

⁷ Xuanwu Hospital, Capital University of Medical Sciences, China

⁸ University of Illinois at Urbana-Champaign, USA

Abstract. In this chapter, we outline a vision of Web Intelligence (WI) research from the viewpoint of Brain Informatics (BI), a new interdisciplinary field that systematically studies the mechanisms of human information processing from both the macro and micro viewpoints by combining experimental cognitive neuroscience with advanced information technology. BI studies human brain from the viewpoint of informatics (i.e., human brain is an information processing system) and uses informatics (i.e., WI centric information technology) to support brain science study. Advances in instrumentation, e.g., based on fMRI and information technologies offer more opportunities for research in both Web intelligence and brain sciences. Further understanding of human intelligence through brain sciences fosters innovative Web intelligence research and development. WI portal techniques provide a powerful new platform for brain sciences. The synergy between WI and BI advances our ways of analyzing and understanding of data, knowledge, intelligence, and wisdom, as well as their interrelationships, organizations, and creation processes. Web intelligence is becoming a central field that revolutionizes information technologies and artificial intelligence to achieve human-level Web intelligence.

1 Introduction

The term “Web Intelligence (WI)” was first introduced in 2000 [88]. As a new field of study, it presents excellent opportunities and challenges for the research and development of new generations of Web-based information processing technology, as well as for exploiting Web-based advanced applications [38,91,93]. In a previous paper [76], we discussed several perspectives of WI research:

WI may be viewed as applying results from existing disciplines (e.g., Artificial Intelligence (AI) and Information Technology (IT)) to a totally new domain - the World Wide Web (the Web for short); WI may be considered as an enhancement or an extension of AI and IT; WI introduces new problems and challenges to the established disciplines.

WI has been recognized gradually as a new research field on studying intelligence on the Web and intelligence for the Web.

Although WI related topics have been investigated separately in several existing disciplines, such as AI, Cognitive Science, and Neuroscience, there is a lack of a unified framework so that intelligence can be systematically studied for developing *human-level* Web intelligence. Brain Informatics (BI) is an emerging interdisciplinary field to systematically investigate human information processing mechanisms from both macro and micro points of view, by cooperatively using experimental, computational, cognitive neuroscience, and advanced WI centric information technology. It attempts to understand human intelligence in depth, towards a holistic view at a long-term, global vision to understand the principles and mechanisms of human information processing system (HIPS). The main objective of this chapter is to outline such a unified framework by examining what happens when *WI meets BI*. This leads to a new brain informatics perspective of WI research.

As more detailed blueprints and issues of WI are being evolved and specified [38,76,93,100], it becomes evident that one of the fundamental goals of WI research is to understand and develop wisdom Web based intelligent systems. Such systems integrate all human-level capabilities such as real-time response, robustness, autonomous interaction with their environment, communication in natural language, commonsense reasoning, planning, learning, discovery and creativity.

Turing gave the first scientific discussion of human-level machine intelligence [71]. Newell and Simon pioneered studies on programming computers for general intelligence [46]. McCarthy argued that reaching human-level AI requires programs that deal with the commonsense informative situation, in which the phenomena to be taken into account in achieving a goal are not fixed in advance [42]. Laird and Lent argued that interactive computer games are the killer application for human-level AI research, because they can provide the environments for research on the right kinds of problems that lead to the type of incremental and integrative research needed to achieve human-level AI [28].

In this chapter, we argue that human-level intelligence may be achieved by the combination of WI and BI. While the Web and the Web-based intelligent systems provide the necessary infrastructure for supporting BI research, as well as testbeds and applications of BI, BI research provides foundations to WI research. The rest of the paper is organized as follows. Section 2 details a new perspective of WI research. Section 3 examines how studies in two of the most fundamental WI related research areas, namely Autonomy Oriented Computing (AOC) and Granular Computing (GrC), interplay with those in BI. Section 4 describes several high-impact *WI meets BI* research topics. Finally, Section 5 gives concluding remarks.

2 A Brain Informatics Perspective of WI Research

There are urgent needs and great benefits of combining WI and BI research. Fundamental issues in both fields need to be investigated and integrated systematically in order to materialize those benefits.

2.1 What Is Brain Informatics?

Brain Informatics (BI) is an emerging interdisciplinary field to study human information processing mechanism systematically from both macro and micro points of view by cooperatively using experimental, computational, cognitive neuroscience and advanced WI centric information technology. It attempts to understand human intelligence in depth, towards a holistic view at a long-term, global vision to understand the principles and mechanisms of human information processing system (HIPS), with respect to functions from perception to thinking, such as multi-perception, attention, memory, language, computation, heuristic search, reasoning, planning, decision-making, problem-solving, learning, discovery and creativity. BI can be regarded as brain science in WI centric IT age [98,99]. BI is proposing to study human brain from the viewpoint of informatics (i.e., human brain is an information processing system) and use informatics (i.e., WI centric information technology) to support brain science study.

Figure 1 shows the relationship between BI and other brain science related disciplines as well as the WI centric IT. On one hand, although brain sciences have been studied from different disciplines such as cognitive science and neuroscience, BI represents a potentially revolutionary shift in the way that research is undertaken. It attempts to capture new forms of collaborative and interdisciplinary work. In this vision, new kinds of BI methods and global research communities will emerge, through infrastructure on the wisdom Web and knowledge grids that enables high speed and distributed, large-scale analysis and computations, and radically new ways of sharing data/knowledge. On the other hand, some of these lessons in cognitive science and neuroscience are applicable to novel technological developments in BI, yet others may need to be enhanced or transformed in order to manage and account for the complex and possibly more innovative practices of sharing data/knowledge that are made technically possible by the wisdom Web and knowledge grids [37,38,98].

2.2 Key Research Topics of Brain Informatics

In order to study BI systematically and give a global view to answer what is brain informatics, we list several major subtopics in each research area below, which is an extensional description of BI research.

- Thinking centric investigation of HIPS:
 - Human deductive/inductive reasoning mechanism for understanding the principle of human reasoning and problem solving;
 - Human learning mechanism for acquiring personalized student models in an interactive learning process dynamically and naturally.
- Perception centric investigation of HIPS:
 - Human multi-perception mechanism;
 - Auditory, visual and tactile information processing.
- Modeling human brain information processing mechanism:
 - Neuro-mechanism of HIPS;

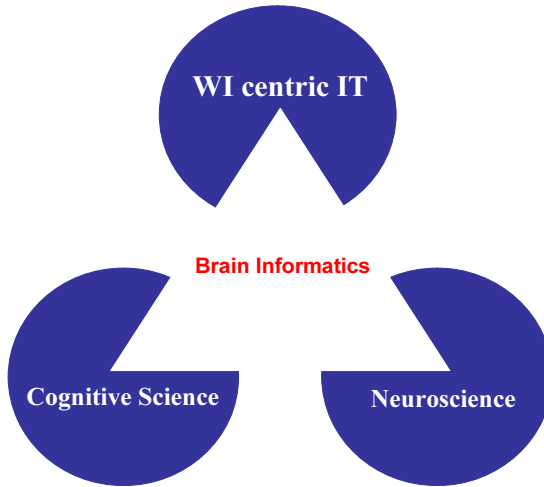


Fig. 1. The relationship between Brain Informatics and other brain science related disciplines as well as the WI centric IT

- Mathematical models of HIPS;
- Cognitive and computational models of HIPS.
- Information technologies for management and use of human brain data:
 - Human brain data collection, pre-processing, management, and analysis;
 - Multi-media human brain data mining and reasoning;
 - Databasing the brain and constructing data brain models;
 - Developing brain data grid and brain research support portals.

As a crucial step in understanding human intelligence, we must fully examine the mechanisms in which the human brain operates. The existing results, as reported over the last few decades about human information processing mechanism, are greatly related to progress of measurement and analysis technologies. Various non-invasive brain functional measurements are possible recently, such as fMRI and EEG. If these measurement data are analyzed systematically, the relationship between a state and an activity part will become clear. Furthermore, it is useful to discover more advanced human cognitive models based on such measurement and analysis. New instrumentation and new data analysis methods are causing a revolution in both AI and brain sciences [45,67].

In summary, BI emphasizes on a *systematic* approach for investigating human information processing mechanisms, including measuring, collecting, modeling, transforming, managing, mining, interpreting, and explaining multiple human brain data obtained from various cognitive experiments by using powerful equipments, such as fMRI and EEG. Human brain is regarded as an information processing system. A *systematic* study includes the investigation of human thinking centric mechanisms, the design of cognitive experiments, human brain data management, and human brain data analysis. Multi-aspect analysis in multiple

human brain data sources based on a conceptual data model of human brain is an important methodology in BI.

2.3 “WI Meets BI” in Principle

As pointed out by McCarthy [42], if we understood enough about how the human intellect works, we could simulate it. However, so far we did not have sufficient ability to observe ourselves or others to understand directly how our intellects work. Understanding the human brain well enough to imitate its function requires experimental and theoretical success in cognitive science and neuroscience.

Neuroscience, the study of the brain and nervous system, is beginning to have direct measurement and observation of ourselves or others to understand directly how our intellects work. These measurements and observations are, in turn, challenging our understanding of the relation between mind and action, leading to new theoretical constructs and calling old ones into question. New instrumentation (fMRI etc.) and advanced information technologies are causing an impending revolution in WI and brain sciences. This revolution is bi-directional:

- **WI for BI:** WI based technologies (e.g., the wisdom Web, data mining, multi-agent, and data/knowledge grids) will provide a new powerful platform for brain sciences;
- **BI for WI:** New understanding and discovery of the human intelligence models in brain sciences (e.g., cognitive science, neuroscience, and brain informatics) will yield new WI research and development.

The first aspect means that WI technologies provide an agent based multi-database mining grid architecture on the wisdom Web for building a brain-informatics portal [95,98]. A conceptual model with three levels of workflows, corresponding to a grid with three layers, namely data-grid, mining-grid, and knowledge-grid, respectively, is utilized to manage, represent, integrate, analyze, and utilize the information coming from multiple, huge data and knowledge sources. Furthermore, wisdom Web based computing provides not only a medium for seamless information exchange and knowledge sharing but also a type of man-made resource for sustainable knowledge creation, and scientific and social evolution. The new generation Web will enable humans to gain practical *wisdoms* of living, working, and playing, in addition to information search and knowledge queries. The wisdom Web relies on multi-layer knowledge grids based service agencies that self-organize, learn, and evolve their courses of actions, in order to perform service tasks, as well as their identities and interrelationships in communities [37,38,91]. These service agencies cooperate and compete among themselves in order to optimize their own as well as others’ resources and utilities. The proposed methodology attempts to change the perspective of cognitive/brain scientists from a single type of experimental data analysis towards a holistic view at a long-term, global field of vision.

The second aspect of the new perspective on WI means that the new generation of WI research and development needs to understand multi-nature of

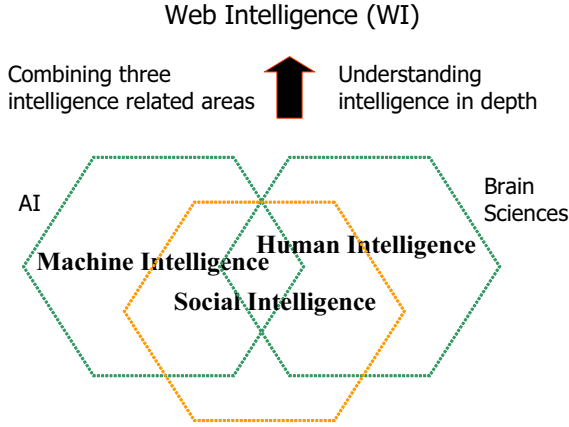


Fig. 2. The relationship between WI and other three intelligence related research areas

intelligence in depth, by studying together the three intelligence related research areas: machine intelligence, human intelligence, and social intelligence, as illustrated in Fig. 2, towards discovery of new cognitive and computational models for developing truly human-level Web intelligence.

Machine intelligence (or Artificial Intelligence) has been mainly studied as computer based technologies. Various computational models and knowledge based systems have been developed for automated reasoning and learning. Unfortunately, most of such models and systems will not work well when dealing with large-scale, global, distributed, multiple information sources on the Web. BI presents new opportunities as well as challenges to solve the difficulty, that is, to find a good way of bridging the huge gap between classical automated reasoning/learning and biologically plausible reasoning/learning.

On the other hand, human intelligence is concerned with the nature of intelligence towards our understanding of intelligence. The capabilities of human intelligence can be broadly divided into two main aspects: perception and thinking. So far, the main disciplines with respect to human intelligence are cognitive science that mainly focuses on studying mind and behavior based cognitive models of intelligence, as well as neuroscience that mainly focuses on studying brain and biological models of intelligence. In cognitive neuroscience, although many advanced results with respect to “perception oriented” study have been obtained, only a few of preliminary, separated studies with respect to “thinking oriented” and/or a more whole information process have been reported [17]. Study of HIPS from the *WI meets BI* point of view should be “thinking oriented”.

Furthermore, social intelligence needs a combination of machine intelligence and human intelligence for establishing social networks that contain communities of people, organizations, or other social entities [91]. One of the important implications of the Web is that it introduces a social network where a set of people (or organizations or other social entities) are connected. The connections are based on a set of social relationships, such as friendship, co-working or information

exchange with common interests. In other words, it is a Web-supported social network or virtual community. In this sense, the study of WI is of social network intelligence (social intelligence for short).

Figure 3 shows the relationship between WI and BI research. The synergy between WI with BI will yield profound advances in our analyzing and understanding of the mechanism of data, knowledge, intelligence and wisdom, as well as their relationship, organization and creation process. It means that WI fundamentals and technologies will be studied as a central topic and in a unique way. It will change the nature of information technologies in general and artificial intelligence in particular, working towards a new understanding and development of human-level Web intelligence.

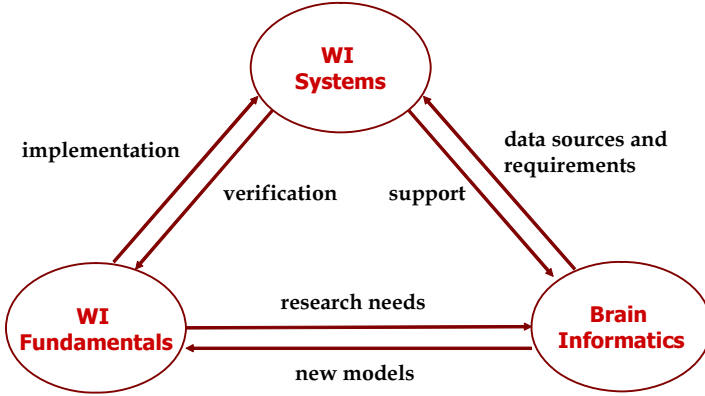


Fig. 3. The relationship between WI and BI research

3 “WI Meets BI” in Fundamental Research

Making a detailed plan for *WI meets BI* research (including the detailed systematic fMRI/EEG experiment plan) is one of the most urgent tasks. Unfortunately, there is a gap between WI and brain science research. Finding a better way to bridge this gap is the key to success in *WI meets BI*. Figure 4 provides a schematic diagram that depicts three facets of WI related research and development, i.e., fundamentals, technologies, and applications. In what follows, we will take an in-depth look at the fundamental facet, and in particular, describe how studies in two of the most fundamental WI related research areas, namely Autonomy Oriented Computing (AOC) and Granular Computing (GrC), will interplay with those in BI, hence narrowing such a gap.

3.1 The AOC Dimension

Autonomy Oriented Computing (AOC): What and How? The goals of AOC [40,34] are: (1) to discover and understand the working mechanisms that

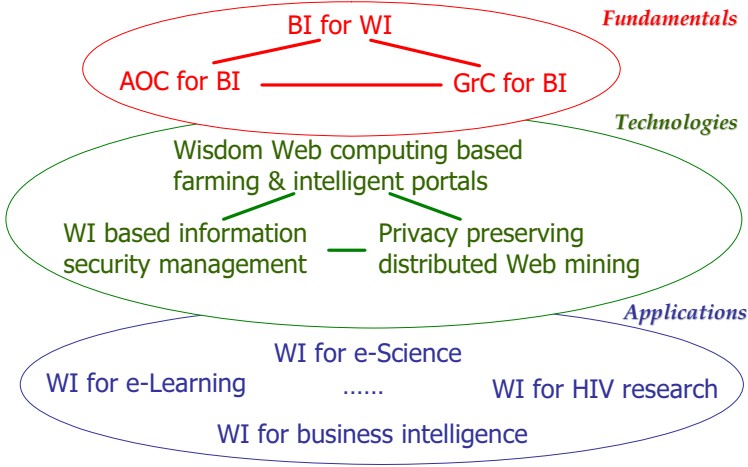


Fig. 4. Three facets of WI related research and development

lead to emergent behavior in complex systems (e.g., the dynamics of WWW, social networks, and immune systems), and (2) to design and develop scalable solutions to large-scale, distributed computational problems (e.g., distributed optimization in sensor network data routing, robot world modeling, and dynamic grid resource allocation).

In doing so, AOC emphasizes the modeling and characterization of *autonomous* entities in the systems or problems at hand, and thereafter, creates a computational system that enables the modeled entities to locally interact following certain nature or real-world inspired rules. An AOC system is an open, non-equilibrium system in which autonomous entities are free to react to external stimuli and to actively carry out information exchanges or utility updates based on the predefined behavioral rules. As a result, certain behavior of the entities and/or their effects will be nonlinearly aggregated and amplified as opposed to others. This process is known as *self-organization*, which can be regarded as the *core* of AOC.

The Distinct Characteristics of AOC. In complex systems modeling, AOC enables the process of self-organization to effectively recap certain empirically observed emergent behavior and hence provide a model for its working mechanism [36]. In complex problem solving, AOC utilizes self-organization to efficiently converge to a *desired* solution state [35,33]. The benefits of AOC in either case can be summarized as follows:

1. AOC lends itself very well for *natural formulation*, since many complex systems or problems at hand are locally-interacting, autonomous, and distributed in nature;
2. AOC provides an *easy-to-implement* computing or programming means, as autonomous entities can readily be developed and deployed;

3. AOC offers *scalable performances* in both systems modeling and problem solving, as the spirit of self-organization lies in the fact that the larger the scale, the more effective and efficient the process should become.

From the above-mentioned characteristics and/or benefits, we can note that AOC differs from conventional agent-based computing, as the latter is aimed primarily at providing a distributed software or system development methodology that is based on the models of rational agents. At the same time, it also differs from conventional agent-based simulation, as the goals of AOC are made both more explicit and broader.

The Basic Elements of AOC. As may be noted from the above, there are several basic elements essential to an AOC system, some of which are outlined below.

1. **Environment:** As one of the main components in an AOC system, the environment serves not only as the domain for autonomous entities, but also as an indirect communication medium among entities.
2. **Autonomous Entities:** An autonomous entity reacts to other entities as well as the environment. It modifies its exert changes to the environment, and/or affect other entities. Central to an autonomous entity is its *local behavior* and *behavioral rules* that govern how it should act or react to the information
3. **Interactions:** The emergent behavior of an AOC system originates from its internal interactions. Generally speaking, there are two types of interactions, namely, *interactions between entities and their environment* and *interactions among entities*. Different AOC systems may have different ways of direct or indirect interactions among their entities. Indirect interactions are implemented through the communication medium role of the environment.

For further readings on AOC, e.g., comprehensive surveys of related work, formal descriptions of the AOC approaches and formulations, and detailed discussions of examples, please refer to [40].

AOC for BI. Human brain is one of the most complex systems that we have ever encountered. Central to human consciousness, cognition, and emotion is the emergence of mental complex behavior based on the self-organization of neurons and neural activities; some of the emergent patterns with respect to certain specific conditions or tasks can be experimentally measured and visualized using brain imaging technologies.

Many interesting observations have been made regarding the complex nature of how we think, and can be found in [55].

The goals of BI are twofold: (1) to measure, characterize, and predict how humans think and reason, and in turn (2) to provide new insights into how to develop and build real-world, large-scale AI systems, e.g., in the context of WI. Both goals complement well with those of AOC as mentioned above. This can be viewed in the following two important aspects.

AOC Informs BI. AOC provides an ideal means for the advancement of BI. AOC can be used to hypothesize and model the underlying interactions and tempospatial interrelationships among different regions and/or functions in the human brain with respect to performing various cognitive tasks, such as learning, reasoning, and decision making. In other words, AOC can *inform* neuroscience and BI studies as to what areas, interactions, information exchange mechanisms, and/or cognitive strategies to look into, following the AOC’s *white-box* modeling results.

For instance, based on the AOC-based approach to regularity characterization [36], it is now possible to explicitly model WWW users, including their interest profiles, motivations, and navigation strategies or decision-making processes, and to study how those attributes lead to different emergent regularities on the WWW. The approach views users as *information foraging entities* categorized into three different categories: *recurrent*, *rational*, and *random users*. Users of different categories will apply different sets of navigation behavior. The AOC-based results match well with the known empirical results; in other words, they offer a white-box model of user navigation behavior that can be used to explain the self-organized WWW regularities. As a natural extension, we may further validate the modeled user navigation behavior through BI studies, so as to confirm the AOC findings.

BI Guides AOC. Findings from BI studies can, on the other hand, provide a guidance to the design and formulation of new AOC models and methods, e.g., for identifying the behavioral rules of underlying autonomous entities/components and their self-organizing processes [51].

For instance, one of the interesting AOC studies in the context of WI is to develop autonomous search and reasoning mechanisms, whose performance can scale well in light of large-scale, distributed, and heterogeneous content resources, as in the case of WWW. As mentioned above, the core of the AOC-based approach is to create a non-linear process of self-organization that can make the search or reasoning ‘behavior’ not only possible but also *scalable*. In this regard, a better understanding of human cue-based association, extraction and abstraction behavior would be helpful in either developing new AOC-based search and reasoning mechanisms or building a basis for further AOC-based characterization.

How Are Things Connected? *Nature* has unfolded many interesting regularities and properties, some of which are quite ubiquitous [26]. A well-observed example is self-organized criticality, as described in Bak’s *How Nature Works* [4]. At the same time, there also exist other things in nature, about which we are so fascinated but still know very little. The most obvious example is our *brain*; references on some of the existing findings and hypotheses can be found in Pinker’s *How the Mind Works* [55].

As a nature inspired computing paradigm, autonomy oriented computing *naturally* brings the two together. On one hand, AOC explicitly draws on and utilizes the metaphors of autonomy as offered by nature and the natural law of self-organization as the advocated computational means for complex systems

modeling and problem solving. On the other hand, as mentioned in the preceding subsections, AOC can readily play an important role in unveiling and explaining human brain's mental regularities.

3.2 The GrC Dimension

Granular computing (GrC) is a multi-disciplinary study of human-centered and knowledge-intensive problem solving at multiple levels of granularity [5,6,25,32,48,53,54,56,57,63,75,89]. A unified framework of granular computing can be established by focusing on the philosophical, methodological and computational perspectives [77,81,82]. This framework may provide insights into our understanding of the working mechanism of human information processing and would eventually lead to an abstract and conceptual model of the brain.

Overview of Granular Computing. Granular computing deals with problem solving in terms of the parts and the whole. Parts may be viewed as granules situated in the context of a whole. Furthermore, a part can also be a whole consisting of smaller parts. Thus, granular computing focuses on multiple hierarchical structures of granules [5,10,77,81,82,86].

Hierarchical organizations and structures are abundant in the real world. They can be found in many natural, social, and man-made systems [1,2,52,62,64]. Human perception and understanding of the real world depends, to a large extent, on such nested and hierarchical structures. We perceive and represent the world using various grain sizes, and abstract only those things that serve our present interests [23,41,62,83,84]. The ability to conceptualize the world at different granularities and to switch among these granularities is fundamental to human intelligence and flexibility [23].

Granular computing explores the hierarchical structures for an understanding of human intelligence in problem solving and apply it for the design and implementation of intelligent information processing systems. Granular computing may be studied based on the following three interrelated perspectives [81]:

- Philosophical perspective: structured thinking;
- Methodological perspective: structured problem solving;
- Computational perspective: structured information processing.

The multiple levels structures, mathematically defined by a partial ordering, are the central notion that links the three perspectives. In a nutshell, granular computing studies ways of thinking, information processing and computing using structures.

Importance of a Conceptual Model of the Brain. It is a common practice to compare the a human and a computer in order to gain an understanding of the one with the aid of the other. For example, one may draw many correspondences between them, including CPU to the brain, input/output devices to human perceptive organs, and memory to memory. At even lower levels, logic gates are compared to neurons and wires are compared to neuron connections. While such

an understanding is sufficient for certain purposes, it may be inadequate for the understanding of natural intelligence emerged from the human brain.

A conceptual brain model is perhaps still a less studied and understood problem in brain informatics research. In this respect, the von Neumann architecture of computers may shed some light. We can easily convince ourselves that the von Neumann architecture is the foundation of modern day computer. If a different conceptual model was used, we would have a much different type of computers today. Furthermore, without the von Neumann architecture, it will be much more difficult to obtain an understanding of a computer. On the one hand, we may have equipment that allows us to measure physical properties and to observe the behavior of a computer. On the other hand, an understanding of a computer cannot easily be obtained from such measurements and observations. We need conceptual models, like the von Neumann architecture, to put all puzzle pieces together.

The study of human brain is in a similar situation. On the one hand, we have achieved extensive results in neural science and cognitive science. We have detailed description and in-depth understanding of the brain at the neuron level and the cortex region level. The new instruments, such as fMRI, make the observation of the brain more accurate and detailed. On the other hand, there is still a lack of a commonly agreed conceptual model that enables us to see the high level working principles of the brain.

Some researchers have in fact made very promising progress towards conceptual models of the brain. For example, Hawkins uses the notion of a cortical hierarchy for deriving a memory-prediction framework for explaining intelligence [22]. A conceptual model of cortex is proposed by highlighting its hierarchical connectivity and information flow up and down the hierarchy [22]. To some extent, this conceptual model is closely related to the study of granular computing [82].

An urgent task of brain informatics research is therefore to build a conceptual framework of the brain. As a minimum requirement, this framework must cover information processing in the abstract, in the brain, and in the machine. By studying information processing in the abstract, we may be able to build conceptual models that explain how the brain works; the understanding of the brain and its information processing mechanism will provide new insights into how to implement intelligence and information processing in machines.

The Relevance of Granular Computing. The basic principles of granular computing, namely understanding and working with multiple levels of granularity, may capture the essential features of human problem solving and human intelligence. Based on such principles, one may study brain informatics from multiple views and at multiple levels in each view.

The power of evolution. The human brain may be viewed as a natural system in the long history of evolution. Some of the functions of the brain are to perceive, analyze, synthesize, store and retrieve information about its environments, as well as using such information in decision making. As a result, it is reasonable to assume that the brain is good at processing structures and patterns that are

abundant in nature. Furthermore, the brain is able to deal with structures at differing levels of granularity.

A very common and dominant structure is a hierarchical structure, mathematically defined by a partial order. The levels in the structure show the differing degrees of abstraction, control, orderness, structuredness, details, and so on. The brain has amazingly capacity to deal with such structures. Some researchers have suggested hierarchical modeling of the brain. The cortex hierarchical organization is perhaps generally accepted for human vision [41]. Hawkins argues that this view may be more generally applicable [22]. According to him, the human brain can be interpreted as a hierarchical structure that stores a model of the hierarchical structure of the real world. In other words, the real world's nested structure is mirrored by the nested structure of our cortex.

As an emerging multi-disciplinary study, granular computing deals with hierarchical structures with multiple levels of granularity. This makes it very relevant to the study of brain informatics. If we believe in the power of evolution, we need to accept that the working principles of the brain can be related to the working principles of the real world. This is the foundation of our search for a conceptual model of the brain based on the ideas of granular computing.

Roles of languages in information processing and natural intelligence. Concepts are the basic unit of human thoughts and play a central role in our understanding of the world. Languages, either special purpose languages such as logic, mathematics, and programming languages or general purpose natural languages, are developed as ways to represent our thoughts and to communicate with each other. Human information processing and natural intelligence depend crucially on our language ability.

Hierarchical structures of a language can be observed from two aspects. In one aspect, a language itself consist of a set of components that can be arranged hierarchically. For example, in English we have letters, words, phrases, sentences, paragraphs, and articles. They are nested within each other to form a hierarchy. In the other aspect, we can easily observe the hierarchical structure of things described by a language. Similar to Hawkins' argument, the hierarchical structures of a language is determined by the things it is intended to describe. That is, the structures of language again mirror the structures of the real world.

It seems reasonable to start from an assumption that human brain processes information by exploring the hierarchical structures. For example, when reading an article, our eyes may see a text a sequence of letters at a lower level, our short term memory may view it as a sequence of words or sentences at a higher level, and our long term memory may view it as a set of ideas at an even higher level. Such a sequence of abstraction may provide clues on human information processing and intelligence. Human vision may be similarly interpreted, if we view pictures and images as things described by a different type of language.

Data, Information, Knowledge and Wisdom hierarchy. From the practical side, results from the study of granular computing for brain informatics may be applied to design and implement better Web-based intelligent information systems.

In computer science, we have studied many approaches for representing and processing information and knowledge. Typically, knowledge representation methods must deal with the issues of granularity. Data, information, knowledge, and wisdom may be viewed as the descriptions of the same world at different levels of granularity. Since we have the same underlying world, it is possible to transform data into information, information into knowledge, and knowledge into wisdom. The brain has the ability to process and work with all those types. While data is more related to our perceived sensory level ingredients, others are related to synthesized results.

The data, information, knowledge, and wisdom hierarchy also captures the evolution of the World Wide Web, change from data Web, to information Web, to knowledge Web, and to wisdom Web [38,80,91]. With a deeper understanding of information processing in the abstract, in the brain and in machine, we may be moving closer to the wisdom Web [38,91].

There is evidence supporting our views presented here. However, we are still far away from a real framework. The main objective is to stimulate discussion and further research. Although the details may be refined or change with more understanding of the problem, research on such a conceptual model is perhaps on the right track.

4 Impending “*WI meets BI*” Research

We briefly report our high-impact research projects to demonstrate the potentials of combining WI and BI.

4.1 Reasoning Centric, Thinking Oriented Studies of HIPS

Figure 5 presents a schematic diagram of reasoning centric, thinking oriented functions and their relationships. It represents an attempt for a systematic examination of human thinking centric mechanisms. The core issue is to investigate human deduction, induction, and abduction related reasoning mechanism, including commonsense and non-monotonic reasoning, as shown in the central of Figure 5. Heuristic search, autonomy (related to AOC), information granulation (related to GrC), attention and memory are some component functions to implement human reasoning, as well as emotion, uncertainty and stability are some interesting characteristics, which should be investigated with respect to human reasoning, as illustrated in the middle circle of this figure. Furthermore, decision-making, problem-solving, planning, computation, language, learning, discovery and creativity are the major human thinking related functions, which will be studied systematically, as illustrated outside the middle circle of this figure.

Research in cognitive neuroscience, including brain lesion and brain imaging studies of normal subjects, have made some preliminary progresses for the brain mechanism of deductive/inductive reasoning. Patients studies can provide some rough localization result of inductive reasoning. For example, Gazzaniga and colleagues administered simple inductive reasoning tasks to split-brain patients

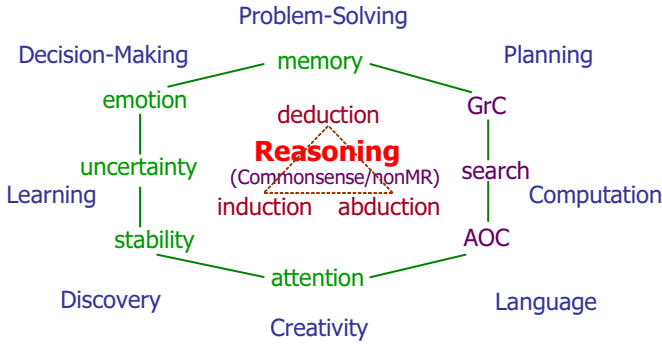


Fig. 5. A systematic illustration of reasoning centric, thinking oriented functions and their relationships (GrC: Granular Computing [78]; AOC: Autonomy Oriented Computing [40]; nonMR: non-monotonic reasoning)

and concluded that reasoning is a left hemisphere phenomenon [15]. Furthermore, Gazzaniga postulated a “left brain interpreter” [16]. Varley and colleagues suggested that induction may recruit dorsolateral and medial aspects of the prefrontal cortex [72]. However, the results of patients may be insufficient to tell us more details about the spatiotemporal characteristics of inductive reasoning. Nowadays, brain imaging of normal subjects is the most popular way for researchers. Goel and colleagues used PET to examine human deductive/inductive reasoning when subjects performed sentential inductive tasks for the first time [18]. With the new event-related fMRI technique, Goel and colleagues performed similar experiments, and indicated that left dorsolateral prefrontal gyrus showed a greater activity during induction than deduction [20]. Christoff and colleagues hypothesized that the process of relational integration is a component process of inductive reasoning [11]. Some adapted RPM (Raven’s Progressive Matrices) tasks were used in their experiment, and the results indicated that the process of relational integration is specifically associated with the bilateral rostrolateral PFC (RLPFC, BA 10) and the right dorsolateral PFC (DLPFC, BA 9 and 46). Goel and colleagues adopted novel animals as induction stimuli, and the fMRI results showed that rule inference was specifically associated with the bilateral hippocampal activation while the task characterized by difficulty interaction was associated with activation in the right lateral orbital prefrontal cortex [19].

However, there are limited experimental studies in cognitive neuroscience. In particular, the aforementioned brain imaging studies mainly have two flaws. Firstly, for sentential tasks and RPM tasks, individual difference induced by background knowledge of tasks is significant, i.e., the effects of background knowledge cannot be well counterbalanced. Secondly, all these studies mainly employed PET or fMRI technique to explore the functional dissociation among activated brain regions. Due to their low temporal resolution, we still know little about the time course of inductive reasoning process in human brain.

Our purpose is to understand activities of HIPS by investigations in the following two levels:

- investigating the spatiotemporal features and flow of HIPS, based on functional relationships between activated areas of human brain for each given task;
- investigating neural structures and neurobiological processes related to the activated areas [61].

More specifically, at the current stage, we want to understand:

- How a particular part (one or more areas) of the brain operates at a specific point of time;
- How the operations change over time;
- How the activated areas are linked, indexed, and navigated functionally?
- How the activated areas work cooperatively to implement a whole information processing functionality;
- How a cognitive process is supported by neurobiological processes;
- What are individual differences in performance.

The key question is “can we find a new cognitive model for developing human-level Web based network reasoning and problem solving?”.

Human and Web Problem Solving and Reasoning. As an example of thinking oriented studies for WI meets BI, we describe a study of human and Web problem solving and reasoning in a unified way. The objective is to develop Web based problem solving system with human-level capabilities.

Problem-solving is one of main capabilities of human intelligence and has been studied in both cognitive science and AI, where it is addressed in conjunction with reasoning centric cognitive functions such as attention, control, heuristic search, reasoning, learning, and so on, using a logic based symbolic and/or connectionist approach. Logic based problem-solving may be viewed as theoretic models that are mathematical systems with no real time and memory constraints. Web-based problem-solving systems need real-time response and deal with global, multiple, huge, distributed information sources.

A more concrete issue of WI is the development and application of a Web-based problem-solving system for portal-centralized, adaptable Web services [38,69,91,93]. The core of such a system rests on the Problem Solver Markup Language (PSML) and PSML-based distributed Web inference engines for network reasoning. PSML should support at least the following essential functions:

- Supporting complex adaptive, distributed problem solving;
- Performing automatic reasoning on the Web by incorporating globally distributed contents and meta-knowledge, automatically collected and transformed from the semantic Web and social networks, with locally operational knowledge-data bases;
- Representing and organizing multiple, large-scale knowledge-data sources for distributed network reasoning;

- Combining multiple reasoning methods in PSML representation and distributed inference engines efficiently and effectively;
- Modeling user behavior and representing/managing it as a personalized model dynamically;
- Developing the Web based reasoning and problem solving system, with a consideration of an emotional factor.

As a first step, a possible way for implementing certain distributed reasoning capabilities of the future PSML is to make use of an existing logic language coupled with agent technologies. We have demonstrated one possible implementation of such capabilities. A preliminary version of an implementation, called β -PSML, is based on the combination of OWL with Horn clauses and is able to couple global semantic Web/social networks with local information sources for solving problems in a large-scale distributed Web environment [68,69].

In order to develop a Web based problem-solving system with human-level capabilities, we need to better understand how human being does complex adaptive, distributed problem solving and reasoning, as well as how intelligence evolves for individuals and societies, over time and place [67,85,99]. Ignoring what goes on in human brain and focusing instead on behavior has been a large impediment to understand complex human adaptive, distributed problem solving and reasoning.

In the light of BI, we need to investigate specifically the following issues:

- What are the existing thinking/reasoning models in AI, cognitive science, and neuroscience?
- How to design fMRI/EEG experiments and analyze such fMRI/EEG data to understand the principles of human reasoning and problem solving in depth?
- How to build the cognitive model to understand and predict user profile and behavior?
- How to implement human-level reasoning and problem solving on the Web based portals that can serve users wisely?

As a result, the relationships between classical problem solving and reasoning and biologically plausible problem solving and reasoning need to be defined and/or elaborated.

In summary, human and Web problem solving and reasoning needs to be studied in a unified way, based on the following approach: (1) to investigate the mechanisms of human problem solving and reasoning from a BI perspective; (2) to investigate AOC based problem solving and reasoning; (3) to investigate GrC based problem solving and reasoning; (4) to develop and validate new cognitive/computational models for Web-based problem solving and reasoning, based on the results from the studies of (1)-(3).

Studying Lower and Higher Functions of HIPS. Functions of HIPS can be classified as the lower and higher functions. This classification is not a mechanistic one, but relative and from a global view. Thinking oriented study in HIPS needs to understand relationship between lower and higher functions of HIPS.

If we want to understand the mechanism of higher functions, such as reasoning, problem-solving and decision-making, we need to understand the mechanism of lower functions, such as attention, memory, vision, and their relationship. We give some research examples in order to explain the relationship between lower and higher functions, from the viewpoint of a perception oriented study.

1. **Visual, Auditory and Calculated Functions.** Human visual and auditory systems have been evolving through the long history and have attained very high performance. However, artificial systems for those are often poorer in performance despite of remarkable progress in their efficiency. In order to construct artificial systems with high performance, such as human visual and auditory systems, it is necessary to elucidate the mechanisms of human visual and auditory systems. In order to investigate the characteristics of human visual and auditory information processing, the authors have measured the function of human brain related to calculation. For the tasks of measuring the human information processing involving calculation, the characteristics of the visual and auditory information processing were measured by using the mental arithmetic problems as the visual and auditory stimuli by psychological experiment. Activated areas of human brain related to the calculation were also measured by fMRI. The measurement results clearly show the difference between the visual and auditory calculation. They provide evidence for explaining the mechanism of visual and auditory information processing and may have an impact on the construction of artificial systems [29,44].
2. **Auditory and Language Functions.** Function segregation in the left inferior frontal gyrus was investigated in an fMRI experiment using a passive listening task, in which Japanese subjects were required to passively listen to words and non-words in Japanese and English. Listening to English words and non-words activated the bilateral dorsal inferior frontal gyrus more extensively than listening to Japanese words and non-words, implying an automatic articulatory representation access for harder perception effort of the nonnative language stimuli. In both languages, word listening activated the left ventral inferior frontal gyrus more extensively than nonsense word listening, suggesting a lexical or semantic processing function. This function segregation in the left inferior frontal gyrus has also been found in studies using visual tasks [8,74].

On the other hand, from the viewpoint of thinking oriented study, we need to consider what are the relationship and difference between the perception and thinking oriented studies. Based on BI methodology, we can develop an approach to unify the perception and thinking oriented studies in some cognitive level. In other words, by a systematic design of cognitive experiments in BI methodology, the data obtained from a cognitive experiment and/or a set of cognitive experiments may be used for multi-task/purpose, including for both lower and higher functions. For example, it is possible that the experiment for investigating visual, auditory and calculated functions as mentioned above is re-designed to meet requirements of both thinking and perception oriented studies for multi-task/purpose (e.g. investigating the

mechanisms of human visual and auditory systems, computation, problem-solving, and the spatiotemporal feature and flow of HIPS).

Cognitive Architecture Meets fMRI/EEG. According to Newell [47], cognitive architecture would explain how all the components of the mind work together to generate coherent human cognition [3]. Therefore, it is also a kind of effort towards conceptual modeling of the brain. ACT-R, one of the best known, fully implemented, and free to public cognitive architecture models, is a theory on how the structure of brain achieves the function of adaptive cognition in a system level, as well as a platform to build computational models to simulate/predict human cognitive behavior, including performing complex dynamic tasks, which usually emphasize perceptual-motor components and their coordination with other cognitive components (learning and memory, reasoning, and so on) and with strong time pressure [3,65], as usually required by Web tasks.

Consistent with granular computing, ACT-R has two levels. The subsymbolic level deals with fine-graded models of learning and performance. In the symbolic level, ACT-R consists of modules. Consistent with Autonomy Oriented Computing, the processes of the interaction between the inside modules and the outside world as well as the interaction among these modules. The output of an ACT-R model is a time course for when and how long the activations of each module involved in the task are. Based on it, one can predict the performance of the subjects (such as reaction time and accuracy). Recent years, the modules in ACT-R have been mapped to brain areas, and methods have been developed to predict Blood Oxygenation Level-Depend (BOLD) effect in fMRI experiments, as shown in the chapter of ACT-R meets fMRI.

Technologies have been developed to check the synchronous neural oscillations from EEG data, which may reveal the dynamic cognitive processes in brain [73]. Thus EEG and ACT-R (with fMRI) will provide two aspects to infer the cognitive processes and their neural basis. Therefore, combining cognitive architecture, fMRI and EEG might be an appropriate approach to help us to reach the goal described in this section, investigating the spatiotemporal feature and flow of HIPS and investigating neural structures and neurobiological processes related to the activated areas. In addition to the tasks mentioned in other sub-sections, we need to systematically perform a set of behavior, fMRI and EEG experiments, and with the help of ACT-R modeling, to explore the cognitive processes and their neural bases of human interacting with the Web and to have a better understanding of WI and BI.

4.2 Studying the Full Process from fMRI/EEG Experiments to New Cognitive WI Models

Figure 6 shows a full process from designing fMRI/EEG experiments based on WI needs to discovering new cognitive WI models. It offers a systematic approach for measuring, collecting, modeling, transforming, managing, and mining multiple human brain data obtained from various cognitive experiments by using fMRI and EEG [96,98].

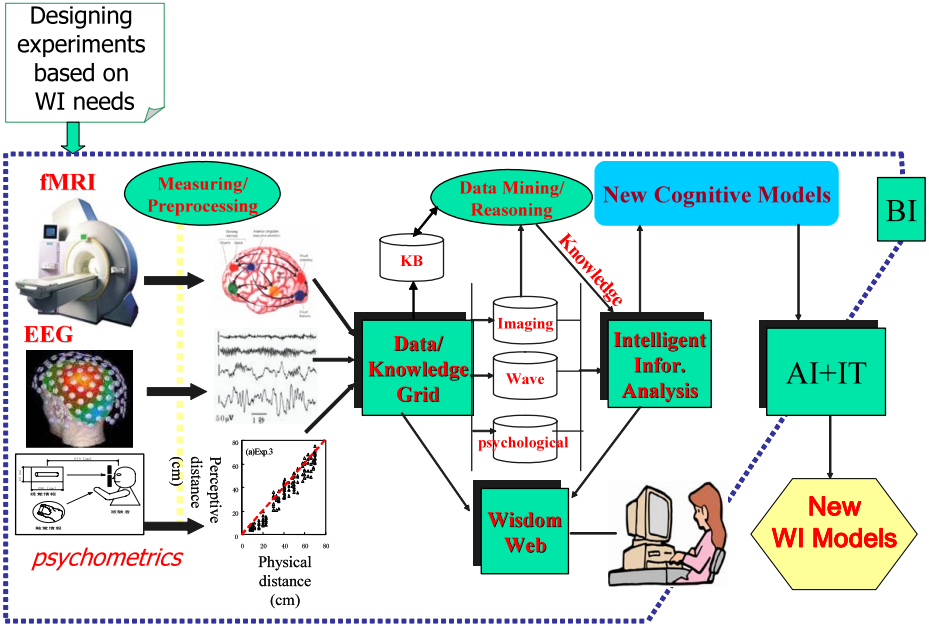


Fig. 6. From fMRI/EEG experiments to new cognitive WI models

Each of fMRI and EEG technologies has its strength and weakness from the aspects of time and space resolutions. fMRI provides images of functional brain activity to observe dynamic activity patterns within different parts of the brain for a given task. It is excellent in the space resolution, but inferior in the time resolution. On the other hand, EEG provides information about the electrical fluctuations between neurons that also characterize brain activity, and measurements of brain activity at resolutions approaching real time. In order to discover new knowledge and models of human information processing activities, not only individual data source obtained from a single measuring method, but also multiple data sources from various practical measuring methods are required.

The future of BI will be affected by the ability to do large-scale mining of fMRI and EEG brain activation data. The key issues are how to design the psychological and physiological experiments for systematic obtaining various data from HIPS, as well as how to analyze and manage such data from multiple aspects for discovering new models of HIPS. Although several human-expert centric tools such as SPM (MEDx) have been developed for cleaning, normalizing and visualizing the fMRI images, researchers have also been studying how the fMRI images can be automatically analyzed and understood by using data mining and statistical learning techniques [43,45,66,70,96]. We are concerned with how to extract significant features from multiple brain data measured by using fMRI and EEG in preparation for multi-aspect data mining that uses various data mining techniques for analyzing multiple data sources.

Building a Brain-informatics Portal. Building a brain-informatics portal is, in fact, to develop a data mining grid centric multi-layer grid system on the wisdom Web for multi-aspect data analysis. The wisdom Web [37,38] and Grid computing [9,14] have provided the ideal infrastructures, platforms, and technologies for building such a brain-informatics portal to support cognitive/brain scientists in multi-aspect analysis in multiple, large-scale data sources. We need to study experimental cognitive neuroscience, data mining, intelligent agents, data and knowledge grids, the semantic Web and wisdom Web in a unified way.

Middleware, as a new platform, is required to cope with multiple huge, distributed data sources for multi-aspect analysis in building a brain-informatics portal on the wisdom Web. It is necessary to create a grid-based, organized society of data mining agents, called a *data mining grid* on the Grid computing platform (e.g. the Globus toolkit) [14]. This means:

- developing various data mining agents for various services oriented multi-aspect data analysis;
- organizing the data mining agents into a Grid with multiple layers such as data-grid, mining-grid, and knowledge-grid, under the OGSA (Open Grid Services Architecture) that firmly aligns with service-oriented architecture and Web services, in order to understand the user's questions, transform them to data mining issues, discover the resources and information about the issues, and get a composite answer or solution;
- using a conceptual model with three levels of workflows, namely data-flow, mining-flow, and knowledge-flow, corresponding to the three-layer Grid, respectively, for managing data mining agents for multi-aspect analysis in distributed, multiple data sources and for organizing the dynamic, status-based processes of brain informatics study.

The data mining grid is made of many smaller components that are called *data mining agents*. Each agent by itself can only do some simple thing. Yet when joining these agents on the Grid, more complex tasks for brain informatics study can be carried out.

Ontologies are also used for the description and integration of multiple human brain data sources and grid-based data mining agents in data mining process planning [30,90,92]. It is necessary to provide:

- a formal, explicit specification for the integrated use of multiple human brain data sources in a semantic way;
- a conceptual representation about the types and properties of data/knowledge and data mining agents, as well as relations between data/knowledge and data mining agents;
- a vocabulary of terms and relations to model the domain and to specify how to view the data sources and how to use data mining agents;
- a common understanding of multiple human brain data sources that can be communicated among grid-based data mining agents.

A Data-Brain Model and Its Construction. The Data-Brain is a conceptual brain data model, which represents functional relationships among multiple

human brain data sources, with respect to all major aspects and capabilities of HIPS, for systematic investigation and understanding of human intelligence. The Data-Brain is helpful for understanding the principles and mechanisms of HIPS [27,98,99].

The key questions are how to obtain such data by systematic fMRI/EEG experiments, how to manage such huge multimedia data for systematic investigation and understanding of human intelligence, as well as how to analyze such data from multi-aspect and multi-level for discovering new cognitive models. A new conceptual model is needed to represent complex relationships among multiple human brain data sources, which are obtained by systematic fMRI/EEG experiments. The following supporting capabilities are requested to build such a Data-Brain:

- It is a grid-based, simulation and analysis oriented, dynamic, spatial and multimedia database;
- It deals with multiple data sources, multiple data forms, multiple levels of data granularity;
- It provides multiple views and organizations;
- It includes various methods for data analysis, simulation, visualization, as well as the corresponding knowledge and models.

Agents for data collecting, storing and retrieving are deployed on the Grid platform, like Globus, as a standard Grid service. OGSA-DAI is used to build database access applications [14,102]. The aim of OGSA-DAI is to provide the middleware glue to interface existing databases, other data resources and tools to each other in a common way based on the Open Grid Services Architecture (OGSA). This middleware is based on the GGF-defined OGSi specification and layered on top of the Globus toolkit 3 OGSi implementation (GT3 Core).

Multiple data sources are collected by various cognitive fMRI/EEG experiments, modeling and transformation. They are recorded to the corresponding databases through the Grid service on the distributed sites. The data-flow is a collection of descriptions for the dynamic relationship among multiple data sources on the data-grid. Data sources from cognitive fMRI/EEG experiments, to be collected on the data-grid, include:

- human multi-perception mechanism for studying the relevance between auditory and visual information processing;
- human deductive/inductive reasoning mechanism for understanding the principle of human reasoning and problem solving in depth;
- human computation mechanism as an example of human problem solving system;
- human decision-making mechanism from developing Web based decision-making support system with an emotional factor;
- human learning mechanism for acquiring personalized student models in an interactive learning process;
- human heuristic search in problem solving and reasoning;
- human emotion factors in higher cognitive functions.

In order to build a Data-Brain, a systematic methodology of cognitive experimental design needs to be developed, so that multiple human brain data sources obtained by fMRI/EEG experiments are interrelated and utilized for multi-purpose. Event-related experimental designs have become an important methodology in EEG/fMRI research to evaluate the high level characteristics of HIPS in the central nervous system [60]. There are, at present, two main methods called event-related potential (ERP) and event-related fMRI for event-related experimental designs. ERP is a tiny signal embedded in the ongoing EEG. By averaging the traces, investigators can extract this signal, which reflects neural activity that is specifically related cognitive events [21]. ERPs are best suited for addressing questions about the time course of cognition rather than elucidating the brain structures that produce the electrical events. ERPs also provide physiological indices of when a person decides to response, or when an error is detected. On the other hand, event-related fMRI follows the same logic as used in ERP/EEG studies and provides the spatial resolution. Thus, event-related fMRI will further allow fMRI and EEG to be combined in paradigms that are identical across methods. By using such techniques, it is now becoming possible to study the precise spatiotemporal orchestration of neuronal activity associated with perceptual and cognitive events [60], as well as systematic collection of human brain data for building a Data Brain.

Multi-aspect Human Brain Data Analysis on a Multi-layer Grid. *Multi-aspect analysis* in a multi-phase mining process is an important methodology for knowledge discovery from multiple human brain data [90]. There are two main reasons why a multi-aspect analysis approach needs to be used. First, we cannot expect to develop a single data mining algorithm for analyzing all main aspects of multiple human brain data towards a holistic understanding, due to the complexity of human brain. Various data mining agents (association, classification, clustering, peculiarity-oriented, main-fold, etc), deployed on the mining-grid, need to be cooperatively used in the multi-phase data mining process for performing multi-aspect analysis as well as multi-level conceptual abstraction and learning. Second, when performing multi-aspect analysis for complex brain informatics problems, a data mining task needs to be decomposed into sub-tasks. These sub-tasks can be solved by using one or more data mining agents that are distributed over different computers on the Grid. The decomposition problem leads us to the problem of distributed cooperative system design.

Such a methodology needs to be implemented on a wisdom Web based brain-informatics portal, which supports a multi-phase mining process based on a conceptual data model of human brain. Generally speaking, several kinds of rules and hypotheses can be mined from different data sources by multi-aspect mining. The results cannot be directly utilized to support brain scientists' research activities and applications until they are combined and refined into more general ones to form *active knowledge*, through an explanation-based reasoning process. On the other hand, from the viewpoint of applications, distributed Web inference engines under the knowledge-flow management will employ such active knowledge with various related knowledge sources together to implement knowledge

services for supporting brain scientists' research activities on the brain-informatics portal [98].

In the multi-tier architecture of the brain-informatics portal, lower levels provide middleware support for higher level applications and services, thereby opening the door to developing more complex, flexible, and effective systems. The three-level workflows are generated dynamically, based on the conditions (situations), data quality analysis, and a multi-phase mining process.

We emphasize that both pre-processing and post-processing steps are important before/after using data mining agents. In particular, informed knowledge discovery, in general, uses background knowledge obtained from experts (e.g., cognitive/brain scientists) about a domain (e.g., cognitive neuroscience) to guide a spiral discovery process with multi-phase such as pre-processing, rule mining, and post-processing, towards finding interesting and novel rules/features hidden in data. Background knowledge may be of several forms including rules already found, ontologies, taxonomic relationships, causal preconditions, ordered information, and semantic categories. Such brain-informatics related knowledge, the generated hypotheses are deployed on the knowledge-grid and the knowledge-flow is utilized to generate, evaluate, refine, and employ knowledge on the knowledge-grid for various knowledge-based reasoning [68,69,95]. From the top-down perspective, the knowledge level is also the application level with the support from both the mining level and the data level to serve cognitive/brain scientists and the portal itself updating. From the bottom-up perspective, the data level supplies the data services for the mining level, and the mining level produces new rules and hypotheses for the knowledge level to generate active knowledge.

On the other hand, from the application viewpoint, distributed Web inference engines under the knowledge-flow management will utilize such active knowledge with various related knowledge sources together to implement knowledge services for supporting cognitive/brain scientists' research activities on the brain-informatics portal [68,69,98].

For a case study, our purpose is to investigate the spatiotemporal features and flow of HIPS as mentioned in Section 4.1. In our preliminary experiments, we have observed that fMRI brain imaging data and EEG brain wave data extracted from HIPS are *peculiar* ones with respect to a specific state or the related part of a stimulus. Accordingly, we proposed a way of *peculiarity oriented mining (POM)* for knowledge discovery in multiple human brain data, without using conventional imaging processing to fMRI brain images and frequency analysis to EEG brain waves [96,98]. The proposed approach provides a new way for automatic analysis and understanding of fMRI brain images and EEG brain waves to replace human-expert centric visualization. The mining process is a multi-step one, in which various psychological experiments, physiological measurements, data cleaning, modeling, transforming, managing, and mining techniques are cooperatively employed to investigate HIPS.

Further research will include studying the neural structures of the activated areas and trying to understand how a peculiar part of the brain operates and

how it is linked functionally to individual differences in performance by combining various data mining methods with reasoning. Some lessons in cognitive neuroscience are applicable to novel technological developments in BI, yet others may need to be enhanced or transformed in order to manage and account for the complex and possibly more innovative practices of sharing, analyzing and creating data/knowledge that are made technically possible by the wisdom Web and knowledge grids [38,98,99].

Figure 7 gives the global picture of an example about how to investigate the spatiotemporal features and flow of HIPS. In the cognitive process from perception (e.g., a cognitive task by vision stimuli) to thinking (e.g., reasoning), data are collected in several event-related time points, and transformed into various forms in which POM centric multi-aspect data analysis (MDA) can be carried out efficiently and effectively. Furthermore, the results of separate analysis can be explained and combined into a whole flow.

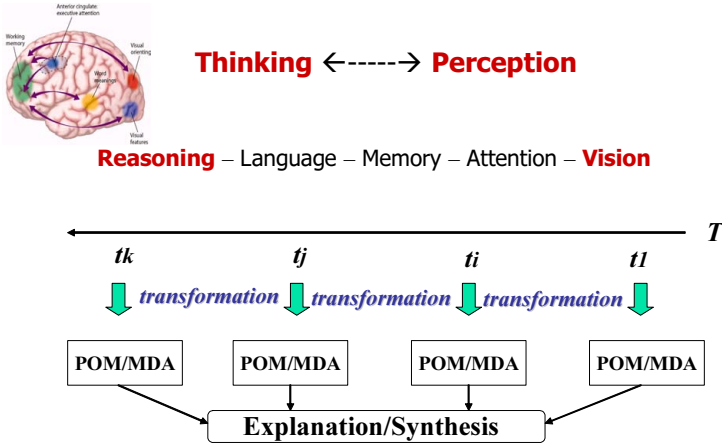


Fig. 7. Investigating the spatiotemporal features and flow of HIPS

5 Conclusion

As two related emerging fields of research, Web Intelligence and Brain Informatics mutually support each other. Their synergy will yield profound advances in the analysis and understanding of data, knowledge, intelligence and wisdom, as well as their relationships, organization and creation process. When WI meets BI, it is possible to have a unified and holistic framework for the study of machine intelligence, human intelligence, and social intelligence.

BI emphasizes on a *systematic* approach for investigating human information processing mechanisms, including measuring, collecting, modeling, transforming, managing, and mining multiple human brain data obtained from various cognitive experiments by using fMRI and EEG. In other words, human brain is regarded as an information processing system, and a *systematic* study including

the investigation of human thinking centric mechanisms, the design of cognitive experiments, data management, and data analysis. Multi-aspect analysis in multiple human brain data sources based on a data brain model is an important methodology in BI.

The proposed methodology attempts to change the perspective of cognitive/brain scientists from a single type of experimental data analysis towards a holistic view at a long-term, global field of vision to understand the principles and mechanisms of HIPS. New generations of WI research and development need to understand multi-nature of intelligence in depth. The recently designed instrumentation (fMRI etc.) and advanced IT are causing an impending revolution in both WI and BI, making it possible for us to understand and develop human-level Web intelligence.

Acknowledgments

We are grateful to all our research collaborators, assistants, and students who have, over the years, together contributed to the development of Web Intelligence (WI) and Brain Informatics (BI). We are very grateful to people who have joined or supported the WI and BI communities, members of the WIC advisory board, WIC technical committee, and WIC research centres, as well as keynote/invited speakers of IEEE/WIC/ACM WI-IAT conferences. This work is partially supported by the grant-in-aid for scientific research (No. 18300053) from the Japanese Ministry of Education, Culture, Sports, Science and Technology, and discovery Grants from Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

1. Ahl, V., Allen, T.F.H.: *Hierarchy Theory, a Vision, Vocabulary and Epistemology*. Columbia University Press (1996)
2. Allen, T.F.: A Summary of the Principles of Hierarchy Theory, (accessed March 11, 2005), <http://www.isss.org/hierarchy.htm>
3. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An Integrated Theory of the Mind. *Psychological Review* 111(4), 1036–1060 (2004)
4. Bak, P.: *How Nature Works: The Science of Self-Organised Criticality*. Copernicus Press (1996)
5. Bargiela, A., Pedrycz, W.: *Granular Computing: An Introduction*. Kluwer Academic Publishers, Boston (2002)
6. Bargiela, A., Pedrycz, W.: The Roots of Granular Computing. In: *Proceedings of 2006 IEEE International Conference on Granular Computing*, pp. 806–809 (2006)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* 284, 34–43 (2001)
8. Cai, C., Kochiyama, T., Osaka, K., Wu, J.: Lexical/Semantic Processing in Dorsal Left Inferior Frontal Gyrus. *NeuroReport* (in press, 2007)
9. Cannataro, M., Talia, D.: The Knowledge Grid. *Communications of the ACM* 46, 89–93 (2003)

10. Chen, Y.H., Yao, Y.Y.: Multiview intelligent data analysis based on granular computing. In: Proceedings of 2006 IEEE International Conference on Granular Computing, pp. 281–286 (2006)
11. Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J.K., Holyoak, K.J., Gabrieli, J.D.E.: Rostrolateral Prefrontal Cortex Involvement in Relational Integration During Reasoning. *NeuroImage* 14(5), 1136–1149 (2001)
12. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Heidelberg (2001)
13. Fensel, D., Harmelen, F.: Unifying Reasoning and Search to Web Scale. *IEEE Internet Computing* 11(2), 94–96 (2007)
14. Foster, I., Kesselman, C. (eds.): The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
15. Gazzaniga, M.S., Smylie, C.S.: Dissociation of Language and Cognition. *Brain* 107(1), 145–153 (1984)
16. Gazzaniga, M.S.: The Mind's Past. University of California Press, Berkeley, CA (1998)
17. Gazzaniga, M.S. (ed.): The Cognitive Neurosciences III. MIT Press, Cambridge (2004)
18. Goel, V., Gold, B., Kapur, S., Houle, S.: The Seats of Reason? An Imaging Study of Deductive and Inductive Reasoning. *NeuroReport* 8(5), 1305–1310 (1997)
19. Goel, V., Dolan, R.J.: Anatomical Segregation of Component Processes in an Inductive Inference Task. *Journal of Cognitive Neuroscience* 12(1), 1–10 (2000)
20. Goel, V., Dolan, R.J.: Differential Involvement of Left Prefrontal Cortex in Inductive and Deductive Reasoning. *Cognition* 93(3), B109–B121 (2004)
21. Handy, T.C.: Event-Related Potentials, A Methods Handbook. The MIT Press, Cambridge (2004)
22. Hawkins, J., Blakeslee, S.: On Intelligence. Henry Holt and Company, New York (2004)
23. Hobbs, J.R.: Granularity. In: Proceedings of the Ninth International Joint Conference on Artificial Intelligence, pp. 432–435 (1985)
24. Hu, J., Zhong, N.: Organizing Multiple Data Sources for Developing Intelligent e-Business Portals. *Data Mining and Knowledge Discovery* 12(2-3), 127–150 (2006)
25. Inuiguchi, M., Hirano, S., Tsumoto, S. (eds.): Rough Set Theory and Granular Computing. Springer, Berlin (2003)
26. Kauffman, S.: At Home in the Universe: the Search for Laws of Complexity. Oxford University Press, Oxford (1996)
27. Koslow, S.H., Subramaniam, S. (eds.): Databasing the Brain: From Data to Knowledge. Wiley, Chichester (2005)
28. Laird, J.E., van Lent, M.: Human-Level AI's Killer Application Interactive Computer Games. *AI Magazine*, 15–25 (2001)
29. Li, C., Kochiyama, T., Wu, J., Chui, D., Tsuge, T., Osaka, K.: Attention Systems and Neural Responses to Visual and Auditory Stimuli: an fMRI Study. In: Proc. 2007 IEEE/ICME International Conference on Complex Medical Engineering, pp. 1515–1519 (2007)
30. Li, Y., Zhong, N.: Mining Ontology for Automatically Acquiring Web User Information Needs. *IEEE Transactions on Knowledge and Data Engineering* 18(4), 554–568 (2006)
31. Liang, P., Zhong, N., Wu, J.L., Lu, S., Liu, J., Yao, Y.Y.: Time Dissociative Characteristics of Numerical Inductive Reasoning: Behavioral and ERP Evidence. In: Proc 2007 International Joint Conference on Neural Networks (IJCNN 2007), IEEE Press (in press, 2007)

32. Lin, T.Y., Yao, Y.Y., Zadeh, L.A. (eds.): Data Mining, Rough Sets and Granular Computing. Physica-Verlag, Heidelberg (2002)
33. Liu, J., Tang, Y.Y., Cao, Y.C.: An Evolutionary Autonomous Agents Approach to Image Feature Extraction. *IEEE Transaction on Evolutionary Computation* 1(2), 141–158 (1997)
34. Liu, J.: Autonomous Agents and Multi-Agent Systems: Explorations in Learning, Self-Organization, and Adaptive Computation. World Scientific, Singapore (2001)
35. Liu, J., Han, J., Tang, Y.Y.: Multi-agent Oriented Constraint Satisfaction. *Artificial Intelligence* 136(1), 101–144 (2002)
36. Liu, J., Zhang, S., Yang, J.: Characterizing Web Usage Regularities with Information Foraging Agents. *IEEE Transactions on Knowledge and Data Engineering* 16(5), 566–584 (2004)
37. Liu, J., Zhong, N., Yao, Y.Y., Ras, Z.W.: The Wisdom Web: New Challenges for Web Intelligence (WI). *Journal of Intelligent Information Systems* 20(1), 5–9 (2003)
38. Liu, J.: Web Intelligence (WI): What Makes Wisdom Web? In: *Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pp. 1596–1601 (2003)
39. Liu, J., Jin, X., Tang, Y.: Multi-agent Collaborative Service and Distributed Problem Solving. *Cognitive Systems Research* 5(3), 191–206 (2004)
40. Liu, J., Jin, X., Tsui, K.C.: Autonomy Oriented Computing: From Problem Solving to Complex Systems Modeling. Springer, Heidelberg (2005)
41. Marr, D.: Vision, A Computational Investigation into Human Representation and Processing of Visual Information. W.H. Freeman and Company, San Francisco (1982)
42. McCarthy, J.: Roads to Human Level AI? Keynote Talk at Beijing University of Technology, Beijing, China (September 2004)
43. Megalooikonomou, V., Herskovits, E.H.: Mining Structure-Function Associations in a Brain Image Database. In: Cios, K.J. (ed.) *Medical Data Mining and Knowledge Discovery*, pp. 153–179. Physica-Verlag (2001)
44. Mizuhara, H., Wu, J., Nishikawa, Y.: The Degree of Human Visual Attention in the Visual Search. *International Journal Artificial Life and Robotics* 4, 57–61 (2000)
45. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to Decode Cognitive States from Brain Images. *Machine Learning* 57(1-2), 145–175 (2004)
46. Newell, A., Simon, H.A.: Human Problem Solving. Prentice-Hall, Englewood Cliffs (1972)
47. Newell, A.: Unified Theories of Cognition. Harvard University Press (1990)
48. Nguyen, H.S., Skowron, A., Stepaniuk, J.: Granular Computing: A Rough Set Approach. *Computational Intelligence* 17, 514–544 (2001)
49. O'Reilly, R.C.: Biologically Based Computational Models of High-Level Cognition. *Science* 314(5796), 91–94 (2006)
50. Ohshima, M., Zhong, N., Yao, Y.Y., Liu, C.: Relational Peculiarity Oriented Mining. *Data Mining and Knowledge Discovery*, Springer (in press)
51. Van Orden, G.C., Holden, J.G., Turvey, M.T.: Self-organization of Cognitive Performance. *Journal of Experimental Psychology: General* 132, 331–350 (2003)
52. Pattee, H.H. (ed.): Hierarchy Theory, The Challenge of Complex Systems. George Braziller, New York (1973)

53. Pawlak, Z.: Granularity, Multi-valued Logic, Bayes' Theorem and Rough Sets. In: Lin, T.Y., Yao, Y.Y., Zadeh, L.A. (eds.) *Data Mining, Rough Sets and Granular Computing*, pp. 487–498. Physica-Verlag, Heidelberg (2002)
54. Pedrycz, W. (ed.): *Granular Computing: An Emerging Paradigm*. Physica-Verlag, Heidelberg (2001)
55. Pinker, S.: *How the Mind Works* (1997)
56. Polkowski, L.: A Model of Granular Computing with Applications: Granules from Rough Inclusions in Information Systems. In: *Proceedings of 2006 IEEE International Conference on Granular Computing*, pp. 9–16 (2006)
57. Polkowski, L., Skowron, A.: Towards Adaptive Calculus of Granules. In: *Proceedings of 1998 IEEE International Conference on Fuzzy Systems*, pp. 111–116 (1998)
58. Qin, Y., Sohn, M.-H., Anderson, J.R., Stenger, V.A., Fissell, K., Goode, A., Carter, C.S.: Predicting the Practice Effects on the Blood Oxygenation Level-dependent (BOLD) Function of fMRI in a Symbolic Manipulation Task. *Proceedings of the National Academy of Sciences, USA* 100(8), 4951–4956 (2003)
59. Qin, Y., Carter, C.S., Silk, E., Stenger, V.A., Fissell, K., Goode, A., Anderson, J.R.: The Change of the Brain Activation Patterns as Children Learn Algebra Equation Solving. *Proceedings of the National Academy of Sciences, USA* 101(15), 5686–5691 (2004)
60. Rosen, B.R., Buckner, R.L., Dale, A.M.: 'Event-related functional MRI: Past, Present, and Future. *Proceedings of National Academy of Sciences, USA* 95(3), 773–780 (1998)
61. Shulman, R.G., Rothman, D.L.: Interpreting Functional Imaging Studies in Terms of Neurotransmitter Cycling. *Proceedings of National Academy of Sciences, USA* 95(20), 11993–11998 (1998)
62. Simon, H.A.: The Organization of Complex Systems. In: Pattee, H.H. (ed.) *Hierarchy Theory, The Challenge of Complex Systems*, pp. 1–27 George Braziller, New York, (1963)
63. Skowron, A., Stepaniuk, J.: Information Granules: Towards Foundations of Granular Computing. *International Journal of Intelligent Systems* 16, 57–85 (2001)
64. Skowron, A., Synak, P.: Hierarchical Information Maps. In: Ślęzak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) *RSFDGrC 2005. LNCS (LNAI)*, vol. 3641, pp. 622–631. Springer, Heidelberg (2005)
65. Sohn, M.-H., Douglass, S.A., Chen, M.-C., Anderson, J.R.: Characteristics of Fluent Skills in a Complex, Dynamic Problem-solving Task. *Human Factors* 47(4), 742–752 (2005)
66. Sommer, F.T., Wichert, A. (eds.): *Exploratory Analysis and Data Modeling in Functional Neuroimaging*. MIT Press, Cambridge (2003)
67. Sternberg, R.J., Lautrey, J., Lubart, T.I.: *Models of Intelligence*. American Psychological Association (2003)
68. Su, Y., Zheng, L., Zhong, N., Liu, C., Liu, J.: Distributed Reasoning Based on Problem Solver Markup Language (PSML): A Demonstration through Extended OWL. In: *Proc. 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE 2005)*, pp. 208–213. IEEE Press, Los Alamitos (2005)
69. Su, Y., Liu, J., Zhong, N., Zheng, L., Liu, C.: A Method of Distributed Problem Solving on the Web. In: *Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005)*, pp. 42–45. IEEE Press, Los Alamitos (2005)

70. Tsukimoto, H., Morita, C.: The Discovery of Rules from Brain Images. In: Arikawa, S., Motoda, H. (eds.) DS 1998. LNCS (LNAI), vol. 1532, pp. 198–209. Springer, Heidelberg (1998)
71. Turing, A.: Computing Machinery and Intelligence. *Mind* LIX (236), 433–460 (1950)
72. Varley, R., Siegal, M.: Evidence for Cognition without Grammar from Causal Reasoning and ‘Theory of Mind’ in an Agrammatic Aphasic Patient. *Current Biology* 10(12), 723–726 (2000)
73. Ward, L.M.: Synchronous Neural Oscillations and Cognitive Processes. *TRENDS in Cognitive Sciences* 7(12), 553–559 (2003)
74. Wu, J., Cai, C., Kochiyama, T., Osaka, K.: Function Segregation in the Left Inferior Frontal Gyrus: a Listening fMRI Study. *NeuroReport* 18(2), 127–131 (2007)
75. Yao, J.T.: Information Granulation and Granular Relationships. In: Proceedings of the IEEE Conference on Granular Computing, pp. 326–329 (2005)
76. Yao, Y.Y., Zhong, N., Liu, J., Ohsuga, S.: Web Intelligence (WI): Research Challenges and Trends in the New Information Age. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) WI 2001. LNCS (LNAI), vol. 2198, pp. 1–17. Springer, Heidelberg (2001)
77. Yao, Y.Y.: Information Granulation and Rough Set Approximation. *International Journal of Intelligent Systems* 16, 87–104 (2001)
78. Yao, Y.Y., Zhong, N.: Granular Computing Using Information Tables. In: Lin, T.Y., Yao, Y.Y., Zadeh, L.A. (eds.) Data Mining, Rough Sets and Granular Computing, pp. 102–124. Physica-Verlag (2002)
79. Yao, Y.Y.: A Partition Model of Granular Computing. In: Peters, J.F., Skowron, A., Grzymala-Busse, J.W., Kostek, B., Świniarski, R.W., Szczuka, M. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 232–253. Springer, Heidelberg (2004)
80. Yao, Y.Y.: Web Intelligence: New Frontiers of Exploration. In: Proc. 2005 International Conference on Active Media Technology (AMT 2005), pp. 1–6 (2005)
81. Yao, Y.Y.: Three Perspectives of Granular Computing. *Journal of Nanchang Institute of Technology* 25, 16–21 (2006)
82. Yao, Y.Y.: ‘The Art of Granular Computing. In: Kryszkiewicz, M., et al. (eds.) Rough Sets and Intelligent Systems Paradigms. LNCS (LNAI), vol. 4585, pp. 101–112. Springer, Heidelberg (2007)
83. Zadeh, L.A.: Towards a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic. *Fuzzy Sets and Systems* 19, 111–127 (1997)
84. Zadeh, L.A.: Some Reflections on Soft Computing, Granular Computing and Their Roles in the Conception, Design and Utilization of Information/Intelligent Systems. *Soft Computing* 2, 23–25 (1998)
85. Zadeh, L.A.: Precisiated Natural Language (PNL). *AI Magazine* 25(3), 74–91 (2004)
86. Zhang, B., Zhang, L.: Theory and Applications of Problem Solving. North-Holland, Amsterdam (1992)
87. Zhang, L., Zhang, B.: The Quotient Space Theory of Problem Solving. *Fundamenta Informaticae* 59, 287–298 (2004)
88. Zhong, N., Liu, J., Yao, Y.Y., Ohsuga, S.: Web Intelligence (WI). In: Proc. 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC 2000), pp. 469–470. IEEE Press, Los Alamitos (2000)

89. Zhong, N.: Multi-database Mining: a Granular Computing Approach. In: Proceedings of the Fifth Joint Conference on Information Sciences (JCIS-2000), pp. 198–201 (2000)
90. Zhong, N., Liu, C., Ohsuga, S.: Dynamically Organizing KDD Process. *International Journal of Pattern Recognition and Artificial Intelligence* 15(3), 451–473 (2001)
91. Zhong, N., Liu, J., Yao, Y.Y.: In Search of the Wisdom Web. *IEEE Computer* 35(11), 27–31 (2002)
92. Zhong, N.: Representation and Construction of Ontologies for Web Intelligence. *International Journal of Foundations of Computer Science* 13(4), 555–570 (2002)
93. Zhong, N., Liu, J., Yao, Y.Y. (eds.): *Web Intelligence*. Springer, Heidelberg (2003)
94. Zhong, N., Yao, Y.Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. *IEEE Transaction on Knowledge and Data Engineering* 15(4), 952–960 (2003)
95. Zhong, N.: Developing Intelligent Portals by Using WI Technologies. In: Li, J.P., et al. (eds.) *Wavelet Analysis and Its Applications, and Active Media Technology*, vol. 2, pp. 555–567. World Scientific, Singapore (2004)
96. Zhong, N., Wu, J.L., Nakamaru, A., Ohshima, M., Mizuhara, H.: Peculiarity Oriented fMRI Brain Data Analysis for Studying Human Multi-Perception Mechanism. *Cognitive Systems Research* 5(3), 241–256 (2004)
97. Zhong, N., Liu, J. (eds.): *Intelligent Technologies for Information Analysis*. Springer, Heidelberg (2004)
98. Zhong, N., Hu, J., Motomura, S., Wu, J.L., Liu, C.: Building a Data Mining Grid for Multiple Human Brain Data Analysis. *Computational Intelligence* 21(2), 177–196 (2005)
99. Zhong, N.: Impending Brain Informatics (BI) Research from Web Intelligence (WI) Perspective. *International Journal of Information Technology and Decision Making* 5(4), 713–727 (2006)
100. Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies (iIT) from the Stand-Point of Web Intelligence (WI). *Communications of the ACM* 50(3), 89–94 (2007)
101. Zhong, N.: Ways to Develop Human-Level Web Intelligence: A Brain Informatics Perspective. In: Franconi, E., Kifer, M., May, W. (eds.) *The Semantic Web: Research and Applications*. LNCS, vol. 4519, pp. 27–36. Springer, Heidelberg (2007)
102. The OGSA-DAI Project: <http://www.ogsadai.org.uk/>

Neuroscience: New Insights for AI?

Tomaso Poggio

McGovern Institute for Brain Research
Center for Biological and Computational Learning
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract. Understanding the processing of information in our cortex is a significant part of understanding how the brain works and of understanding intelligence itself, arguably one of the greatest problems in science today. In particular, our visual abilities are computationally amazing and we are still far from imitating them with computers. Thus, visual cortex may well be a good proxy for the rest of the cortex and indeed for intelligence itself. But despite enormous progress in the physiology and anatomy of the visual cortex, our understanding of the underlying computations remains fragmentary. This position paper is based on the very recent, surprising realization that we may be on the verge of developing an initial quantitative theory of visual cortex, faithful to known physiology and able to mimic human performance in difficult recognition tasks, outperforming current computer vision systems. The proof of principle was provided by a preliminary model that, spanning several levels from biophysics to circuitry to the highest system level, describes information processing in the feedforward pathway of the ventral stream of primate visual cortex. The thesis of this paper is that – finally – neurally plausible computational models are beginning to provide powerful new insights into the key problem of how the brain works, and how to implement learning and intelligence in machines.

I have always believed that theoretical results from information theory, theory of computation, and learning theory will play an important role in our understanding of how the brain processes information and how intelligent behavior arises from a large number of neurons. At the same time, I felt that the gap between computer science and neuroscience was still too large for establishing a direct connection. Until a few months ago, I always tried to keep separate the projects in my lab focusing on computer vision, i.e. developing engineered systems for image recognition, from the projects focused on the functions of visual cortex.

A few months ago, for the first time in my career, my perspective changed in a dramatic way. The turning point was a surprising discovery: a preliminary model implementing the theory of visual cortex on which we have been working for the last five years, in close cooperation with a number of anatomical and electrophysiological labs, turned out to perform as well or better than the best engineering systems and as well as humans on difficult recognition tasks involving

natural, complex images. In my mind this meant that we may be closer to a basic understanding of how visual cortex recognizes objects and scenes than I ever thought possible. It also means that the AI community should follow this kind of developments in neuroscience quite closely. Let me first describe the problem, its importance, and then the approach that I propose.

Specific problem: The human visual system rapidly and effortlessly recognizes a large number of diverse objects in cluttered, natural scenes. In particular, it can easily categorize images or parts of them, for instance faces, and identify a specific one. Despite the ease with which we see, visual recognition – one of the key issues addressed in computer vision – has remained quite difficult for computers and is indeed widely acknowledged to be a very difficult computational problem. Object recognition in primate cortex is thought to be mediated by the ventral visual pathway running from primary visual cortex, V1, over extrastriate visual areas V2 and V4 to inferotemporal cortex, IT. IT in turn is a major source of input to PFC involved in linking perception to memory and action. Over the last decade, several physiological studies in non-human primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. Given the wealth of physiological and behavioral data do we understand how visual recognition is done? Can we develop a theory leading to computer models capable of processing images as visual cortex does?

Why developing a theory is both difficult and important: After the breakthrough recordings in V1 by Hubel and Wiesel there has been a noticeable dearth of comprehensive theories attempting to explain the function and the architecture of visual cortex beyond V1. The reason of course is that a comprehensive theory is highly constrained by many different data from anatomy and physiology at different stages of the ventral stream and by the requirement of matching human performance in complex visual tasks such as object recognition. Thus, developing a consistent, quantitative theory is difficult. However, it would be extremely useful. Even a partial understanding of visual cortex is likely to provide powerful insights in how other parts of cortex work. Finally, theoretical foundations would be of key importance for the AI community because ultimately we want to understand the information processing involved in seeing and be able to replicate it in machines.

Preliminary results: One of the first models of visual object recognition, Fukushima's Neocognitron (Fukushima, 1980), followed the basic Hubel and Wiesel hierarchy (Hubel and Wiesel, 1968) in a computer vision system. Building upon several conceptual proposals (Perrett and Oram, 1993; Wallis and Rolls, 1997; Mel, 1997), we developed (Riesenhuber and Poggio, 1999; Serre et al., 2002; Giese and Poggio, 2003) a similar computational model. The present theory (Serre et al., 2005) has evolved over the last 6 years from that initial model. The theory is the outcome of computer simulations, trying to quantitatively account for a host of recent anatomical and physiological data. It is mainly the result of collaborations and interactions with several neuroscience experimental

labs (N. Logothetis in the early years and now D. Ferster, E. Miller, J. DiCarlo, C. Koch, I. Lampl, W. Freiwald, M. Livingstone, E. Connor). The architecture of the model resulting from the theory is shown in Fig. 1. It is qualitatively and quantitatively consistent with (and in some cases actually predicts) several properties of cells in V1 (Lampl et al., 2004), V2, V4 (Gawne and Martin, 2002) and IT (Logothetis et al, 1995; Hung et al., 2005) as well as fMRI and psychophysical data (Riesenhuber et al., 2004). The present theory bridges several levels of understanding, from computation and psychophysics to system physiology and anatomy, to the level of specific microcircuits and biophysical properties. The key extension with respect to the original model by Riesenhuber and Poggio is an unsupervised learning of the tuning of each unit at the S2, S2b and S3 levels (possibly corresponding to V4 and PIT, see Fig. 1) on a set of natural images unrelated to the task. In the present model, units (of the simple type) become tuned to the neural activity induced by natural images within their receptive field.

The most interesting, new result is that the model (see Fig. 1) outperforms the best computer vision systems on several different recognition tasks on real-world natural images. In fact, this is perhaps the first time that a model of cortex does as well as humans on a natural image recognition task. Even more surprisingly, the model mimics human performance when tested for rapid categorization without eye movements. The full theory and the results above are still unpublished (apart from a technical report titled, “A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex” by Serre, Kouh, Cadieu, Knoblich, Kreiman and Poggio, 2005).

I should emphasize that this is still far from solving the problem of vision. First, vision is more than object recognition and the visual cortex is more than the ventral stream. Second, the model in its present form cannot account for normal, everyday vision which involves eye movements and complex attentional top-down effects which must be mediated by higher brain centers and the extensive anatomical backprojections found throughout visual cortex. However, this theory may account for the immediate recognition of single pictures – a task humans can perform very well.

The open question – beyond establishing the basic aspect of the feedforward model – is whether it can be extended in the next few years to become a full theory of normal vision. The obvious approach involves physiology, psychophysics, fMRI and AI. In this project graduate students from computer science would work along with neuroscience students trying to solve simultaneously the problem of how visual cortex works and how to build a machine that sees.

In summary, the time may have come for AI to learn from serious neuroscience.

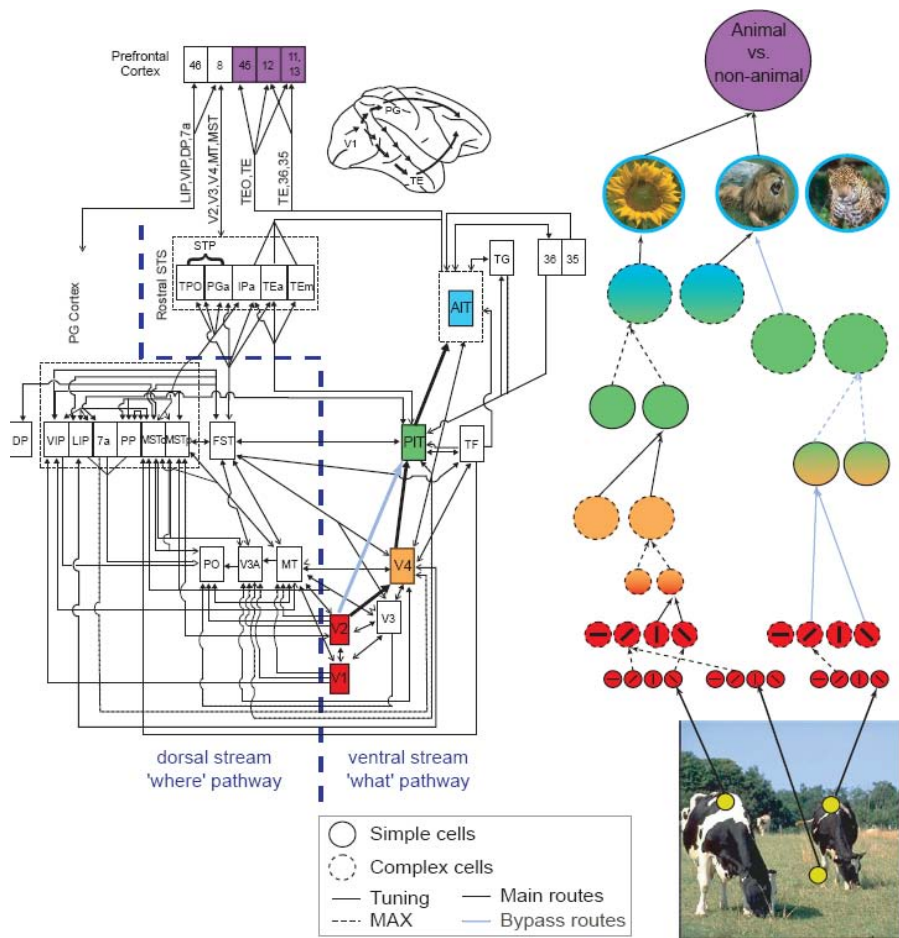


Fig. 1. The basic architecture of the model of the ventral stream (right). The figure provides a tentative mapping to the structural primitives of the ventral stream in the primate visual system (left). The theory assumes that one of the main functions of the ventral stream is to achieve a trade-off between selectivity and invariance. There are two basic operations iterated throughout the hierarchy. Stages of “simple” (S) units with Gaussian-like tuning (plain circles and arrows), are interleaved with layers of “complex” (C) units (dotted circles and arrows), which perform a max operation on their inputs and provide invariance to position and scale. Developmental-like unsupervised learning, on a set of natural images, determines the tuning of the simple units in the S2 and S3 layers (corresponding to V4 and PIT, respectively). Learning of the synaptic weights from S4 to the top classification units is the only task-dependent, supervised learning stage in this architecture. The total number of units in the model is in the order of 2^7 . Colors indicate the correspondence between model layers and cortical areas. The table on the right provides a summary of the main properties of the units at the different levels of the model. The diagram on the left is modified from Van Essen and Ungerleider (with permission).

Network Thinking and Network Intelligence

Deyi Li¹, Liping Xiao², Yanni Han³, Guisheng Chen¹, and Kun Liu²

¹ Institute of Electronic System Engineering, Beijing 100039, China
ziquinli@public2.bta.net.cn

² Institute of Command Automation
PLA University of Science and Technology, Nanjing 210007, China
xlp-ping@126.com

³ State Key Lab of Software Development Environment
Beijing University of Aeronautics and Astronautics, Beijing 100083, China
hyn@nlsde.buaa.edu.cn

Abstract. Networks interact with one another and are recursive. Network intelligence and networked intelligence, as the way of knowledge representation, have become very active recently. What topological measures can be used to characterize properties of networks? What properties do different structures of real-world networks share, and why? How did these properties come about? How do these properties affect the dynamics of such networks? How to use network topology to extend other dimensions? Given a real-world network with certain properties, what are the best ways to search for particular nodes? Furthermore, some specific implementations and examples of network intelligence will be given in this paper. Such as mining typical topologies, discovering sensitive links and important communities from real complex networks, networked control, and making a virtual reality of emergence phenomenon in complex systems.

1 Introduction

With the development of science and technology at the information age, our society depends more on a diversity of networks rather than those isolated systems. Network becomes the engine of the 21st century's scientific research. Network has been researched in two main aspects: nodes and interaction between the nodes. Especially in recent years, many people take great interests in the network with complex regularities and characters [1][2]. The reason appears obviously. For one thing, networks are everywhere and have different shapes, including Internet and its close cousin the World Wide Web, network of air lines, large-scale electric power grid, the structure of a piece of Very Large-Scale Integration (VLSI), human social relationships and even artificial works, among others. For another, what is more important is that people seek to explore the common law in network. Most complex problems in real world can be represented in the form of network topology, which is composed of nodes and links. By this way, we can map the problems in real world to those in the topology representation of complex network. These are just the hot topics in current research, network thinking and network intelligence.

Network thinking is distinct from symbol thinking, which is a special form of imagery thinking. It implies a novel thinking methodology, which can be used to describe the research object of complexity and the inherent laws of dynamics by means of the foundational theory of complex network and the novel knowledge representation with network topology. Thus the basic viewpoints of network thinking may be described as follows:

1) In complex network, more attentions will be paid to the interaction and inter-affection between nodes rather than those isolated nodes or links.

2) Each sub-system, or reduced view of the topology at different scale, has the characteristics of self-similarity; self-organizing and the network structure may evolve towards a certain object.

3) Through the gradual interaction and inter-affect between nodes of networks, the whole complex networks will behavior the phenomenon of emergence, such as synchronization among nodes.

We reconsider the science and technology of AI based on the new method of network thinking. Generally speaking, it doesn't mean to apply the artificial intelligence technology to the context of network and distributed system directly, but to emphasize the idea of cognition such as "Global First" and "Topology First". The essence of network intelligence lies in exploring the general laws among the uncertainty in the network and the self-organizing, self-similarity and the mechanism of emergence during the evolvement process of network. By this way, the intelligent behavior from the collective features of network can be discovered and the knowledge can also be found out.

It turns out that many interesting and worthwhile work have been done about the complex network in recently years. Especially a series of important characteristics of small world network or scale-free network imply that the scale of network may be very large, whereas the fundamental information to describe the topology of network may be little. In the next section some remarkable characters among typical topologies in complex network will be introduced [3]. Then four aspects of network intelligence including networked data mining, networked control system, networked information processing and storage, and emergence computation will be presented.

2 Network Topology Modeling and Features Analysis

There has been a strong upsurge in the study of complex network in many fields, ranging from nature science and sociology to computer science and communications, etc. Stephen Haking, one of the most influential ideologists, has pointed out that "I think the next century will be the century of complexity". Evidences widely existing in nature indicate that we can't deal with the complex system only by reductionism, and we also can't predict the abundant behaviors of the whole complex system just from the sum of many isolated objects, which means the whole is larger than the sum of all parts. In a word, the research of complex network is a new territory of contemporary scientific research.

2.1 General Concepts of Network Modeling

The real world is a complex system with the properties of multi-parameter, non-linear, time-varying and unstable. From the viewpoint of network thinking, the topology of complex network can abstractly describe many popular systems in nature and society. The concepts, entities and phenomena involving in complex network system are filled with uncertain factors. There are several resulting statistics parameters such as degree distribution, average path length between pairs of nodes, clustering coefficient, which play important parts in the process of analyzing the complex network [4]. The degree of a node is topologically defined as the total number of its connections. Therefore, the larger the degree, the more important role the node plays. The average of every node's degree over the network is called the average degree of the network. And the statistical distribution of all nodes degree is characterized by a probability distribution function. In fact, many empirical results demonstrate that for most large-scale real networks, such as the World Wide Web, the degree distribution deviates significantly from the Poisson distribution and follows the power-law distribution. Furthermore, the distance between two nodes is defined as the number of edges along the shortest path connecting them. While the average path length L is the mean distance over all pairs of nodes. Here L determines the effective size of a network and represents the most typical separation of one pair of nodes. The inherent tendency of clustering is quantified by the clustering coefficient. For a given node i , the nodes directly connected with i are the neighbors of i . If the node i has k neighbors, these neighbors can have C_k^2 edges theoretically. Given the number of edges within these neighbors is m , the clustering coefficient C_i of the node i is defined as m_i/C_k^2 . The average of all nodes' clustering coefficients is called the clustering coefficient of the network. Many evidences show that most real networks have the feature of clustering. Presenting the real world by these parameters, we can abstract the details of network topology. Then we can study the characteristics of relevant model theoretically. Thus, in the next section, we give an outline of the collective features and evolution laws of some typical topology networks.

2.2 Theoretical Analysis of Typical Topology Models

The results of numerous studies make scientists realize what's more important are to study the network topology rather than isolated nodes. Although networks in the real world have different shapes and seem independent, the structure and general laws may be similar in nature [5]. Different topology structures, as clearly seen, represent different models in reality and imply different knowledge patterns and locomotion rules. There have been several typical topologies presented in the earlier years, such as pure random graph, small world network, scale-free network, hub network and star network, and so on. The following sections elaborate on three of them: random network, small world network and scale-free network.

Random network

The research pioneers of random graphs are two Hungarian mathematicians Paul Erdős and Alfred Rényi [6]. In some sense, the construction of this graph can be abstractly regarded as the evolvement of a random network. A random network can be generated by the following steps: each node can get connected to any other arbitrary nodes with some probability p . The topology of real large networks (i.e. Internet, WWW, telephone networks, ecological networks) substantially differs from the topology of random graphs produced by the simple model [7]. On the one hand, a random network doesn't exhibit the strong clustering of nodes seen in the real world. Empirical studies show that clustering coefficients in real world are often orders of magnitude higher than the value in a random network with the same number of nodes and links. On the other hand, the degree distribution of a random network is approximately Gaussian which has the characteristic of short tails and clear cutoffs. In contrast, the power-law degree distribution is often observed in real world, which has a long tail without cut-off points. Consequently, a large scale of random network often exhibits the following characteristics: low clustering coefficients and low average path length.

Because of the difference between random network and real world, it turns out that new methods, tools and models needed to be developed. In the past few years we have witnessed dramatic advances in these aspects [8].

Small world network

Despite the large size of network, it commonly happens that there is a comparatively short distance among some pairs of nodes. Thus in 1998, Watts and Strogatz introduced a new network model named small world network, which meant the shortest path length (number of link hops) between two nodes scales logarithmically with the size of network for fixed average degree [9]. To build such a network, each node should be linked to k of its nearest neighbors in the context of network consisting of N nodes arranged in a ring. For each link, we rewire one end of the link to a randomly chosen node in the network with small probability p . The small world network can be highly clustered as regular lattices and has small characteristic path lengths like random graphs. An often-cited example is known as "six degrees of separation" [10]. The properties of small world have been suggested to have adaptive meanings, which can be of great importance in recognizing evolutionary paths, the origins of homeostatic stability and the sensibility to perturbations in biological networks [11]. By employing the idea of network thinking, we can solve many problems in real world. Small world network with low but non-zero p also exhibits a high clustering coefficients and low average path length.

A common feature of the earlier network models (such as ER or WS model) is that they predict the probability distribution of nodes connectivity $P(k)$ has an exponential cutoff and a characteristic size k . Here we demonstrate a new model that is free of scale and follows a power-law distribution over many orders of magnitude.

Scale-free network

Barabási and Albert developed an alternative network model, of which the distribution has no cutoff and the instances of scales are all presented, so designated the term “scale-free”. Scale-free network can be easily obtained provided that the links to a new node are made preferentially from nodes that already have many links. That means the more connections of the node; the earlier it has been incorporated. So there are many variations in the number of links per node. In the scale-free network, there are a small number of nodes with many links coexisting with a much larger number of nodes with the smaller number of links. Meanwhile, it has two distinct characteristics. First, the mechanism of growth: the network starts out with a small number of initial nodes. At each time, a new node is added and is connected to the existing nodes. The probability of making one of those connections to such a node is proportional to the degree of this node. Second, the mechanism of preferential attachment: network grows in such a way that nodes with higher degree receive more links than nodes with lower degree. The above mechanisms are universal which can be applied to a lot of complex systems, including business networks, social networks and transportation networks.

More generally, “small world” and “power-law distributions” are generic properties of network topology. It is a great advancement to represent the network topology in the form of knowledge representation and simulate it by means of computer than representing it by logic symbols before [12]. The models of EA, WS and BA accordingly present the mathematical ways to generate the random network, small world network and scale-free network. On the contrary, there exist many uncertainty factors in the evolutionary process of complex network in the real world. Meanwhile, there is no remarkable boundary between different network topology patterns. So it is hard for us to create an ideal and independent model in the language of mathematics. There are two fundamental problems in the process of using knowledge representation to describe the network topology: How to abstract the typical network topology in accordance with the real word? How to simulate it with computer as much as possible? From the viewpoint of network thinking and network intelligence, we propose a new approach to combining several typical networks of complex system and reveal phenomena in nature vividly.

The imagery of complex network systems pervades modern culture. In fact, the system in real world is not fixed at the original scale. It grows up dynamically and evolves into larger complex network according to some rules. The interrelation and interaction between nodes in the system can be enhanced. Different sub systems are self-similarity with the collective system. The dynamics behavior of local nodes can lead to the synchronization of the whole system after propagation in the network. Meanwhile the networks exhibit resilience to the loss or failure of nodes or links and to determine their vulnerabilities. By building the topology model and analyze the dynamics, we can explore the inherent essence and laws, which shed new light on the solution of AI problems. We present four interesting directions of network intelligence for future research and the related work in details below.

3 Networked Data Mining

Data mining, as commonly understood, is the process to extract hidden information and knowledge, which is previously unknown and potentially useful, out of a large quantity of data, which is incomplete, fuzzy and noisy [13]. Through ten-year research, although a lot of effective methodologies and tools in data mining have been developed, for instance recursive analysis, discriminate analysis, learning from demonstration, genetic algorithm, neural networks, multi-dimensional data analysis, and attribute-oriented inductive method etc., there is still a large gap between the performance of the existing methodologies and the expectation by the public. Most of those methodologies are based on the hypothesis of independent same-distribution data, and can't open out anfractuous relationship between actual data, just like super linking relationship of webpage, social relationship, and interaction between proteid and so on. Thus it has become a top problem how to construct relationship between data samples in large scope and mine inherent knowledge and patterns inside different objects or network [14].

3.1 From Relational Data to Networked Data

Networked data mining is a new direction that mines useful data in web environment. Data in the process of networked data mining is not traditionally related data any longer but is networked data, which is represented by network topology. So it is the precondition and basis of networked data mining to represent data by network topology [15]. Generally speaking, when problems in a real world are extracted by complex network methods, the research objects are represented as nodes and relationship between them which are expressed as links, without thinking about the detail features of nodes and material shapes of links. In order to improve the representation ability of network topology, it is very necessary to deeply research how to discover features of nodes and links and realize associated networked data mining.

Figure 1 shows the process of networked data mining. First, related data in huge raw data of mining objects is extracted and organized by traditional relationship model. Second, those data is represented by the pattern of network topology, such as degree, betweenness and community structure etc., via network analysis and representation. When the topology structure is got, patterns hiding in network can be found by community detecting and backbone retrieving. At last, valuable knowledge is obtained by interpreting those topology patterns automatically [16].

Compared with traditional data mining, networked data mining maybe contains following main characteristics:

- 1) Raw data to be processed is represented and treated from related data model to networked data model, and knowledge being mined may be interpreted topologically, which reflects inherent relationship between those data objects. It means that as long as valuable topology knowledge is discovered, knowledge and information reflecting inherent relationship of things can be understood and explained well at last.

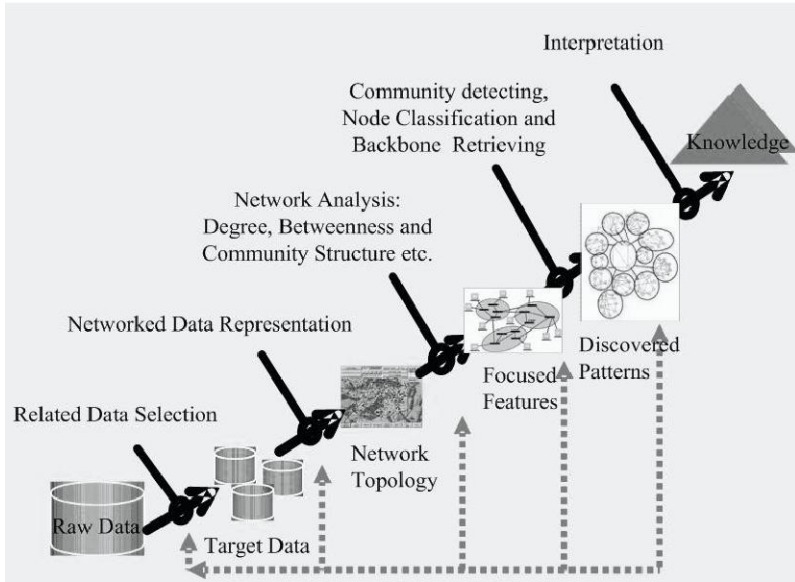


Fig. 1. The process and key issues of networked data mining. Firstly, topological analysis permeates along each steps; secondly, while the raw data is represented with topology, the collective characteristics are implied in the figure at different scales; finally, the consequent result are conveniently interpreted, understood and reflected back to the original data space.

2) The mining procession of networked data mining is based on the theory of complex network that through analyzing characters of network topology, mining the typical topology structure, and reducing layer network iteratively, then important topology models are obtained.

3) Furthermore, the networked representing data and knowledge can be visualized easily, and networked data mining can realize alternating visualized data mining, which also integrates human ability of cognition. Therefore the efficiency may be largely improved by the alternating supervision and experts' inductions in mining process.

3.2 Mining Sensitive Links and Important Communities

Many networks are inhomogeneous, consisting of lots of undifferentiated nodes, and these nodes may be combined to a number of distinct groups. So, Classifying backbone nodes and sensitive links, discovering important communities from a real world network are two key issues in networked data mining.

Discovering the backbone nodes and sensitive links means to discover the nodes and links which are so most important that the whole network would fall into pieces if they are attacked or destroyed. By computing the betweenness of links, a complicated topology structures may be reduced to concise backbone

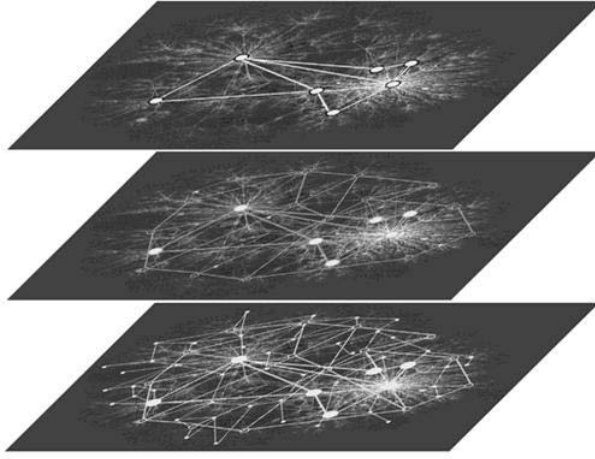


Fig. 2. Discover knowledge through mining sensitive links and important communities in a network topology gradually from one common scale to another reduced backbone scales. It illustrates that different backbone topology may be mined in different scales; this works just like that human intelligence discovers knowledge from varying views, which probably shows network intelligence in data mining.

networks at multi-scales on demand [17]. Links with low betweenness and nodes with small degrees are deleted and neglected, then the remainders are sensitive ones which are very meaningful to research the complex networks (see Fig. 2).

To discover communities in complex networks is the other key issue in networked data mining. Many networks in the real world turn out to be highly clustered. Finding the most significant communities make great sense to uncover the inherent features of a complex network.

Community model can capture the hierarchical features of the Network (see Fig. 3). There are lots of communities in a network, and inside one community the connections of nodes are abundant and close but between communities the connections are less and relaxant. But such links between different communities are usually more important than those inside the same communities. They connect different communities and are bridges between them, and without those links the whole network maybe broken into a set of multiple communities not an organic one any more.

One method of mining community based on link betweenness is the links removal method. With this method, links with the highest betweenness are cut off, so the most important links of different communities are removed firstly. And then the links with less betweenness are removed in turn until all links connecting different communities are removed.

The purpose of networked data mining is to find the communities and kernels with high potential in a complex networks. Finding those is a way to uncover the some mystery of the complex network and is much helpful to research, use and

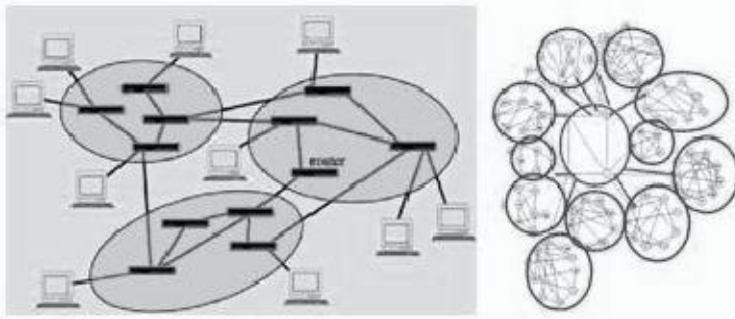


Fig. 3. Community detection and classification. The figure shows that closer ties exist inside the community but mere links occur between communities. The extent and intensity of communication via links may be measured by betweenness of links, thus betweenness of links can be taken as one criterion of community detection. The left figure shows the case of a typical computer network, and the right one shows the case of friendship between native minority people, both of them have cluster features.

design complex networks. Several experiments based on real networked data sets have been shown the encouraging results on discoverer important communities [18].

Considering the complex networks in the real world, study on only above factors, nodes, links and communities, can not meet various demands. For instance, for Web mining the bandwidth of a link and throughput of a node, for a traffic network the scale of population in a city and the geographical distance between two cities, for a virus propagation network the epidemic spreading rate of a virus, etc. usually can not be neglected. Therefore, we need to extend some more properties of networks including the quality of a node, physical distance between two nodes, the age of a node, betweenness of a link and betweenness of a node, in order to discover different kinds of knowledge.

3.3 Mining Typical Topology from a Real Work

Mathematically speaking, any typical topology is only an approximate mapping from the reality [19]. There are five typical topologies: pure random graph, small world network, scale free network, hub network and star network. We may order the different typical topologies in the degree of randomness [20]. Figure 4 describes, from left-bottom to right-up, the pure random with Poisson distribution, semi random with power-law distribution and none random with delta distribution. At the left-bottom area, the corresponding network is more uniform, robust and chaos; at the right-up area, the corresponding network is more uneven, fragile and controllable.

As a matter of fact, a real complex network may have hybrid nature of all these topologies. However, with different information granularity or at different views, an evolutionary and growth network may be by and large characterized by an ideal typical model.

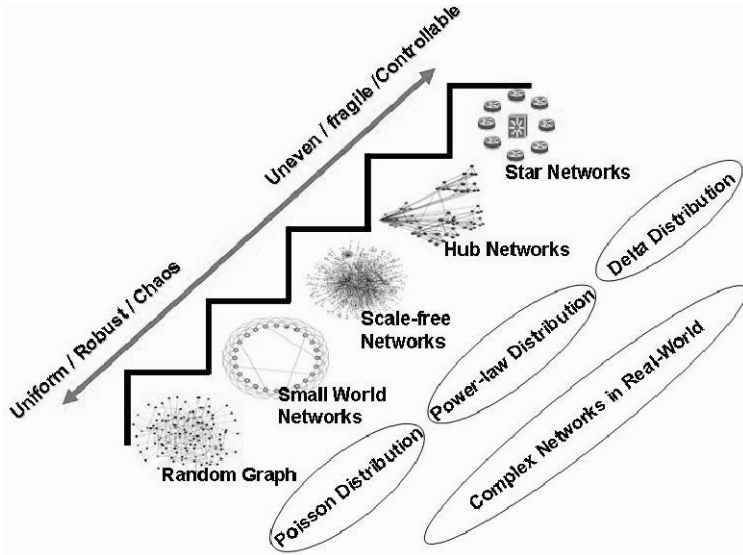


Fig. 4. The topological features and their evolutionary tendency of five typical network models. From left-bottom to right-up, the pure random with Poisson distribution, semi random with power-law distribution and none random with delta distribution. At the left-bottom area, the corresponding network is more uniform, robust and chaos; at the right-up area, the corresponding network is more uneven, fragile and controllable.

4 Networked Control Systems

4.1 From Classical Cybernetics Engineering to Networked Cybernetics Engineering

With the growing development of classical cybernetics engineering, human being can well control many stable or unstable objects by simply using their approximate, qualitative knowledge and experiences, such as riding a bicycle, making a long flexible pole upright on his forehead. An acrobat even can balance very complex objects on his own strange way. As shown in Figure 5, to stabilize an inverted pendulum system has become a typical and hot subject in imitating such a human-intelligence and automation for about thirty years. The kernel of inverted pendulum control systems is very much the same as robot control, space aircraft control, and many other unstable systems. The typical features of this control system include localized control structure, independent control object and local close-loop. It is completed maybe all by ONE industrial PC [21].

Nowadays, Network thinking and network intelligence is also considered in the field of real physical control system. With the rapid increasing developments ranging from control systems to computer technologies, and from network

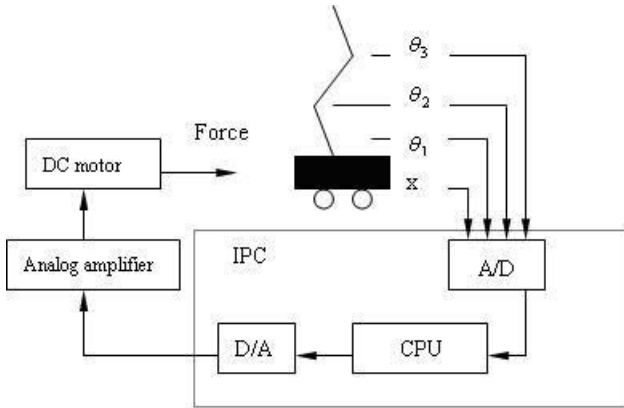


Fig. 5. Classical cybernetics engineering. Here shows the experimental block diagram of triple Inverted pendulum control system. A pole is hinged to a motor-driven cart that moves on rail tracks to its right and left. The pole has only one degree freedom (rotation about the hinge point). The primary control tasks are to keep the multi-link inverted pendulum vertically balanced and keep the cart within the rail track boundaries. The dynamics of the inverted pendulum system are characterized by eight state variables: x : position of the cart on the rail; $\dot{x}$: velocity of the cart; and upper pole to the vertical axis respectively; angle velocities with the lower, middle and upper pole to the vertical axis respectively.

communication technologies to large scale integrated circuits, many smarter sensors and actuators have been integrated into a networked control systems (NCS) via communication interfaces. In the NCS, not only a sensor or actuator itself is designed as a node, but also a subsystem composed of one more closed-loops is treated as ONE node of a network at different scales. The process from input variants to output gain, for example, is no longer locally realized, by contrary, remote and distributed sensing, long-distance measurement, remote control and heterogeneous data fusion become to be key issues of NCS.

The outstanding characteristic of a networked control system is having one or more close-loops. Each of the close-loops may be treated as ONE node in complex network, and more attention is now paid to the interaction and inter-affection between nodes via their links rather than to the node itself (i.e. one close-loop). Typically when the words networking and control are used together, the focus is commonly on the control of networks, but from the viewpoint of network thinking, it doesn't means control of networks but networked control, or control through networks. Therefore, although the node itself plays a subordinate but important role in an NCS, the feedback loops should be designed to links in the network since the network will have many other possibly essential characteristic, and NCS design objectives revolve around the performance and stability of a target physical device, which is linked to network, rather than of the computer network. For example, the problem of stabilizing one node within a close-loop

4.2 A Challenge to Make More Reliable Control Based on Networks with Uncertainty

With control systems being networked, the traditional close-loop feedback within a node is extended to two-dimension network topology. Control systems get larger and more open and complex, relation and interaction between different control systems need more and more attention to be paid. It is a great challenge to make more reliable control if it is based on networks with uncertainty. When two stable control systems interact with each other, as shown in Figure 7, it is valuable to study how to make sure the whole system is still stable.

To make more reliable control based on networks with uncertainty, the following key issues have to be well treated and it has been a great challenge for researchers in recent decades.

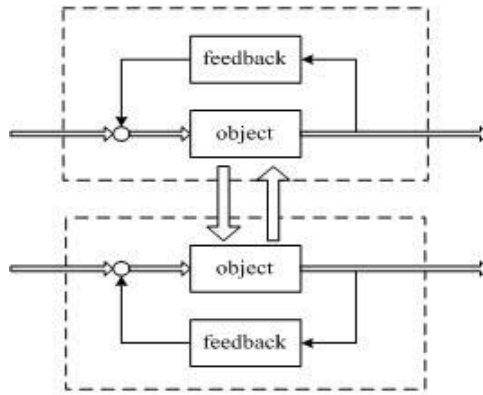


Fig. 7. Two stable control systems interact with each other. With control systems being networked, the traditional close-loop feedback is extended to two-dimension network topology. Control systems get larger and more complex.

1) The time-varying transporting period and sampling period through a network with uncertainty may result in conflict, unconventionality and abnormality of the dynamics of nodes.

2) The scheduling of shared resources in complex networks may interfere the control algorithm.

3) The time-delay in closed-loops induced by networks may reduce the performance of NCS, and finally leads to instability of system.

Generally, the field of networked control systems is technologically driven and motivated by cost, maintenance, and reliability. The use of networks will be common in practice, for researchers, the field of networked control systems has much to offer. Developing techniques to dynamically analyze the characteristic of nodes and links in NCS remains a challenge. A detailed analysis would be more useful to both design engineers and theory researchers. Ultimately, research in this area promises the interesting combination of Artificial Intelligence both in control theory and complex networks.

5 Networked Information Processing and Storage

5.1 Networked Information Storage

In complex networks, the methods of information storage are much different because there are various topologies in them. In general nodes in a network can be divided into four types: hubs of modules (nodes with many connections within their community or “module”), ultra-peripheral nodes (all links are within the node’s module), peripheral nodes (most links are within the node’s module), non-hub connectors and non-hub kinless nodes (links homogeneously distributed among all modules). Usually ultra-peripheral nodes are nodes within module and play most non-structural roles, and when the system changes or updates, they are easily abandoned. In contrary, the others nodes which play more structural roles in different degrees will be conserved with much more possibility [23][24][25].

The networks are constructed by the mechanism of “growth by preferential attachment”, and the networks grow in such a way that the nodes with high degree receive more new links than others. Stations of information in complex network are different because it is stored on different nodes with unlikeness functions. Information which is stored in structural nodes tends to be referred and has widely affection and plays more important roles, even definitive roles, in the whole complex networks. But information which is stored in non-structural nodes like ultra-peripheral nodes maybe neglected or abandoned when the network changes or some accidents happen to the network. The same information in different nodes will have much different force and affection in networks. That is the reason that vaccinating a hub is more likely to slow the spread of a disease than vaccinating a randomly chosen person. What is important is not the hub person himself but is his station in the network [26].

So based on above analysis, we can bring forward some improvement to configure information storage in network:

- 1) Optimize information storage in network. Because node degree distribution is non-uniform, we can optimize information storage in network, and try our best to dispose important information in backbone nodes. Usually important information in network should be transmitted and referred in large scope. Therefore it is helpful to accelerate important information propagating by disposing them in backbone nodes, let them scatter to all network in a short time, and improve their affection to the collective network.

- 2) Pay more attentions to backbone nodes. The non-uniform network is fragile because there are backbone nodes. If these backbone nodes are attracted through malice, the whole network will be affected. So the backbone node should be paid more attentions than others. At any time we should make sure the backbone nodes are in good conditions so that the whole network can work normally. If disaster or some accidents happens, the backbone nodes and information in them both should be protected well. Disaster Resume Program (DRP) is one system that records significant information and help companies resuming to normal condition from the devastation of disaster rapidly.

5.2 Networked Information Processing

As information stored in different nodes has large different affection to the whole network, these characteristics lead to the information propagating non-uniformly, too. Non-uniform distribution network is fragile when it faces diseases or infection. While in uniform distribution network there is a critical point, just when the propagation speed is bigger than the critical point, diseases will spread in the network. In non-uniform distribution network there isn't this kind of critical point. The computer diseases will spread and remain in a balance at last, as long as its propagating speed is bigger than zero. This phenomenon is resulted in by the characteristics of the network, which means that in non-uniform distribution network, like Internet, few computers infected by diseases would influence the whole network. And due to small world features of network, such disease will spread very quickly, and live in a long time.

The WWW is a typical complex network which is familiar to us. Despite its increasing role in modern society, the WWW is still uncontrollable: any individual or institution can create a website with any number of documents and links. Much work has been done to research various properties of the web, including its diameter, degree distribution, connected components and macroscopic structure. Because of the method of its growth, the Web can be studied as a directed graph or an undirected graph, whose nodes correspond to static pages on the web and whose links correspond to hyperlinks between these pages. Although it is a large and complex, the average of the shortest path between two nodes, d , that is the number of links which must be followed to navigate from one documents to the other, has direct proportion with logarithm of the total number of pages, N . It indicates the Web forms a small-world network. With normal N , the diameter d is only about 19! But the existence of the shortest doesn't mean that it is easy to be found. People with only local knowledge of the network have much difficulty in searching for short path A search engine also needs to search near 10% of the whole network to find a document at distance d .

It is necessary to understand the dynamics of propagation of information for Websites to be visited. The web turns out to be a scale-free network with power-law distribution, where 80% Web information is only located on 20% Websites. There are lots of factors which affect information processing in the web, such as: the Web pages are added and updated all the time, the connection in Web pages located in a Website show some communities, and authority and credit standing of the Website owner. Beside of the Web structure and topology, all of those can affect the dynamic of information processing in network and they make it uncertain on a certain degree.

So we should research the dynamics of process on network, mine and discover backbone nodes, deploy important information on them, and offer more search strategy to Web search engine and speed up the information propagation Consensus war in military affairs, drumbeating in amusement calling, all of that make use of the dynamics of the propagation of information in networks, let purposeful information propagate in the whole network rapidly, and attain their goals in will.

5.3 Web Intelligence

With Internet and the World Wide Web (WWW) developing, we have entered into an information age. The Web provides a total new communication media, like audio media, video media, and figures etc. The Web impacts all aspects in our lives, from academic research to ordinary daily life. It revolutionizes the way in which information is gathered, stored, processed, presented, shared, and used. The Web offers new opportunities and challenges for many areas [27][28][29].

Web Intelligence is firstly introduced and formally announced in the new Web Intelligence conference in a position paper at a Panel on Data Mining and Web Information Systems at the 24th Annual International Computer Software and Applications Conference (IEEE COMPSAC) in 2000 by Zhong et al. WI encompasses Web information systems environments and foundations, ontological engineering, human-media interaction, Web information management, Web information retrieval, Web agents, Web mining and farming, and emerging Web-based applications. It also aims at deepening the understanding of computational, logical, cognitive, physical, and social foundations as well as the enabling technologies for developing and applying Web-based intelligence and autonomous agents systems (Liu et al., 2003). As a new research field, WI joins AI and IT, and makes use of them on the new platform of Web and Web-related products, systems, services and activities. On the one hand, it may be viewed as applying results from these existing disciplines to a totally new domain. On the other hand, WI may also introduce new problems and challenges to the established disciplines [30][31][32].

Web Intelligence presents excellent opportunities and challenges for the research and development of new generation Web-based information processing technology. Web mining is the process to discover and extract potentially useful information and knowledge from Web data. It encompasses tasks such as mining Web search-engine data, analyzing the Web's link structures, classifying Web documents automatically, mining Web page semantic structures and page contents, etc. [33].

Now current search engine, like Google, Yahoo, are index-based Web search engine which crawls on the Web, indexes Web pages, and builds and stores huge keyword-based indices that help locate sets of Web pages that contain specific keywords. However they suffer from several deficiencies. First, a topic of any breadth user input can easily contain hundreds of thousands of documents, which can lead to a search engine returning a huge number of document entries, many of which are only marginally relevant to the topic or contain only poor-quality materials. Second, many highly relevant documents might don't contain keywords that explicitly define the topic, a phenomenon known as the *polysemy problem* [34].

So we hope to build a new data mining system which can enhance the quality of Web searches. To do so, one of methods is to construct meta search-engine. For example, a search for the keyword and its synonyms can be performed by normal index-based Web search-engines and they return the result respectively. Then, we can cluster or classify this result set of documents via some intelligent

text cluster arithmetic, which can obtain more relative Web documents [35]. This method has been used in some degree, but because there is little excellent text cluster and classify arithmetic, development of meta search-engine are still limited in a certain degree.

6 Emergence Computation

Emergence is a common feature of complex systems. The mechanism of emergence is still a problem which is difficult to understand. Emergence computation is one of the methods to study emergence. By computation it exploits the self-organizing, collective, competitive, and cooperative phenomena arising by the interaction of components following certain dynamical laws [36]. Macroscopic features of complex system can be gained by setting up and running emergence computation system. Emergence computation is a type of computation which is bottom-up, not programmed globally or totally. In emergence computation system, local information or limited information is taken as a unit of computation, and certain collective behavior may emerge in an unexpected way. This process is similar with self-organization in chemical processes, such as Belusov-Zabotinsky reaction or physical processes, laser oscillation. In brief, emergence computation is to study how high-level representation and information processing emerge via interactions of vast simple individuals.

6.1 From Pattern Recognition to Pattern Emergence

As a new direction of pattern recognition [37], pattern emergence attracted many researchers recently. Pattern is representation of emergence in complex system [38]. A system, consisting of many basic components, can emerge collective synchronization via interaction among them. Usually synchronization represents as a kind of pattern, where emergence is one of the reasons that such pattern forms.

Emergence is a phenomenon that arises by interactions among components in complex systems and can be observed in a system view [39]. Emergence is a sudden shift from one state to another, also called phase transition in physics. Emergence focuses on the mechanism that the whole is more than the sum of the parts [40]. Emergence phenomenon exists extensively in each scale and various domains, from nucleon to the universe, from living organizations to human society, from chemistry to physics. For example, fireflies synchronize their flashes; adaptive path minimization by ants; fish schooling and bird flocking; wasp and termite nest building; pattern formation in animal coats; women living together synchronize their menstrual cycles; the audience in assembly halls spontaneously clap in synchrony without any active control; and synchronization of the Huygens' pendulum clocks, etc. [41].

Emergence is a type of interaction with coupling and front-back association above all. This kind of interaction mainly is self-adaptability and condition adaptability of individuals, with no central control. Components obey respectively regulation, interact with each other, improve self-action, and consequently

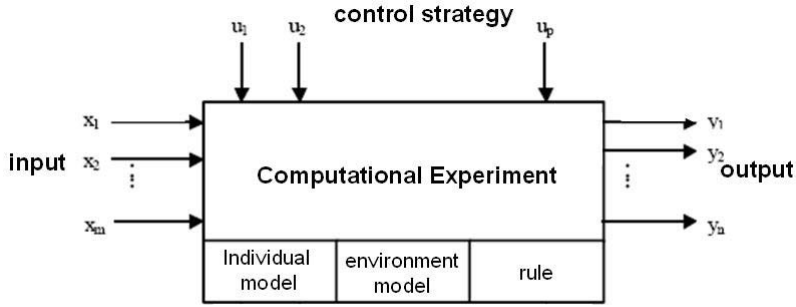


Fig. 8. General model of computational experiments. This mathematical model includes individual model, environment model and evolvement rule. The sets of x_i and y_i are respectively inputs and outputs of emergence computation system, where $i=1,2,\dots,m$, and $j=1,2,\dots,n$. The set of are control factors, where $i_k=1,2,\dots,p$.

ameliorate collective behaviors. Emergence computation has been applied to research emergence phenomenon. Usually setting up emergence computation system is a key and difficult question in emergence computation and computational method of experiments is also a good assistant in it.

6.2 Computational Method of Experiments

Emergence computation is generally characterized by the interaction between relatively simple entities. And from a system view, emergence computation is system exhibiting emergent properties of complex systems. Emergence computation system is bottom-up and there is usually no central control to govern the collective behavior of the system. Typically examples in emergence computation applications include cellular automata, neural networks, genetic algorithms, and agent-based systems [42]. In general, a complex system can also be taken as a large network composed by simple components with no central control, where emergence phenomena are exhibited. Complex systems usually have layered structures and behave large differently different scales.

Computational experiment is a method to research emergence phenomenon in complex systems via setting up a computation system [43]. The purpose of computational experiment is to simulate the behavior of individual in complex system by the method of computer simulation. In the process of it, a group of individuals are let to interact with each other and evolve in the virtual environment constructed by computers, and complex collective behaviors of system can may emerge from button to top [44]. A general model of computational experiments is given by Figure 8.

The process of computational experiment is complicated, which can be divided into following five stages by and large:

- 1) Build up the mathematical models of research objects;
- 2) Research the computation algorithms;
- 3) Set up the platforms of computational experiments;

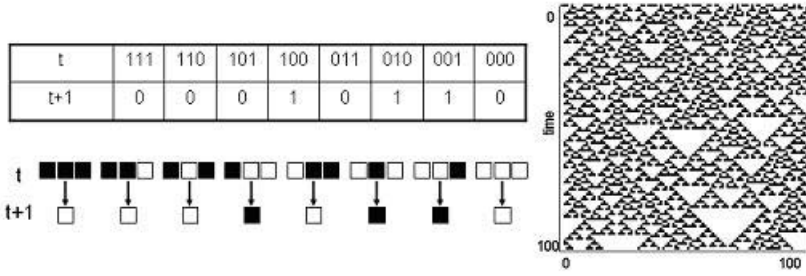


Fig. 9. (The former): Rule table of a cellular automata with periodic boundary conditions. At each time step, state of each cell is updated as a function of the cell own state and those of its two nearest neighbors according to the CA rule; **(The latter): A space-time diagram illustrates typical behaviors of elementary CAs.** The one-dimensional lattice of cells is displayed horizontally and time increases down the figure.

4) Run the computation experiments;

5) Analyze the results of computational experiments, modify mathematical models in goal circumstances and return to step 1) again.

The computational experiment is not only a kind of computation but also a kind of virtual reality. Much work based on this method has been done by us.

Cellular automata (CA) can be considered a good example of emergence computation. A cellular automata, in its simplest incarnation, is a one-dimensional lattice of sites or cells, each of which is colored either black or white, with white coded as “0” and black coded as “1”. The color, white or black, is treated as the state of the cell, and can change with time in discrete steps. In each time-step every cell updates its state—it either maintains its color from the previous step or takes on the other color—as a function of its previous state and that of its two nearest neighbors. The list of mappings of each three-cell neighbors to update the state of center cell is called as the rule table (see Figure 9 (the former)). The initial configuration and rule are both very simple, but cellular automata can emerge complex patterns after a period of time. The space-time behavior of this cellular automata starting with a random initial configuration is given by Figure 9 (the latter), when it runs on a lattice of 100 cells after 100 time-steps.

As a case study about emergence, the spontaneously synchronized clapping in concert halls is a typical case which can be perceived directly by the human. Maybe everybody has such an experience: after a good performance, the audience bursts into a thunderous applause, and in a very short time, this tumultuous clapping suddenly turns into a rhythmic clapping, as if the audience were driven by certain mysterious force. This is a typical self-organizing synchronization phenomenon. Synchronized clapping offers a wonderful example of emergence.

In order to investigate and uncover the emergence mechanism in details, we construct a mathematical model of interaction among audience brought on by conformity psychology of people. We take sound as the representation to

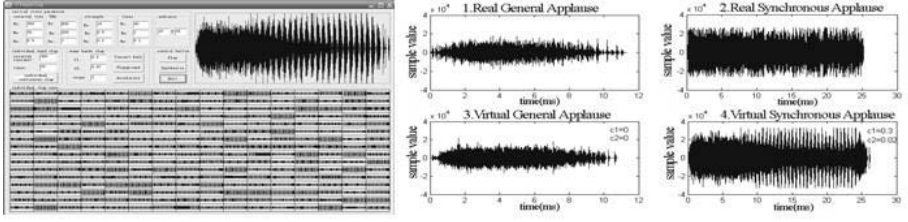


Fig. 10. (The former): Experimental platform of Clapping. We can set the initial state of clapping, select different coupling modes, adjust the coupling coefficient, and get different types of emergence, such as general applause, spontaneous synchronous applause; **(The latter): Comparison between virtual applause and recorded applause.** It is difficult to distinguish the virtual general applause from the real one in sound effect and wave. It is also difficult to distinguish the virtual synchronous applause from the real one in sound effect, although there is little difference in sound wave because of noise.

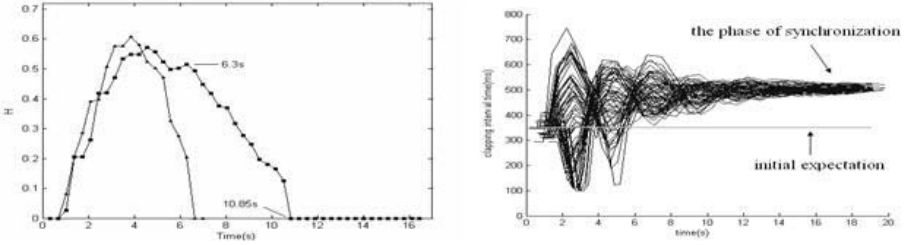


Fig. 11. (The former): Comparison of entropy curves of courteous applause and spontaneous synchronization applause in concert halls. Triangle (▲) represents Courteous applause and rectangle (■) represents spontaneous synchronization applause; **(The latter): Clapping interval time slows down in the process of synchronization.** The line is initial expectation and a group of curve is Clapping interval time of many hands clapping in phase of synchronization. The Clapping interval time in phase of synchronization is greater than initial expectation the phase of synchronization.

simulate the applause in concert halls, and set up an experimental platform of emergence computation which solely depends on interaction rather than global coordination. The virtual applause is found to be hardly distinguishable from the real one (see Fig. 10).

By analyzing data in the process of many hands clapping, we discover the critical area of entropy function from chaos to order, the clapping interval time slow down in the process of synchronization, the degree of which is related with coupling intensity of clapping time (see Fig. 11). Compared with “period doubled” result of paper-the sound of many hands clapping authored by Nédá, Z. and colleagues published in 2000 in *Nature* [45], our results are more universal.

7 Conclusion

It is of great significance to research network topology as two-dimension knowledge representation. In the process of solving real problems, the objects are firstly represented as nodes, and the relationships between objects are represented as links, and the locations and characteristics of nodes and links can all be endowed with given physical rights respectively. The nodes in network stand for various entities; the characteristics of nodes can be expressed referring to particle's characteristic in physics (Amplitude, Frequency and Phase and so on). So dynamics mechanism and prevention of attack-induced cascading breakdown in complex networks can be studied, which is called Networked Intelligence. It differs from traditional AI, it is also known from resolving network route arithmetic, web search engine, distributed computer and so on.

Topology is a special type of figure for human intelligence to represent knowledge and understand them; it is largely different from symbolism. In the past 50 years, however, attention on interacted and evolutionary networks with uncertainty was less paid unfortunately. In fact, Social networks are built upon information networks which built upon communications networks which in turn are built on physical networks. Moreover, this layered structure of interaction built over other networks is reflected directly in the diversity of communities: sociologists, managements, economists, biologists, chemists, physicists, and a wide variety of engineers. Networks evolution and growth drive the fundamental issue that forms our view of network representation and network intelligence. We are witnessing a revolution in the making as scientists from all different disciplines discover that complexity has a strict architecture.

Nowadays, network becomes the engine of scientific research activities in 21st century. Networks interact with one another and are recursive. We have come to grasp the important knowledge of networks. Network is the key to representing the complex world around us. Small changes in the topology, affecting only a few of the nodes, can open up hidden doors, allowing new possibilities to emerge.

In summary, while network thinking and network intelligence is considered in this article, it is always stressed and focused on two kernel ideas, one is global first (i.e. mainly taking the total nodes, links of network as an integration although it is complex, open and dynamically evolved), and the other is topology first (i.e. mainly concerning the interaction and inter-affect among nodes through links and therefore the self-organization, self-similarity and emergency features). The work now we are engaged in is to research and develop cognitive physics, taking network topology as a novel approach of knowledge representation, study various interesting phenomenon in complex network, and do our best to find breakthrough in some fields in AI, especially in the field of cognitive science.

Acknowledgments

This work is partially supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2007CB310804 and the National Natural Science Foundation of China under Grant No. 60675032, No. 60496323.

References

1. Gallagher, R., Appenzeller, T.: Beyond Reductionism. *Science* 284, 79 (1999)
2. Service, R.F.: Complex Systems. Exploring the Systems of Life Science 284, 80–81 (1999)
3. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks. *Adv. Phys.* 51, 1079–1187 (2002)
4. Albert, R., Barabási, A.L.: Statistical Mechanics of Complex Networks. *Review of Modern Physics* 74, 47–91 (2002)
5. Alber, R., Jeong, H., Barabási, A.L.: Diameter of the World-Wide-Web. *Nature* 401, 130–131 (1999)
6. Newman, M.E.J.: Random Graphs as Models of Networks. In: Bornholdt, S., Schuster, H.G. (eds.) *Handbook of Graphs and Networks*, pp. 147–169 (2003)
7. Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. *Science* 286, 509–512 (1999)
8. Wang, X.F., Chen, G.R.: Complex Networks: Small-World, Scale-Free, and Beyond. *IEEE Circuits and Systems Magazine* 3, 6–20 (2003)
9. Watts, D.J., Strogatz, S.H.: Collective Dynamics of “Small-World” Networks. *Nature* 393, 440–442 (1998)
10. Milgram, S.: The Small World Problem. *Psychology Today* 2, 60–67 (1967)
11. Cancho, R.F., Sole, R.V.: The Small-World of Human Language. *Proc. R. Soc. London* 268, 2261–2265 (2001)
12. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-Law Relationships of the Internet Topology. In: *SIGCOMM*, pp. 251–262 (1999)
13. Guha, S., Rasogi, R., Shim, K.: An Efficient Clustering Algorithm for Large Databases. In: *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data* (1998)
14. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco (2001)
15. McCarthy, J.: *Formalizing Common Sense*. Ablex Publishing Co. (1990)
16. Thorsten, J.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Proc. European Conference on Machine Learning*, pp. 170–178 (1997)
17. Han, J., Fu, Y.: Discovery of Multiple-level Association Rules from Large Databases. In: *Proc. Int. Conf. Very Large Data Bases*, pp. 420–431 (1995)
18. Knorr, E.M., Ng, R.: Algorithms for Mining Distance-based Outliers in Large Datasets. In: *Proc. Int. Conf. Very Large Data Bases*, pp. 392–403 (1998)
19. Mitchell, M.: Analogy-making as a Complex Adaptive System. In: Segel, L.A., Cohen, I.R. (eds.) *Design Principles for the Immune System and Other Distributed Autonomous Systems*, p. 335. Oxford University Press (2001)
20. Newman, M.E.J.: The Structure and Function of Complex Networks. *SIAM Review* 45, 167–256 (2003)
21. Li, D.Y., Chen, H., Fan, J.H., Shen, C.Z.: A Novel Qualitative Control Method to Inverted Pendulum Systems. In: *The 14th International Federation of Automatic Control World Congress* (1999)
22. Gregory, C., Walsh, H.Y.: Scheduling of Networked Control Systems. *IEEE Control Systems Magazine*, 57–65 (2001)
23. Ball, P.: Material Witness: Designing with Complexity. *Nature Materials* 3, 78 (2004)
24. Barabási, A.L.: *Linked: The New Science of Networks*. Perseus (2002)

25. Mitchell, M.: Complex Systems: Network Thinking. *Artificial Intelligence* 170(18), 1194–1212 (2006)
26. Cohen, R., Havlin, S., Avraham, D.: Efficient Immunization Strategies for Computer Networks and Populations. *Physical Review Letters* 91, 247–901 (2003)
27. Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies (iT) from the Stand-Point of Web Intelligence (WI). *Communications of the ACM* 50(3), 89–94 (2007)
28. Zhong, N., Liu, J., Yao, Y.Y., Ohsuga, S.: Web Intelligence (WI). In: *Proc. the 24th IEEE Computer Society International Computer Software and Applications Conference*, pp. 469–470 (2000)
29. Zhong, N., Liu, J., Yao, Y.Y.: *Web Intelligence*. Springer, Heidelberg (2003)
30. Yao, Y.Y.: Web Intelligence: new frontiers of exploration. In: *Proceedings of the 2005 International Conference on Active Media Technology*, pp. 3–8 (2005)
31. Zhong, N., Liu, J., Yao, Y.: In Search of the Wisdom Web. *IEEE Computer* 35(11), 27–31 (2002)
32. Zhong, N.: Web Intelligence meets brain informatics: An impending revolution in WI and Brain Sciences (an extended abstract). In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) *AWIC 2005. LNCS (LNAI)*, vol. 3528, pp. 23–25. Springer, Heidelberg (2005)
33. Cooley, R., Mobasher, B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. In: *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence* (1997)
34. Han, J., Chang, C.C.: Data Mining for Web Intelligence. *Computer IEEE*, 54–60 (2002)
35. Christian, S., Doermann, D., Rosenfeld, A.: Classification of Document Pages Using Structure-based Features. In: *IJDAR*, pp. 232–247 (2001)
36. Stephanie, F.: *Emergent Computation*. The MIT Press vol. 21 (1991)
37. Li, D.Y.: Emergent Computation: the Virtual Reality from Chaos to Order. In: *2006 Chinese Conference on Complex Networks* (2006)
38. Luo, J., Liu, Z.: From Synchronization to Emergence. *Complex System and Complexity Science* 2(1) (2005)
39. Holland, J.H.: *Emergence: From Chaos to Order*. Perseus Books (1998)
40. Holland, J.H.: *Hidden Order*. Addison-Wesley, Reading (1995)
41. Holland, J.H.: *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT Press, Cambridge (1992)
42. Brunner, K.A.: What's Emergent in Emergent Computing? In: Trappl, R. (ed.) (2002)
43. Geweke, J.: Computational Experiments and Reality. In: *DYNARE Conference* (2006)
44. Wang, F.Y.: Computational Experiments for Behavior Analysis and Decision Evaluation of Complex Systems. *Journal of System Simulation* 16(5) (2004)
45. Néda, Z., Ravasz, E., Brechet, Y., Vicsek, T., Barabási, A.L.: The Sound of Many Hands Clapping-tumultuous Applause Can Transform Itself into Waves of Synchronized Clapping. *Nature* 403, 849–850 (2000)

Web Intelligence Meets Brain Informatics at the Language Barrier: A Procrustean Bed?

Nick Cercone

Faculty of Science & Engineering
York University
Toronto, Ontario, Canada
ncercone@yorku.ca

Abstract. We take it for granted that computers hold answers to our questions, our information requirements, our needs over the past twenty five years we have learned much about language, about databases, and about how people interact with computers; researchers have made great strides in the construction of human computer interfaces which (relatively) seamlessly integrate modalities, for example, speech and written language, natural language and menu systems, and so on. The next generation of interfaces and browsers, in order to be considered successful, must do more: they must individualize frameworks of meaning in order to provide relevant timely responses to information requests. I want to make several points, perhaps circuitously, but directed as examining some basic tenets regarding our faith in machines. I direct your attention to several problems inherent in representation(s) required to place information into machines for easy (individualized) access, followed up by some larger questions about the inherent capabilities of machines (versus humans).

1 Introductory Remarks

I am bemused. In part, because reflection as broadly as is necessary to make sense what it means for Web intelligence to meet brain informatics, to individualize and comprehend frameworks of meaning in our rapidly changing environment, requires time. In part, because the same blessing that allows each and every one of us to reach out for information instantly is also the curse of responsibility: we must decide how, why and when to use this marvelous blessing.

When I was 8 years old and growing up in Pittsburgh, my friends and I would often sleep on our enclosed front porches in the swelter of the summer's evening heat, gently rocking to sleep on the large porch swing suspended by

chains from the porch roof. Sometimes we would wake in the middle of the night and climb the neighbor's mulberry and peach trees for midnight snacks, only to always have our actions discovered by our mothers in the morning. It took some time for us to realize that mulberries often left telltale purple marks around our mouths after a late night feast. On other nights, I would wander, by myself, to Moore Playground a kilometer away and lie in the centre field grass in the big playground ball field where the semipros played. At midnight no games were being played and the stands were deserted. I wondered then, as I looked up at the stars (yes the renaissance had occurred in Pittsburgh by then after years of industrial neglect of the environment and you could actually see the stars at night). I pondered the questions then that have basically bothered me ever since: How far was far and how big was big? I came to believe that as a grown-up I would ultimately know the answers to these questions. Much later, as a young man at a small Franciscan College in Ohio, I took just about every philosophy course that would fit into an engineering science curriculum that was possible, still hoping to find some answers to these and similarly important questions.

How true it is that language creates special worlds. And systems that can communicate in natural ways and learn from interactions are key to long-term success in Web intelligence. By focusing directly on the Web, researchers in traditional computational (artificial) intelligence areas can help in developing intelligent, user-amenable Internet systems.

The demands of the interactive, information-rich World Wide Web will challenge the most skillful practitioner. The number of problems requiring Web-specific solutions is large, and solutions will require a sustained complementary effort to advance fundamental machine-learning research and to incorporate a learning component into every Internet interaction.

Still, natural language embodies important modalities for human-computer interactions, from simple database interfaces and machine translation to more general answer-extraction and question-answering systems.

The editors of this volume would have us believe that "the synergy between Web intelligence (WI) with brain informatics (BI) will yield profound advances in our analyzing and understanding of the mechanism of data, knowledge, intelligence and wisdom, as well as their relationship, organization and creation process." Our use of language should put this hypothesis to the test.

2 Language and Artificial Intelligence

Most languages are inaccessible to most of us most of the time¹ - we believe that the language of the Eskimo and Inuit describing the many states of snow

¹ Lewis Carroll once wrote: "I'm so glad that I don't like asparagus" said the small girl to a sympathetic friend. "Because, if I did, I should have to eat it, and I can't bear it."

is inaccessible², but we need not feel a loss. Generally what holds for language holds for life also. It is in this sense that language can serve as a mirror in our investigations into the nature of cognitive capabilities so necessary for Web intelligence. For it is not so much why something happens, or how it occurs, as it is to understand why we perceive things to be the way that they are or how we plan activities to occur or even what we ruminate in between all other thoughts that generally holds our interest.

However, for most of us, the world is a world of matter - wysiwyg.³ The superiority of physics to, say, interpersonal communication, massage, etc. derives from the assumption that if we are able to explain the physical, we may be in a position of explaining everything else.

So where is artificial intelligence, Web intelligence, brain informatics situated? In the days of good old fashioned AI [GOF AI], the quest was to find a general intelligence. Representation schemes were recognized as important, serving as the structure(s) by which the systems we built (we did build large programs rather routinely then) could be extended to cover more and more of a particular domain and, if we were lucky, extend to another domain as well. Then something curious appears to have happened. A necessary step in the name of progress in artificial intelligence was to stop experimenting, and start becoming smarter about what systems we were going to build. We began designing various logics for specific purposes, never yet getting back to connecting them all up with the original quest for finding a more general intelligence (surely this is what Turing had in mind when he proposed that extremely boring party game known now as the Turing test - why I haven't even heard that term used for about 10 years now). Is history repeating itself?

Consider the fine earlier work in theorem proving research: after a while researchers gave up trying to use results of this research as the basis of natural language understanding programs. Now we find theorem proving research results embedded in constraint logic programming systems which are at the heart of many fine natural language efforts. When computer science was born, arguably from numerical analysis research, every computer science department

² Geoffrey K. Pullam's book "The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language", 1991, University of Chicago Press] on the great eskimo vocabulary hoax, notwithstanding. Pullam actually wrote "Once the public has decided to accept something as an interesting fact, it becomes almost impossible to get the acceptance rescinded. The persistent interestingness and symbolic usefulness overrides any lack of factuality. . . . In the study of language, one case surpasses all others in its degree of ubiquity, and the present chapter is devoted to it; it is the notion that Eskimos have bucketloads of different words for snow. . . . But the truth is that the Eskimos do not have lots of different words for snow, and no one who knows anything about Eskimo (or more accurately, about the Inuit and Yupik families of related languages spoken by Eskimos from Siberia to Greenland) has ever said they do. Anyone who simply insists on checking their primary sources will find that they are quite unable to document the alleged facts about snow vocabulary (but nobody ever checks, because the truth might not be what the reading public wants to hear).

³ What you see is what you get.

had numerical analysis courses, teachers and researchers. It would be heresy not to have them. Then the unthinkable happened. The numerical analysts became part of applied mathematics departments in a large part of the university systems and pursued more and more fine grained results studying error analysis to the m^{th} degree. Not only could your space vehicle miss its intended destination, but we could tell you exactly by how much you missed it. Over the last decade, numerical analysis has been reborn and its importance reaffirmed, especially as the pendulum of university research swings to more and more applied research.

Is this what is happening to the quest for generality. Sadly, I think the situation is different in this case. Essentially, I am becoming a skeptic, not of the value of artificial intelligence but of the time frame we have given ourselves to produce an artificial intelligence, hence the importance to Web intelligence, and of the need I am convinced of, of encouraging truly multidisciplinary teams of researchers to tackle some of our problems. Let me cite some examples from natural language understanding, admitting that in each case, we could probably devise a system to tackle the particular problem, but generalizing the solution may well prove elusive.

It is common in the best of the Yiddish tradition to answer a question with a question, often the very same question with different emphasis and intonation. For example, in response to the question "Did you buy flowers for your mother on her birthday?" the response would be quite different if a different word were emphasized in the answer. Thus the answer "*Did* I buy flowers for your mother on her birthday?" is quite different from "Did *I* buy flowers for your mother on her birthday?" which is different from "Did I *buy* flowers for your mother on her birthday?" and so on.

Imagine the reasoning required to interpret the following passage and correctly ask the question which Mr. Rosenberg asks of Murray Goldwag at the end of the passage.

One fine Friday afternoon, Murray Goldwag leaves his place of employment in Brooklyn to catch the bus to Wappinger's Falls upstate to spend the weekend with his finance Lennie Rosenfeld. During the trip and elderly gentleman, Mr. Rosenberg, returning to his home in Wappinger's Falls from a visit with his brother in Brooklyn, strikes up a conversation with Murray. "So, you're up to the Falls for the weekend"? "Yes sir", replied Murray. "By yourself?" "Just visiting friends", replied Murray. Mr. Rosenberg then reflects for a while. He is visiting friends, leaving early enough to arrive in time for a late dinner and to get himself a motel room. He must be visiting a girl friend since no one would go out of their way so much for another boy friend. Now who could he be visiting? He is a handsome lad. Could it be the Goldberg twins? - No, Meryl is away at University out of state and Fran is out of town visiting relatives. Could it be Maxine Kriebel? - No, she has a boyfriend. What about Melinda Eaman? Probably not, she is very rich and would not be going out with any young man who had to travel by bus in order to see her. Well, what about Sarah Lavie? That could be the one, Sarah has been on cloud nine recently and acting mysterious, visiting all the shoppes and making preparations for a big day soon. She is not graduating from College, nor celebrating a promotion or new job. That's it, thought Mr. Rosenberg, Murray must be coming to Wappinger's Falls to see Sarah and make plans for their upcoming wedding. Mr. Rosenberg turns to Murray and says "Congratulations, son, on your upcoming

wedding to Sarah Lavie.” “wha—what”, said Murray. “How did you know, we haven’t even told her parents yet.” “My boy”, said Mr. Rosenberg, “its obvious.”

These two examples of sublanguages of English occur with alarming frequency and we are equipped to handle these and other statements which we have never heard before as examples of “simply miraculous machines.” Pity the poor computer, however. If you do not like these examples, I am sure that I could supply hundreds of other examples, perhaps more straightforwardly, but equally obtuse to our efforts thus far in natural language understanding. Consider the following excerpt from Erle Stanley Gardner’s “The Case of the Demure Defendant”:

“Cross-examine.” Hamilton Burger snapped at Perry Mason.

Mason said, “Mr. Dayton, when you described your occupation you gave it as that of a police expert technician. Is that correct?”

“yes sir.”

“What is an expert technician?”

“Well, I have studied extensively on certain fields of science that are frequently called upon in the science of criminology.”

“That is what you meant by an expert technician?”

“Yes sir.”

“Now what is a police expert technician?”

“Well that means that. . . well, it all means the same thing.”

“What means the same thing?”

“An expert technician.”

“An expert technician is the same as a police expert technician?”

“Well I am in the employ of the police department.”

“Oh the police employ you as an expert witness, do they?”

“Yes sir, . . . I mean no, sir. I am an expert investigator, not an expert witness.”

“You are testifying now as an expert witness are you not?”

“Yes sir.”

“Then what did you mean by saying you were an expert technician but not an expert witness?”

“I am employed as a technician but not as a witness.”

“You draw a monthly salary?”

“Yes.”

“And you are being paid for your time while you are on the stand as an expert witness?”

“Well, I’m paid for being a technician.”

“Then you won’t accept any pay for being a witness?”

“I can’t divide my salary.”

“So you are being paid?”

“Of course - as part of my employment.”

“And are you now employed by the police?”

“Yes.”

“And are you an expert witness?”

“Yes.”

“Then you are now being employed as an expert witness.”

“I guess so. Have it your own way.”

“When you described yourself as a police expert technician that means your testimony is always called by the police. Isn’t that so?”

“No, sir.”

“Who else calls you?”

“Well, I . . . I could be called by either party.”

“How many times have you been on the witness stand?”

“Oh, I don’t know. I couldn’t begin to tell you.”

“Dozens of times?”

“Yes.”

“Hundreds of times?”

“Probably.”

“Have you ever been called by the defense as a defense witness?”

“I have not been directly subpoenaed by the defense. No, sir.”

“So that you have always testified for the police, for the prosecution?”

“Yes, sir. That’s my business.”

“That was what I was trying to bring out,” Mason said. . . .

Mr. Dayton needs to understand the subtleties of noun phrases such as “police expert technician”, to answer Mr. Mason’s questions. Understanding such phrases are troublesome to automate since “police”, “expert” and “technician” are all nouns.

Generalizing semantic considerations for such constructions have proven evasive. The compositional approach to natural language understanding favored by logic grammarians becomes combinatorially explosive. Many researchers represent noun-noun constructions as single lexical entries, constraining the computation required to disambiguate them, and circumventing an annoying semantics problem.

When the domain of discourse is well specified and the number of such phrases is small, this approach works adequately. But is it practical? Consider “western region outage log” employed by telecommunications service personnel. Would the designer of their system resort to separate lexical entries for “eastern region outage log”, “southern region outage log”, “northern region outage log”, “northeastern region outage log”, ..., “western district outage log”, ..., “western prefecture outage log”, ..., “western region service log”, ..., “western region outage record”, ...?

Imagine further, the processing required by Perry Mason. Not only must the subtleties of language understanding realized by Mr. Dayton be mastered but also the reasoning capabilities of Mr. Mason and extraction of relevant and salient features of the conversation be identified in order to generate the appropriate next question. Actually, Mr. Mason’s task is simpler than Mr. Dayton’s - to generate an utterance which conveys a presumably preexisting thought. Mr. Dayton’s task as listener is to decide what Mr. Mason must have been thinking in order to motivate his utterance in the particular context in which he uttered it.

Language is difficult; humans are amazing. By the time you have completed reading this sentence you will have understood its meaning. Your achievement and success in understanding is most impressive. The speaker’s task is much simpler - to generate an utterance which conveys a presumably preexisting thought. Your task as listener is to decide what the speaker must have been thinking in order to motivate his utterance in the particular context in which he uttered it. In general, understanding a natural language (NL) is simply miraculous.

NL represents an important modality for human computer interactions, from simple NL interfaces to databases to machine translation to more general answer-extraction and question answering systems. Other important modalities, e.g., speech, pointing devices, graphical user interfaces, etc. remain. The perfection and integration of multimodal systems takes on new importance when we transpose previous solutions to the Internet. Systems which can communicate in natural ways and can learn from interactions are key to long term success transferring *computational to Web intelligence to brain informatics*.

3 Traditional Natural Language Applications

How long will it be before we have systems that can process language as illustrated in the three examples above? It will be instructive to look at the past.

At Roger's Cablesystems Ltd., the vice president for customer service enters the following into his computer terminal, "Give me the Western region outage log for June". Within seconds SystemX [1,2] presents him with a neatly formatted table (or graph) of the data retrieved from Rogers' relational database. He could have said, "What's the outage log for the Western region for June?", or "Tell me the June regional outage log for the West." or "Find the Western outages for June.", etc. SystemX can determine that whichever phrase he uses, he means the same thing. Such flexibility in parsing, applying the logical rules of grammar to determine meaning, is nontrivial. SystemX's parsing techniques are described in [3]. After parsing, SystemX reformulates the question in SQL (structured query language) and data is extracted for presentation from Roger's large central database.

The nontrivial problem described in the preceding paragraph is but one of a large number of very difficult problems of understanding NL by computer. Fortunately, a NL interface is simpler to comprehend. Although one ultimately encounters problems comparable to the unconstrained NL understanding situation, the domain of discourse, and thereby the context, is highly constrained by the database schema. General analysis of language phenomena and much of the ambiguity inherent in NL understanding is limited but complexities arise when building NL capabilities into database interfaces. One quickly comes to realize that domain knowledge is required in order to interpret queries, in order to answer queries, and that modeling the user is important as well.

An example of SystemX accepting an English query from Rogers' vice president, translating the query into SQL, retrieving data from Rogers' database and displaying the data in the format (table or graphical trend) specified by the user in the query is shown in Figure 1. SystemX is able to display responses to requests for trends in statistical data graphically. The user has the choice of inputting his trend request using English, using menus (in the case of "canned" trends) or using a combination of English and menu responses. Various input modalities are provided as a convenience to users. The "canned" trends display data that is predictably desired on a reasonably frequent basis. They may be accessed for a minimum of keystrokes. The "canned" trends are those available through the first eight menu items in the Trend Menu in Figure 1.

Specifying a request for a trend in English may become quite cumbersome if default parameters (specifying timing and so forth) are not employed. The complex statements required are difficult to formulate and demand patience on the part of the users while waiting parsing. The system therefore allows the users to request ad-hoc trends using a combination of English and responses to menus. This combination of modality reduces the task of specifying a complex query into a set of simple tasks that are accomplished in sequence. The system accesses the database in order to be able to present tasks to users in as helpful a manner as possible.

Despite the many search engines available, searching for a relevant site remains difficult. One major reason for this difficulty is that search engines do not

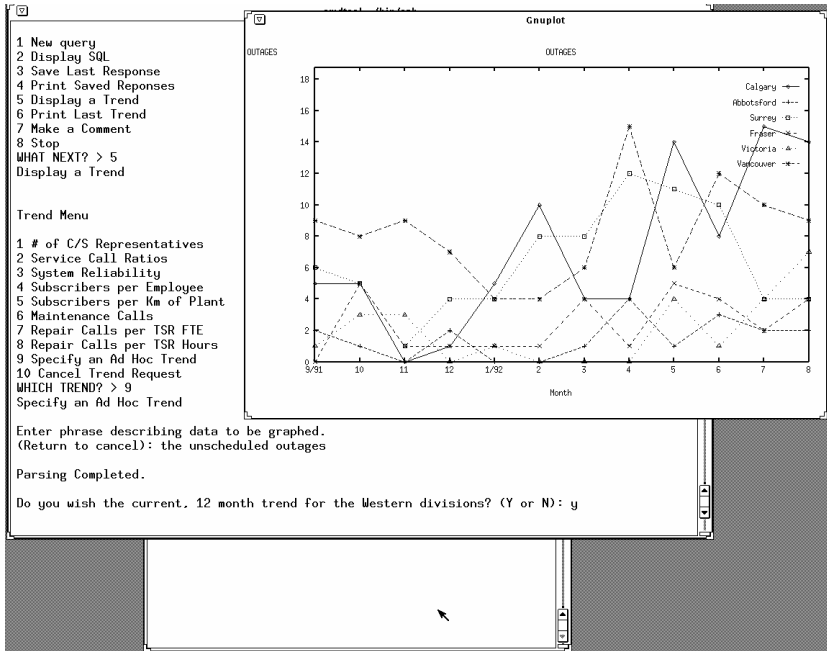


Fig. 1. The SystemX main menu and a trend menu

analyze queries semantically; in contrast, most search engines perform keyword matching.

How can our use of NL semantics improve Internet searching? SystemX was one common application that provided a NL “front-end”, which enables users to access database information without any need to know database structure or any query language, and with no need for query transformation to some other representation. A NL “front-end” to Internet search engines, which allows users to utilize search engines without finding appropriate search terms, is presented in [4,5]. For a search for: “I want to book a flight ticket” or “Show me some sites on online reservation of flight tickets” or phrases like “online reservation of flight tickets”, these queries would yield the same search results.

NLAISE [4] allows users to choose the search engine best suited for their search and enter the query in English. The NL query is analyzed both syntactically and semantically in order to select the most appropriate keywords describing sought information. Keywords are interpreted to provide more meaningful search terms by using keyword synonyms in conjunction with Boolean operators supported by specific search engines.

In NLAISE, the NL query, along with the choice of search engine, is pre-processed in order to transform the query into a form suitable for input to the parser. The parser, in turn, has a description of grammar rules for capturing the constraints of English and a lexicon that contains the words permitted as input. The Head Driven Phrase Structure (HPSG) parser generates a complex

feature structure representing the query. The semantic content of such a complex feature structure is extracted, interpreted and transformed into a form suitable for the search engine that was selected. In a test NLAISE was asked to parse the phrase “I want to schedule a trip to Japan” and generate appropriate keywords for search engine examination. NLAISE was also requested to use Infoseek as the search engine. Inspection of the 1,473 Web pages returned verified that 80% were relevant. Note the choice of keywords “Japan” and “travel” indicates the level of sophistication of NLAISE’s semantic interpretation of the original input phrase.

EMATISE [5] extended NLAISE in 3 user-oriented ways: (1) whereas NLAISE was tied to a single “travel” domain, EMATISE greatly enhanced semantic interpretation to eliminate much ambiguity and toil over multiple domains; (2) EMATISE sent out term expanded queries to multiple search engines in parallel and reranked results returned from these search engines into a single relevant high precision list for the user; and (3) EMATISE’s higher level of abstraction above conventional search services presented the user with a single, central and natural search interface with which to interact.

Consider the following scenario. Imagine picking up the phone in Toronto, dialing your Japanese program co-chairman in Tokyo to explain several papers lost in the shuffle of email systems. You speak English and she speaks Japanese. Fortunately it is 2010 and the English you speak in Toronto is automatically translated into Japanese in the time it takes to transfer your words over the phone lines. Impossible, - probably not. The world of machine translation has both fascinated and frustrated researchers for over 50 years. Recent success in statistical, nonlinguistic and hybrid systems provide hope that we will not be confined to traditional dominant direct, transfer and intralingual approaches. An informative critique of these approaches is given in [6]. We provide an approach following from CS methodology, generate and repair machine translation. (GRMT).

GRMT (Figure 2) is composed of 3 phases: “Analysis Lite Machine Translation (ALMT)”, “Translation Candidate Interpretation (TCI)” and “Repair and Iterate (RI)”. “ALMT” generates translation candidates (TC) by considering syntactic and semantic differences between language pairs without performing any sophisticated analysis. This ensures that the TC can be generated quickly, simply and efficiently. Next, the system interprets the TC to see if it retains the meaning of the SL. If so, that TC will be considered a translation. If not, that TC will be repaired based on the diagnosis that is indicated in the second phase, TCI. Subsequently the repaired TC will be re-interpreted to determine if it still has a different meaning from the SL. These two processes iterate until the TC conveys the same meaning as the SL. The TCI and RI stages ensure the accuracy of the translation result. They also guarantee the accuracy of the translation back from the TL to SL.

GRMT treats SL and TL separately and is aware of the differences between languages. Therefore, if languages can be grouped according to various characteristics, e.g., auxiliary verb, continuous tenses, passive voice, etc., which they

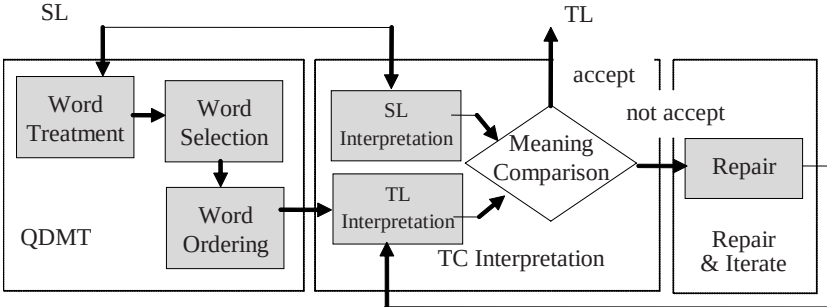


Fig. 2. Organization of Building Block

have in common, then the translation between groups can be performed more simply by GRMT. For example: Group 1 consists of English, French and Spanish, Group 2 consists of Chinese, Japanese and Thai. To perform the translation between these two groups, the transfer approach requires six SL analyzers, six TL generations and 18 sets of transfer rules while GRMT requires six SL TCIs, six TL TCIs and 2 sets of constraint applications.

Table 1. Example of a Generated TC

Example: An old woman lived in the cottage, with a fat black cat and a plump brown hen.

TC: ผู้หญิงแก่ คนหนึ่งไ้ ด้ อยู่ ใน กระโจม หลังนั้นกั๊บ มว สีดำ
อ้วน ตัวหนึ่งแ ละ ไ้ สีน้าตาล อวบ ตัวหนึ่ง
CT: ผู้หญิงแก่ คนหนึ่งไ้ ด้ อยู่ ใน กระโจม หลังนั้นกั๊บ มว สีดำ
อ้วน ตัวหนึ่งแ ละ ไ้ สีน้าตาล อวบ ตัวหนึ่ง
(phúujǐ - woman) (k' - old) (khon- clas) (ny' - an) (dǎj- past) (ju'u- live) (naj- in)
(krath'm- cottage) (la' - clas) (nán- the) (kàb- with) (mw- cat) (si'idam- black)
(?úan- fat) (tua- clas) (ny' - a) (l' - and) (kǎj- hen) (si'inámtaan- brown) (?úab-
plumb) (tua- clas) (ny' - a)

Experiments of ALMT (English to Thai) indicate that TCs can be generated with relative accuracy. Table 1 shows an example of applying ALMT.

4 Brainstorms

Successes mentioned earlier, and others like them, represent contemporary computational intelligence solutions. How do we adapt them to become Web intelligence and brain informatics solutions? We briefly describe current work designed to make useful solutions to computational intelligence problems amenable to such use. Some of this work takes advantage of newer technologies already beginning to show up in Web applications (agent architectures, recommender systems, information extraction tools, etc.). This current work represents an intermediate step along the way to Web intelligence/brain informatics. It necessarily leads to the realization that more adaptable and more general machine learning strategies need developed and incorporated into every aspect of systems. One glaring

example would be learning the meaning of unknown or undefined words, for machine translation and general speech and NL processing.

Java Parsers, Just-in-time Subgrammar Extraction, Modular HPSG's

Stefy is a NL parser implemented in Java, based on HPSGs [7]. It is part of a larger project to implement a NL processing system for Internet information retrieval (IR). This IR task requires Java applets capable of parsing a NL. Earlier we discussed work on developing HPSG parsers. However, Stefy is one of the first implemented in Java. Java was chosen for two reasons. Java supports dynamic class loading and object serialization, which are important features necessary for our concept of distributed NL processing. Java is a good prototyping language, compared to C++ for example, and facilitates easy experimentation with various approaches, which makes this shift in programming language paradigm less drastic.

A drawback of our implementation is that it is not suitable for development of the grammar and lexical resources. Other systems, like ALE [8] and LKB [9], are more appropriate for this task. After a grammar or a lexicon is developed in one of those systems, it is translated into a Java description and used in Stefy.

Stefy represents a new precise and compact description of the HPSG formalism, which is especially suitable for implementation of HPSG parsers in low-level languages. Stefy represents an important step towards applying HPSG formalism in the area of distributed NLP and answer extraction.

Stefy's approach is similar to the filtering techniques, which are a recognized way to improve parser's performance. However, Stefy is different because we insist that the filtered, i.e., extracted, knowledge is in the form of a grammar. This approach is sound, and in practice it provides a clean interface between subgrammar-extraction part and the parser. Keselj gives more arguments for this separation of the subgrammar extraction and parsing [10].

An important part of the HPSG subgrammar extraction is the extraction of the corresponding type sub-hierarchy out of the original hierarchy. Efficient type operations and representation of the types are used in approximate algorithm for subgrammar extraction for HPSGs. Recently, there has been a lot of research activity in the area of grammar modularity. Some of the motivational factors for this work are the following:

- managing complexity. The NL grammars used in NL processing are large and complex. The difficult problems are designing, creating, testing, and maintaining them. Using smaller modules that are combined into larger grammars addresses the complexity problem.
- parsing efficiency. Parsing with a large, wide-coverage grammar is typically not efficient. Quickly extracting a small subgrammar module, and then using it to parse the text can reduce the running-time and space requirements.
- context-based disambiguation. By having a larger grammar we achieve a better coverage, but in the same time it becomes susceptible to ambiguities. Any NL is very ambiguous, and it is well known that humans use world-knowledge and contextual knowledge to do disambiguation. Extracting a subgrammar based on the text to be processed can be viewed as creating a context that can improve disambiguation.

Recommender Systems using ELEM2

Recommender systems suggest information sources, products, services, etc., to users based on learning from examples of their preferences, dislikes, etc. There are two predominant methodologies employed in such systems. *Collaborative (social) filtering methods* base recommendations on other users preferences, e.g., when you order books from Amazon.com, the recommender system may detect other customers who ordered the same books and determine other orders placed by these customers to then enquire whether you may also be interested in acquiring similar material. *Content-based methods* use information about the item ordered/specified in order to make further suggestions to the user. Advantages of content-based methods include the ability to recommend previously unrelated items to users with unique interests and to also provide explanations for recommendations.

For collaborative (social) filtering, we plan to merge information sources to permit more fine-grained analysis and subsequent recommendations. For example, use of the Statistics Canada database on wealth demographics in Canada, which they categorize from richer to poorer by postal code, could conceivably recommend products/services based not only on social preference but also by wealth demographics at the same time.

We especially wish to develop content-based methods since this will provide a new application for ELEM2 [11,12]. Content-based recommender systems provide another unique application for embedded ELEM2. Briefly, a set of documents (Web pages, newsgroup messages, etc.) would have information extracted from an information extraction (word extraction) phase to develop a set of examples. We randomly select a set of examples and choose a subset of these examples from which we determine from a user, positive and negative examples. These positive and negative examples serve as a training set for the user. We apply ELEM2 rule induction process to extract a “user’s profile” and then rank the rest of the examples accordingly. Top ranked examples then serve as an item list for recommendation.

Agents and Agent Architectures

The Internet is a large, distributed, and heterogeneous source of information primarily consisting of on-line World Wide Web documents. It is perceived through a set of applications based on the point-to-point communication links provided by the TCP/IP protocol. Many applications frequently end up with the problem that we want to find a relevant document, relevant item, or, generally, a relevant point in the information space consisting of Telnet sites, news groups, news group postings, FTP (File Transfer Protocol) sites, and WWW documents (pages, movies, radio broadcasts). How can we find out if someone has an e-mail address and how can we find that address? Finding interesting mailing lists is a still better example.

The Internet can be imagined as a low-level structure activated with considerable manual (human) participation. Such an intelligence-assuming environment requires computational intelligence management techniques. The most obvious

example is a simple Web page. If we want to automatically use its content in a fashion more sophisticated than collecting keywords, or collecting embedded links for further navigation, then the most flexible, robust, and appropriate way to do this is to understand some of its content and to reason about it. This is the realm of computational intelligence.

“Agent” has become a computational intelligence term, and a frequent buzzword having a wide range of definitions. Nevertheless, there are some common characteristics that describe an agent. An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment. Furthermore, the development of multi-agent systems is based on work in two areas artificial intelligence and distributed systems.

The combination of NL processing and multi agent system’s is still quite novel and often the terms are used independently. Consider the use of NL processing for information retrieval (IR) over the Internet. This is an attempt to match the meaning of the user’s query to the meaning of retrieved documents. Since this approach relies on higher levels of NL processing, it is difficult to implement. Issues include deciding what is a concept, how to extract concepts from NL texts, and how to do concept matching. The inefficiency of existing NL processing systems is a major obstacle to using them in IR. If we want to use an NLP system to analyze the documents in a large document collection, it has to be efficient and robust to be useful in practice.

A positive approach is to implement distributed NL processing so that the processing cost is widely distributed in the same way as are Internet resources. Multi agent systems are appropriate for this task.

5 Concluding Remarks

Web intelligence/brain informatics requires further research and development into the technologies discussed above and other technologies as well. Adapting existing computational intelligence solutions may not always be appropriate for Web intelligence for a number of reasons, e.g., the magnitude of information available on the Internet and the additional requirements for speedy processing. Computational intelligence solutions which may be adapted must incorporate a more robust notion of learning in order for these solutions to scale to the Web, in order for these solutions to adapt to individual user requirements, and in order for these solutions to personalize interfaces.

We have only briefly touched on a few, albeit important, issues that will be the mainstay of Web intelligence in the near term future. Users will demand access to the Internet that is simple (multimodal interfaces), with language/speech capabilities - both comprehension and, when needed, translation - and personalized (multi agent architectures) Internet use which “learns”.

How soon might we expect to see breakthroughs? One way of considering this question is to recognize that research progress is highly incremental, thus, we are seeing progress every day. I, for one, have great hopes for the future of Web intelligence.

Acknowledgments

The author wishes to acknowledge the support of the Natural Sciences and Engineering Research Council, of Canada.

References

1. Cercone, N., Han, J., McFetridge, P., Popowich, F., Cai, Y., Fass, D., Groeneboer, C., Hall, G., Huang, Y.: SystemX and DBLearn: easily getting more from your Relational Database. invited for Integrated Computer-Aided Engineering. In: Golshani, F. (ed.), vol. 1(4), pp. 311–339 (1994)
2. Popowich, F., McFetridge, P., Fass, D., Hall, G.: Processing Complex Noun Phrases in a Natural Language Interface to a Statistical Database. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992), Nantes, France, pp. 47–52 (1992)
3. McFetridge, P., Cercone, N.: Installing an HPSG Parser in a Modular Natural Language Interface, Computational Intelligence III, pp. 169–178. North Holland, Amsterdam (1991)
4. Mahalingam, G., Cercone, N.: Finding Information Easily is Important for e-Business. In: Kou, W. (ed.) Data Warehousing and Data Mining for Electronic Commerce, pp. 135–168. IBM Press (1999)
5. Hou, L., Cercone, N.: Extracting Meaningful Semantic Information with EMATISE: an HPSG-based Internet Search Engine Parser. In: NLPKE Symposium of the IEEE SMC Conference, Tuscon, AZ (2001)
6. Naruedomkul, K., Cercone, N.: Generate and Repair Machine Translation. Computational Intelligence 18(3), 254–270 (2002)
7. Keselj, V.: Stefy: Java Parser for HPSGs, Version 0.1, Technical Report CS-99-26, Department of Computer Science, University of Waterloo, Waterloo, Canada (2000)
8. Carpenter, B., Penn, G.: ALE, the attribute logic engine, User's Guide (1999), www.sfs.nphil.unituebingen.de/~gpenn/ale.html
9. Copestake, A.: The (new) LKB system, Version 5.2 (1999)
10. Keselj, V.: Just-in-time subgrammar extraction for HPSG, Technical Report CS-2001-08, Computer Science, University of Waterloo, Waterloo, Canada (2001)
11. An, A., Cercone, N.: ELEM2: A Learning System for More Accurate Classifications. In: Mercer, R.E. (ed.) Advances in Artificial Intelligence. LNCS, vol. 1418, pp. 426–441. Springer, Heidelberg (1998)
12. An, A., Cercone, N.: Discretization of Continuous Attributes for Learning Classification Rules. In: Zhong, N., Zhou, L. (eds.) Methodologies for Knowledge Discovery and Data Mining. LNCS (LNAI), vol. 1574, pp. 509–514. Springer, Heidelberg (1999)

Conversational Informatics Where Web Intelligence Meets Brain Informatics

Toyoaki Nishida

Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract. Conversational Informatics is a field of research that focuses on investigating human conversational behaviors and designing conversational artifacts that can interact with people in a conversational fashion. It is aimed at unveiling meaning creation and interpretation through sophisticated mechanisms of the verbal / nonverbal interactions during conversations. I overview major ideas and outcomes of Conversational Informatics and discuss the role of Conversational Informatics as a glue connecting Web Intelligence and Brain Informatics.

1 Introduction

Conversation is the most natural and popular means for people to communicate with each other. Conversations play an important role not only in achieving goals shared by participants or coordinating behaviors of participants, but also in facilitating knowledge creation through such mechanisms as cooperative storytelling or externalizing tacit knowledge. People, mostly in a quite skillful way, combine verbal communication means with nonverbal communication means, such as eye gaze, facial expressions, gestures, etc, in conversation. Conversation encompasses a wide spectrum of interesting phenomena residing at the intersection of the information world and brain activities. Conversation is an effective means for communicating our thought. On the other hand, conversation plays an important role in forming our thought. Embodiment is closely related to both thought and communication.

Conversational Informatics sheds light on the intersection of Web Intelligence and Brain Informatics where entities in the information world are created or interpreted through conversational interactions resulting from brain activities. From the viewpoint of Web Intelligence, Conversational Informatics seeks for better methods of computer-mediated communication, human-computer interaction, and support for knowledge creation that are key to make Web Intelligence effective in the knowledge community. On the other hand, Conversational Informatics can be characterized as a new research area for studying how brain activities manifest as communicative activities.

Conversational Informatics is built up on the foundation of Artificial Intelligence, Pattern Recognition, and Cognitive Science, and is aimed at integrating

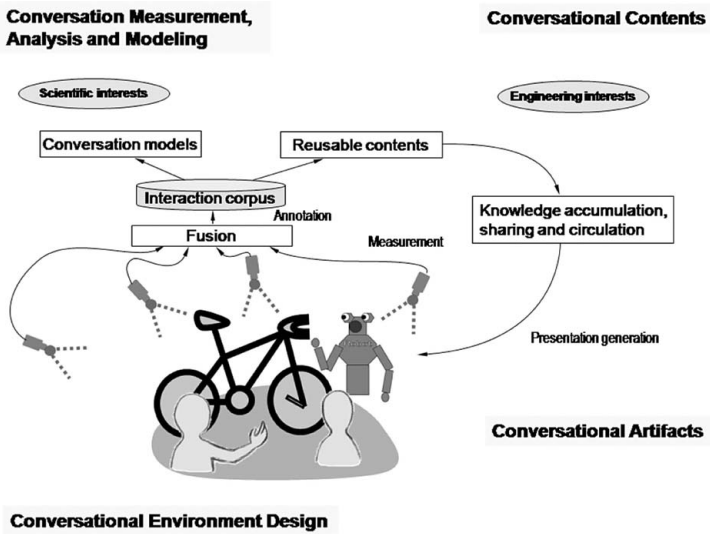


Fig. 1. Conversational Informatics

understanding human conversational behaviors and designing conversational artifacts that can interact with people in a conversational fashion. Current technical developments in conversational informatics center around four subjects, as shown in Figure 1. The first subject is conversational artifacts for mediating human-human communication. The second subject is conversational content that is taken from conversational situations and allows conversational artifacts to produce utterances. The third subject is conversation environment design for capturing and applying conversational contents through conversation. The last subject is conversation measurement, analysis, and modeling that permit us to bring about in-depth understanding of conversational communication.

In this chapter, I present the conceptual framework and major achievements of Conversational Informatics, with particular emphasis on how it meets Web Intelligence and Brain Informatics. I also present some future perspectives.

The rest of this chapter is organized as follows. In the following four sections, I will describe the four subjects of Conversational Informatics in more detail. In section 6, I will discuss the role of Conversational Informatics as glue connecting Web Intelligence and Brain Informatics and give some future perspectives.

2 Conversational Artifacts

The role of conversational artifacts is to mediate the flow of conversational content among people. Major contributions on this subject involve content-driven embodied conversational agents and communicative robots as knowledge media.

2.1 Content-Driven Embodied Conversational Agents

Embodied conversational agents (ECAs) are human-like characters that can interact with the user in a conversational fashion. In order to design an ECA that can mediate inter-human communication, we need to consider many aspects including ease of content management, a virtual environment that provide the context and referents for conversation, and implementation of higher order social concepts such as politeness, friendliness, humor, personality or trust.

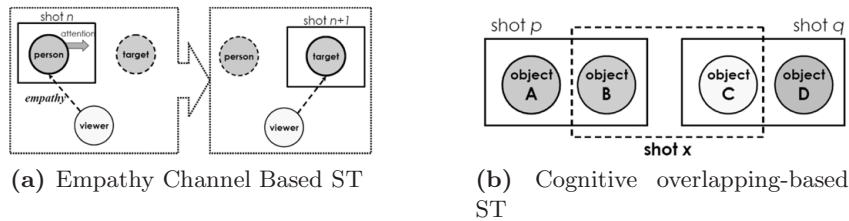


Fig. 2. Shot transitions based on based on the ideas of empathy channel and cognitive overlapping [Okamoto 2005]

We placed particular emphasis on implementing a content-rich ECA, for we believe the users are the most interested in content, rather than other aspects of an ECA, particularly in the context related to knowledge creation. In order to reduce the overhead of knowledge transfer to ECAs, we have developed a number of content-driven ECAs. The earliest system was EgoChat that could take a series of knowledge cards and create a simple presentation with an animated cartoon-like character and a speech sound created with a text-to-speech engine [Kubota 2002]. A powerful dialogue engine is incorporated in EgoChat which permits an EgoChat agent to answer questions by searching for the best match from a potentially large list of question-answering pairs prepared in advance.

The framework of EgoChat has been extended in the subsequent systems called SPOC and IPOC. In SPOC (Stream-oriented Public Opinion Channel), a method of generating nonverbal presentation from the plain-text representation is employed [Nakano 2004]. IPOC permits the user to interact with the embodied conversational agent (the IPOC agent) embedded in a panoramic semi 3D virtual space [Okamoto 2005]. The IPOC agent can move around and talk about objects in the virtual environment. In addition, a camera-work is introduced to enhance the user involvement in the virtual environment.

In order to create user involvement, we introduced a couple of cognitive constraints on determining shot transitions, called empathy channel and cognitive overlapping. An empathy channel is a relationship with an ECA on the screen that allows the user to regard the ECA as her/his surrogate in the virtual world who provides the user with a reference point for acquiring or producing information in the virtual world. An ECA is called an empathized agent when an empathy channel is established with the user. In order for the user to be involved in the virtual world, various perceptual means should be provided for the

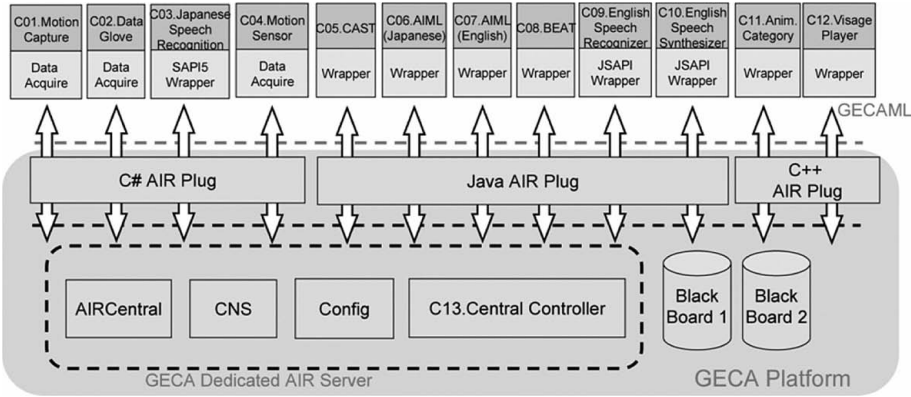


Fig. 3. The GECA platform [Huang 2006]

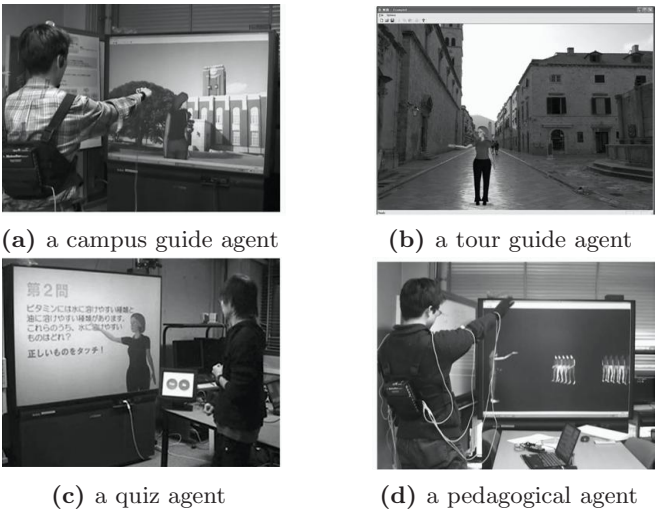


Fig. 4. ECAs built on top of GECA [Huang 2006]

user to regard an ECA on the screen as an empathized agent. It is considered that the effective use of the back image helps the user establish an empathy channel with an ECA. On the other hand, cognitive overlapping is the overlapping based on figure-ground alteration. Cognitive overlapping is considered to help to achieve cognitive reality. Figure 2 illustrates shot transitions based on the ideas of empathy channel and cognitive overlapping.

Three 30-minute TV programs were analyzed to verify the model of shot transition. This result shows that overlapping shot transition occupied 77.9% of the entire transitions. 61.6% of the rest are transitions from person shots to object/person shots in which the empathy channel is considered to be established. 72.7% of them contain attention behaviors, such as a pointing gesture or a gaze toward the object(s) in the next shot.

The GECA (Generic ECA) platform allows one to build an IPOC-like ECA system on multiple servers connected with each other by a computer network (Figure 3) [Huang 2006]. The blackboard model is employed as the backbone of GECA. A high-level protocol (GECAML) is defined to specify the messages exchanged between ECA components such as sensor inputs from the human users, emotion model, dialogue manager, face and body animation, etc.

A number of ECA systems have been implemented on GECA, including a campus guide agent, a tour guide agent, a quiz agent, and a pedagogical agent for teaching cross cultural communication (Figure 4).

2.2 Communicative Robots as Knowledge Media

Robots with a sufficient communicative ability can be characterized as an embodied knowledge medium, where robots bear a role of mediating knowledge among people by acquiring or presenting knowledge contained in the discourse.

In order to explore the idea of communicative robots as knowledge media, we are prototyping listener and presenter robots, as shown in Figure 5. The listener robot is expected to interact with an instructor to acquire knowledge by video-taping important scenes of her/his activities (e.g., assembling/disassembling a machine). The presenter robot, equipped with a small display, will then interact with a novice to show the appropriate video clip in appropriate situations where this knowledge is considered to be needed during her/his work (e.g., trying to assemble/disassemble a machine).

We observed the conversation in detail to obtain a precise model of the way the instructor explains the procedure of assembling/disassembling to the listener (Figure 6). We found that joint attention between the instructor and the listener was achieved very frequently. In fact, when the instructor attended to an object, the listener also attended to it more than 75% of the time. Moreover,

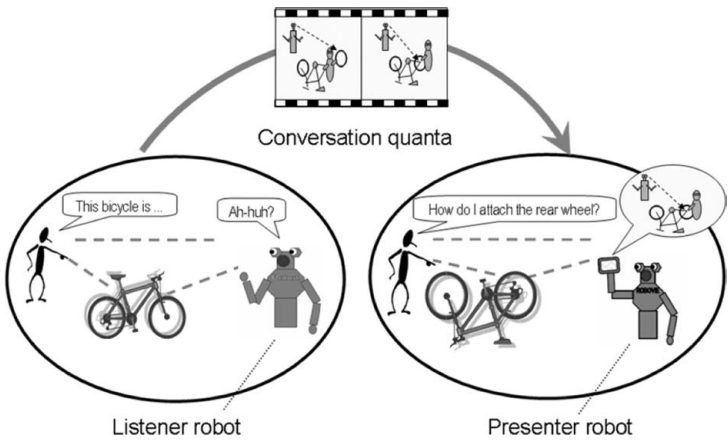


Fig. 5. Listener and presenter robots as embodied knowledge media [Nishida 2006]



Fig. 6. Video data on explanatory scene [Nishida 2006]

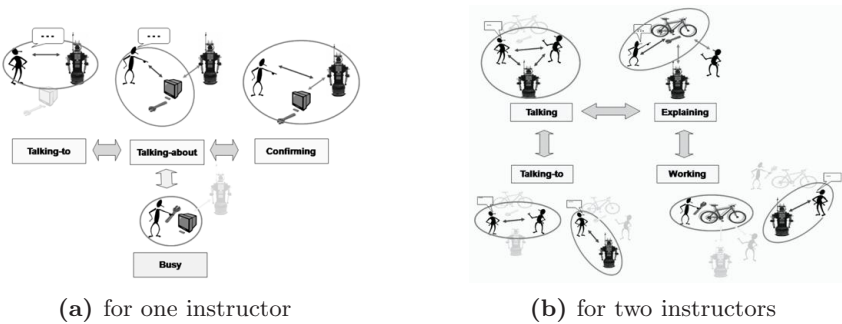
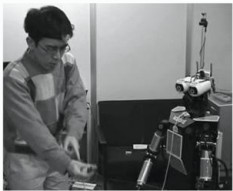


Fig. 7. Communication modes for the listener robot [Hiramatsu 2006]



(a) attention to instructor



(b) joint attention by instructor pointing



(c) confirmation of instructor's intention of drawing attention

Fig. 8. Listener robot interacting with instructor [Nishida 2006]

when the instructor turned his gaze to the listener, he turned his gaze back to the instructor for more than 70% of occasions. In many cases, the listener nodded in concurrence with his gaze, but the frequency differed greatly depending on individuals. It appears quite likely that the instructor's redundant behaviors



(a) starting presentation to listener



(b) confirmation by instructor



(c) attention to pointed object of reference

Fig. 9. Content captured by listener robot and cameras on floor [Nishida 2006]

are intentionally used to establish a joint attention. In addition, it might be reasonable to assume that redundancy of time achieved through repetitive or persistent behaviors are used as a cue to enhance an intention.

We designed the behavior of the listener robot so that it could distinguish four modes of communication, i.e., the talking-to, talking-about, confirming, and busy modes (Figure 7a). In the talking-to mode, the instructor (the human) is mainly watching the listener (the robot) and is involved in the cognitive space based on the relation between them. In the talking-about mode, the instructor is mainly watching the target to be explained, expecting the listener to cognitively share the target. In the confirmation mode, the instructor alternately looks at the listener and the target, suggesting that s/he is interested in whether or not the listener is paying attention to the target. In the busy mode, the instructor is devoting himself to his work without talking to the listener.

Hiramatsu extended the model to cope with the case where two instructors were involved (Figure 7b). The model becomes more complicated to cope with the situations where the robot is watching two instructors are talking with each other. A turn-taking mechanism for multiple-party conversation should be taking into account.

The listener robot was implemented using a humanoid robot called Robovie2, which is almost as tall as a human and can move its hands, head, and eyes. The motions of the instructor and the location of significant objects are captured by a motion capture system. The motion capture system's ten 3D position sensors are attached to the human's body. In addition, several more 3D position sensors are attached to the salient objects in the environment, which will be referred to during the course of explanation.

Figure 8 shows the listener robot interacts with the human user. In Figure 8a shows the initial stage where the listener robot looked at the instructor on the face. In Figure 8b, the listener robot made a joint attention according to the instructor's pointing gesture. In Figure 8c, the listener robot confirmed the object. Figure 9 shows key frames of the video clips captured during the interaction. Figure 9a is a shot taken from a floor camera. Figure 9b illustrates a shot taken by the robot's eye, showing an object confirmed by the instructor. Figure 9c illustrates another shot taken by the robot's eye, showing an object pointed by the user.



Fig. 10. Content captured by listener robot and cameras on floor [Nishida 2006]

Figure 10 shows how the listener robot interacts with two instructors. In Figure 10a, both the instructor in the left and the robot is watching the way instructor in the right is working with the bicycle. In Figure 10b, the robot is looking at the two instructors' conversation. In Figure 10c, the instructor in the left is talking to the robot and the instructor in the right is making a joint attention to the robot.

A presenter robot is designed to efficiently convey knowledge to humans using an installed display integrated into its hand that can be adaptively controlled according to the position of the users [Ohya 2007]. In order for the presenter robot to be able to effectively present its knowledge to the user, it should be able to approach the task area, adjust the display for convenient viewing, and play video to satisfy user's needs.

As for approaching the task area, it is important for the presenter robot to move to a suitable position depending on the user's current position. If the position is too close, the robot may interfere with the user's operation, and if it is too far, the display on the robot's arm may not be visible to the user clearly. Although the robot needs to move whenever the user changes her/his position, the robot's movements should not be too sensitive to the user's position so that the user can view the display in a stable fashion.

As for adjusting the display, the robot needs to control its arm in response to the user's state, posture, and line of sight. If the task area or task operation changes, the user's posture and line of sight may also change. Again, the robot should not be too sensitive to provide the user with a stable view.

As for playing videos to satisfy user's needs, the robot has to understand the user's intention. It is considered to be effective to automatically select videos in response to the positional relation between the user and target objects such as the parts of a bicycle. Moreover, since different users may want to see different sections of the videos showing a similar task operation, it is necessary for the robot to have the ability to identify through interaction the section the current user is interested to see.

Figure 11 shows the architecture of the presenter robot. In order for the presenter robot to work in a noisy environment, a Bayesian net model is used to integrate the intentions of the user and the information collected from different sensors.

Figure 12 illustrates how the user interacts with the presenter robot. Figure 12a shows that the presenter robot makes eye contact in order to coordinate

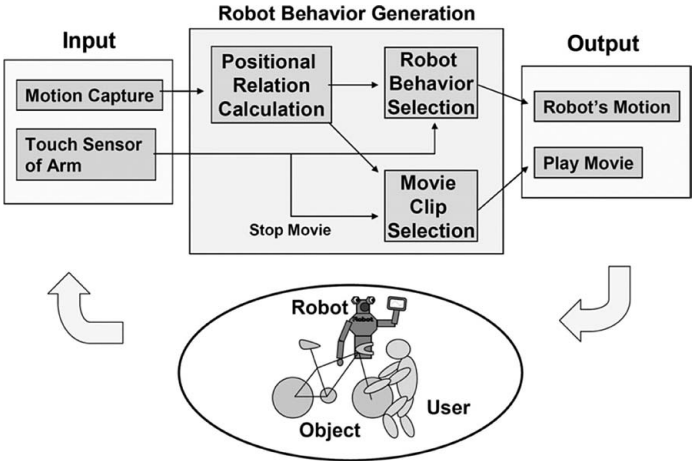


Fig. 11. The architecture of the presenter robot [Ohya 2007]

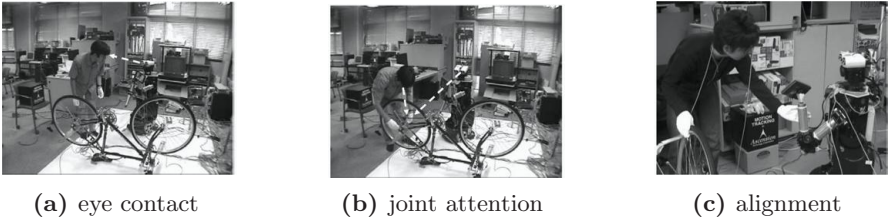


Fig. 12. Interacting with presenter robot [Ohya 2007]

communication. Figure 12b shows that the presenter robot establishes a joint attention during interaction. Figure 12c shows that the user uses a touch sensor to align the position of the display. The experimental evaluation has revealed that the efficiency of the interaction with the user has been much improved in the sense that the time for completing the task with the presenter robot was reduced to about 63% of the baseline setting with the fixed display position, with less (about 50%) interaction time and less frequency (54%) of interaction.

3 Conversational Content

Conversational content represents information and knowledge arising in a conversational situation so that one may reuse it in other situations. We study methods of capturing, accumulating, transforming, and applying conversational content.

We introduce conversation quantization [Nishida 2005] as a conceptual framework for manipulating conversational content. The key idea of conversation

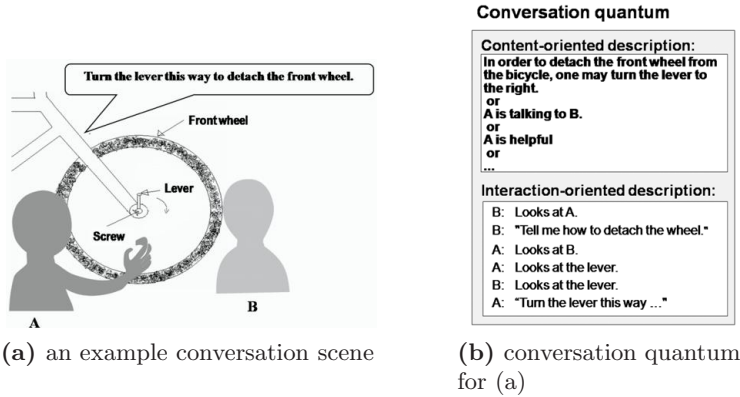


Fig. 13. An example of a conversation quantum [Nishida 2006]

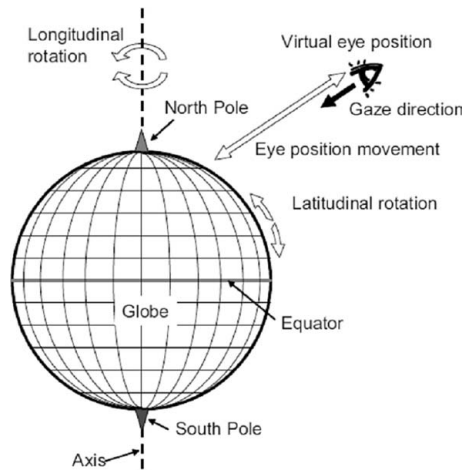


Fig. 14. SKG allows for visual accumulation of conversation contents on the sphere surface

quantization is to use a conversation quantum that encapsulates interaction-oriented and content-oriented views of a conversation scene. The interaction-oriented view is concerned with the way participants interact with each other. On the other hand, the content-oriented view gives a propositional description of the conversation scene.

For example, consider a conversational situation illustrated in Figure 13, where one person on the left is explaining to another person on the right how to detach the front wheel from a bicycle. Figure 13 illustrates a conversation quantum for this conversation scene. The content-oriented description consists of propositions such as “in order to detach the front wheel from the bicycle, one may turn the lever to the right,” or “A is talking to B.” The interaction-oriented

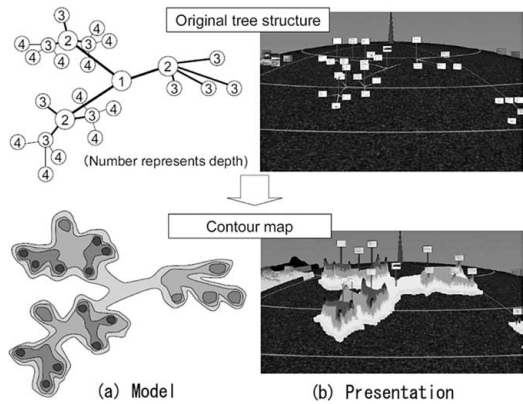


Fig. 15. Nested contour representation for a tree structured data [Kubota 2005b]

description specifies the interactions as a sequence of events such as “B looks at A”, “B asks A to inform how to describe the wheel”, and so on.

The framework of the conversation quantization consists of a spiral of extracting conversation quanta from conversation, accumulating them into a database, and applying them to other conversational situations on demand. Embodied conversational agents or conversational robots may be employed to replay the conversational behaviors on behalf of an actor recorded in a conversation quantum.

In addition to basic operations on conversation quanta, complex operations will produce more content-oriented services such as summarization, complex story building, or spatial arrangement of conversation quanta.

The implementation of the conversation quantization theory depends on the data structure for representing conversation quanta. One may use annotated video clips as representation in order to implement the idea of conversation quantization with minimal technical complexity. Alternatively, one might be interested in using a deep semantic representation such as logical formulas or case frames to increase flexibility of combining conversation quanta.

We have developed a system called the Sustainable Knowledge Globe (SKG) [Kubota 2005b] that permits the user to manipulate a large amount of conversation content on the sphere surface. SKG allows the user to group conversational content into a tree structure so that the user can handle a subtree of content at once (Figure 14).

Conversation contents may also be connected together as a story so that they may be followed sequentially. A graphical user interface is employed to continuously zoom in/out the any region of the sphere surface. A linear zooming method is employed to avoid distortion of the landscape on the sphere surface. In order to help the user visually recognize the tree structure, nesting contours are introduced as shown in Figure 15. In addition, an ECA subsystem is incorporated into the SKG system so that an ECA can navigate the user through the landscape.

4 Conversational Environment Design

Conversational Environment Design is concerned with design and implementation of an intelligent environment for recording and mediating conversations. Currently, we focus on conversation capture, with particular emphasis on capturing conversation quanta from conversation.

In order to understand the process of conversation quantization, we carried out a preliminary experiment on manual materialization and dematerialization [Saito 2005]. We investigated a sample conversation records for six hours obtained from five meetings by two participants held in difference places and weeks, trying to identify minimal segment of conversation that were considered to self-contained in the given discourse. As a result, we identified 41 conversation quanta for one member and 66 conversation quanta for another. The average time length of each quantum for single speech is 42 seconds, while divergence is relatively large. In order to see how much the captured conversation quanta were applicable to a new conversational situation, we made a wizard-of-oz experiment of simulating a conversational system using the obtained conversation quanta. As a result, we found that we can reuse conversation quanta in new conversational situations even though that they were obtained in different situations. In order to understand a conversation quantum, one must share the background knowledge. It is quite likely that the background knowledge should be provided for the user unless s/he shares the context.

We have implemented a small conversational environment called SKG Room for assisting meeting with a human-assisted conversation quanta acquisition facility using auxiliary button devices, as shown in Figure 16a. The architecture of SKG Room is shown in Figure 16b. Each conversation participant is expected to press her/his button when s/he expects or finds a segment of conversation useful and hence intends to record the segment [Saito 2005]. The conversational situation itself is kept recorded during the conversation. Each participant can press the button to indicate the system her/his intention to mark. S/he can specify the in-points and out-points by button-press actions.

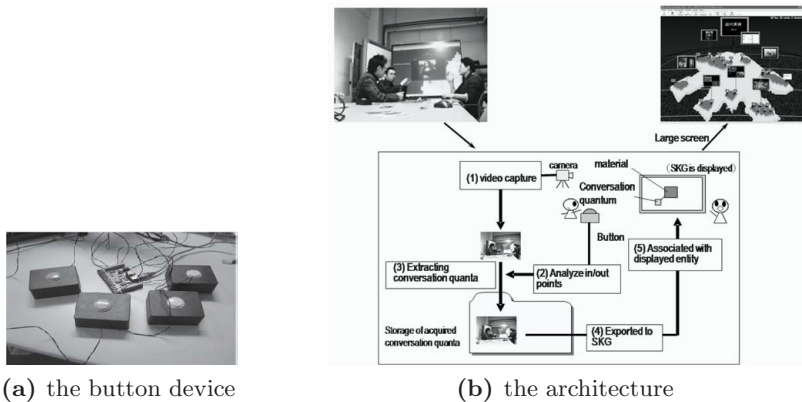


Fig. 16. ASKG Room [Saito 2005]

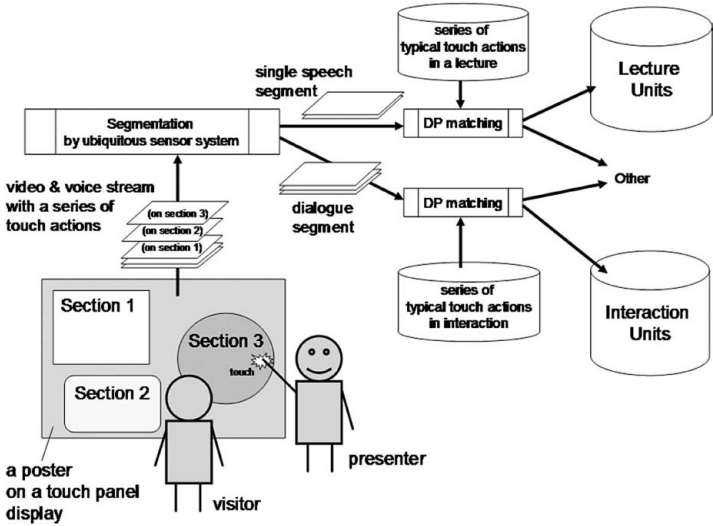


Fig. 17. Conversation quantizing system in a real world poster session [Kubota 2005a]

Ubiquitous sensor room proposed by Sumi et al [Sumi 2005] can be used to further automate the acquisition. It allows conversational events to be captured by environment sensors (such as video cameras, trackers and microphones ubiquitously set up around the room) and wearable sensors (such as video cameras, trackers, microphones, and physiological sensors). In order to supplement the limited capability of sensors, LED tags (ID tags with an infrared LED) and IR tracker (infrared signal tracking devices) are used to annotate the audio/video data with the positional information. Currently, five event types are recognized, including “stay,” “coexist,” “gaze,” “attention,” and “facing”. Events can be recognized by analyzing the behavior of IR trackers and LED tags in the room.

The ubiquitous sensor room was extended to capture conversation quanta. Touch actions on a poster panel were used to estimate conversational segments and their topics (Figure 17) [Kubota 2005a]. Our touch panel presentation system can infer reusable segments by capturing a series of touch actions of the presenter and matching it with typical series of touch actions. The presenter divides the touch panel area into semantic sections of the poster and plays her/his typical presentation in advance. A series of touch actions on the sections are captured while the presenter is explaining the poster to the visitor. The segment is inferred as a “Lecture Unit” type when its series of touch actions are matched well with that of the typical presentation. The system can acquire the conversation topics from the text on the touched poster area. Such information might be helpful for retrieving conversation quanta.

5 Conversation Measurement, Analysis, and Modeling

As is evident from the previous sections, conversation measurement, analysis and modeling constitute a foundation of Conversational Informatics. By exploiting advanced sensor technologies, we are now able to measure various aspects of conversational behaviors, particularly nonverbal behaviors, in greater detail. The insights gained from high-precision quantitative measurement turn out to be very useful in order to better design conversational artifacts and conversational environments. In the rest of this section, I will first overview a project aimed at realizing a communicative robot that permits mutual adaptation with the human user. Then, I will describe the IMADE room project aimed at building a conversation measuring environment that integrates multi-modal, biological, and brain sensing methods.

5.1 A Bootstrapping Approach to Mutual Adaptation

We proposed a bootstrapping approach to mutual adaptation consisting of three stages: a human-human WOZ experiment, a human-robot WOZ experiment, and a human-adaptive robot experiment [Xu 2007a], [Xu 2007b]. Instead of directly diving into the third stage, we observe in detail how people adapt with each other and how people improve the protocols for interacting with robots.

(1) The Human-Human WOZ Experiment. At this stage, we created a setting for a team consisting of two persons in which one member of the team (“the instructor”) instructs another member (“the actor”) to move along a pre-defined path only using gestures. The virtual world consists of the field with obstacles (“bombs”), traffic signals, and targets, as well as the goal and orbit. The traffic signal was used to observe the changing rhythms of the instructor’s gestures and the actor’s movements. The reward would be given to the team if the actor cleared target objects or reached the goal, while penalty would be imposed if s/he exploded bombs or neglected traffic signals.

We considered the following three requirements need to be satisfied. First, it is necessary to limit the available channels between interaction partners to those that are observable, in order to eliminate interference from other communication channels. Second, it is necessary to induce mutual adaptive behaviors between human interaction partners, in order to clarify the mechanism of human-human mutual adaptation. Third, it is necessary to record synchronized data obtained from both the participants (the interaction partners), in order to obtain time series data.

Figure 18 shows the experimental environment we developed. The map indicating the route, targets, traffic signals, and bombs was shown only to the instructor using the GUI monitor in front of the instructor. The position of the actor was sensed by floor pressure sensor and shown on the map. In order to help the instructor observe the actor’s movements more directly, we installed a USB web camera and the scene taken by it was also displayed on the screen with the

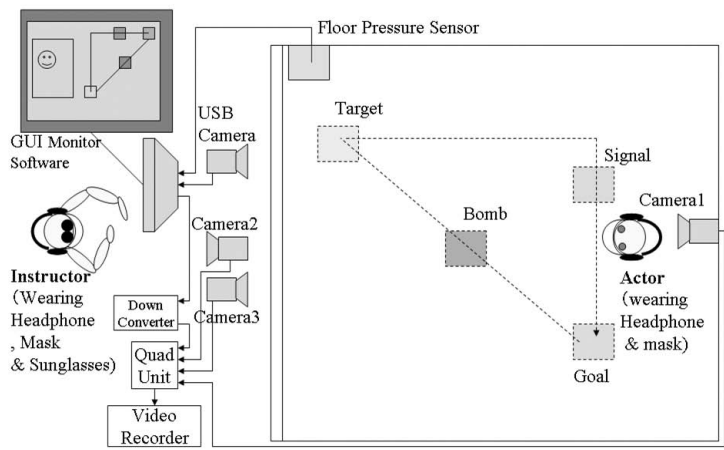


Fig. 18. configuration of human-human WOZ experiment [Xu 2007a]

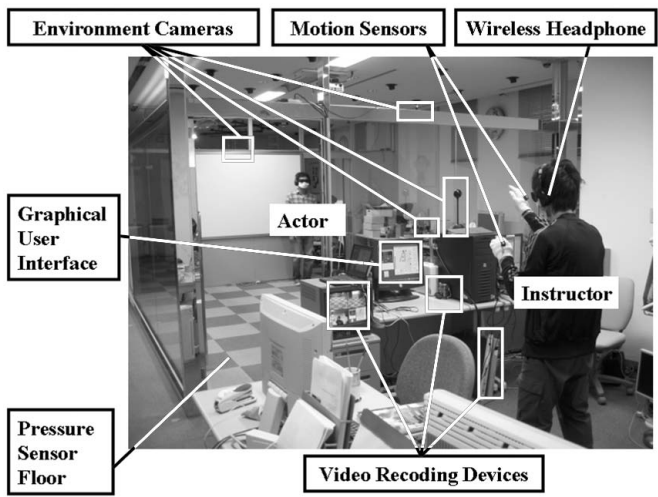


Fig. 19. The environment of human-human WOZ experiment [Xu 2007a]

map and actor’s position overlaid. By comparing the image from the USB camera with the real-time-changed floor sensor information that reflected the actor’s current footsteps, the instructor was able to judge whether the actor was moving along the correct orbit, stepping on a bomb, or approaching the target correctly (Figure 19). The mask and sunglasses were used to prevent interference from other communication channels, such as facial expressions and gaze directions.

By analyzing the experimental results, we obtained three types of mutual adaptive behavior. The first was what we call an alignment-based action in

which two participants align their actions while interacting with each other. For example, this implies that the actor changes his speed and/or steps in order to keep pace with the speed and/or width of the instructor's gesture, and vice versa. Based on the alignment action, two types of behaviors - pace keeping and timing matching - were observed. The second was a symbol-emergent learning in which the instructor used gestures such as symbol-like instructions when he interacted with the actor. The observed symbol-like instructions included the following gestures: "stop", "a bit", and "keep going." The third was an environment learning behavior in which communication between the participants became smoother and more efficient as time elapsed. We also found that gestures of the instructor decreased in the later stages of an experiment session. We suspect that the actor became familiar with the environment when s/he repeatedly moved along the same route, and became more confident in the subsequent rounds.

(2) The Human-Robot WOZ Experiment. At this stage, we considered that the following three conditions need to be satisfied. First, both the instructor’s intentions expressed as gestures and the operator’s intentions expressed as robot’s movements should be observable. Second, some correlations among the data of the instructor, the operator and the robot should be discovered through analyzing the recorded data. Third, the instructor’s gestures, the operator’s feedbacks and the robot’s movements should be recorded synchronously.

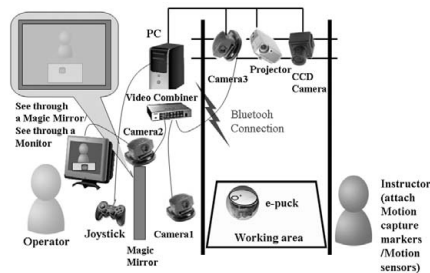


Fig. 20. Configuration of human-robot WOZ experiment [Xu 2007b]

We prototyped the experimentation setting as shown in Figure 20, and conducted some initial experiments and achieved some primary results. The effectiveness of the WOZ experimental environment can be partly proved since several adaptive behaviors of both the instructor and the operator were observed during the procedure of human-robot interaction.

(3) The Human-Adaptive Robot Experiment. At this stage, we developed a preliminary experiment environment to investigate a one-way adaptive human-robot interface, as a first step.

The full configuration of the experimental environment is illustrated in Figure 21. We used an LCD projector fixed on the ceiling to display the maps

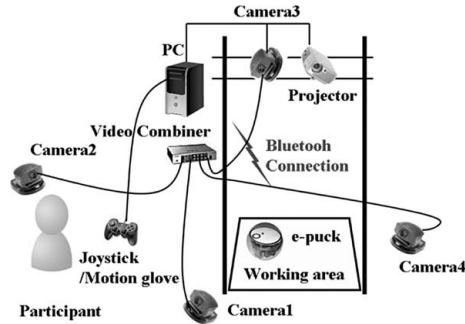


Fig. 21. The configuration of gesture-based human-robot interface experiment [Xu 2007b]

on the floor. Four cameras are used to record the instructor’s gestures/joystick operations and the movement of the robot from four vision angles. The robot controlled through the interfaces is a miniature wireless robot (E-puck).

Two types of interfaces were implemented and compared. The joystick interface consists of a joystick. Since only five types of static gestures are detected by the motion glove, we also limited the joystick to 5 functions: forward and backward respectively mean “go forward” and “move backward”; left and right respectively mean “rotate left” and “rotate right”; finally, the default center position means “stop.” The other interface was based on a motion glove with a 3D motion sensor. It was able to distinguish between five types of gestures: “horizontal flat hand,” “vertical backward flat hand,” “vertical leftward flat hand,” “vertical rightward flat hand” and “forward open palm”. A command was associated with each gesture.

Two types of adaptive behaviors were observed. The first was a shortening behavior in which some redundant behaviors (such as an intermittent stop) were removed as time elapsed in the experiment session. The second was called a replacement behavior in which an action was replaced by another (e.g., “turn right for 90 degree” was replaced by “turn left for 270 degree”). It seemed that the replacement behavior took place because the participant wanted to find a protocol more comfortable to her/him. However, as s/he found the replacement took more time and was not efficient enough, s/he restored the protocol to the previous one.

5.2 IMADE Room Project for Integrated Conversation Measurement

In order to measure conversational situations, we are developing a conversational environment called IMADE (the real world Interaction Measurement, Analysis and Design Environment). We are aiming not only to record conversational situations using ubiquitous cameras/microphones, but also to measure the activities of each participant and their mutual interrelations.



Fig. 22. IMADE (the real world Interaction, Measurement, Analysis and Design Environment) room

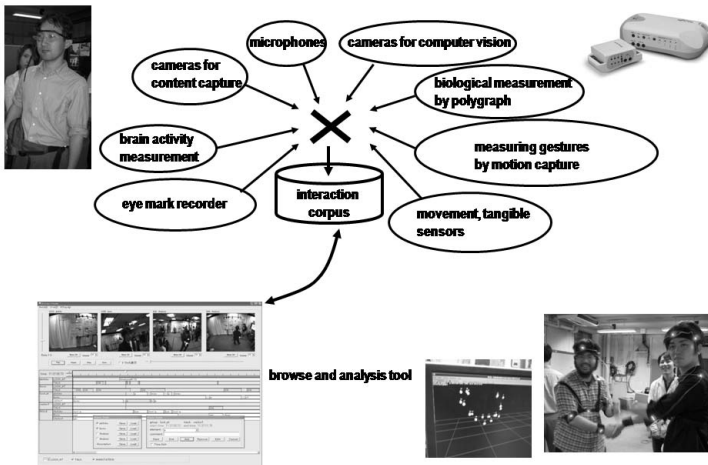


Fig. 23. Facilities to be installed in the IMADE room

So far, we placed much emphasis on multi-modal measurement of conversational situations, as shown in Figure 22 [Bono 2007]. We used 8 wired/wireless microphones to capture audio data, which were aligned in real time recordings. The captured data not only were simultaneously recorded by computer and divided up in a video capture system. We installed 8 cameras on the ceiling of the

IMADE room. The captured data were gathered in the video capture system we developed. We can observe multiple video data that are time aligned. The format is MPEG2/PS at 4 MGbps/sec. We can not only adjust the camera position and angle as we like but also settings stored in the system. We incorporated an optical motion capture system to capture the body direction, head direction, and standing or sitting position by using this system. We chose “MobileEye” as the eye mark recorder.

We plan to introduce biological and brain measurement devices so that we can observe the internal activities and their interdependencies of each participant in a given conversational situation, as shown in Figure 23.

6 Conversational Informatics at the Intersection of Web Intelligence and Brain Informatics

Conversation has been and is attracting significant interest in academic research ranging from communication science to information engineering applications. Conversational Informatics significantly extends the previous work by introducing a measurement-based approach to nonverbal communications and extending the scope of applications to knowledge management in general. The former has created a new frontier that solicits an intensive joint research between scientists and engineers. The latter has created numerous demands for a novel technology.

Web Intelligence encompasses endeavor to build networked intelligence by connecting distributed wisdom and data. Conversational Informatics can interface Web Intelligence with the human society by providing people with a natural means for deriving benefits from Web Intelligence. By closely integrating Web

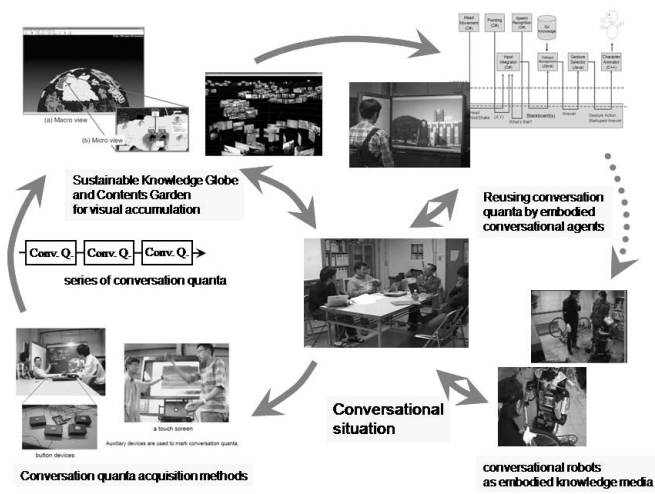


Fig. 24. A conceptual framework for conversational knowledge circulation [Nishida 2007]

Intelligence and Conversational Informatics, we will be able to realize a comprehensive technology for empowering communities by knowledge circulation. It not only enables knowledge sharing and formation of the community-maintained artifacts of lasting value, but also it will increase the value of knowledge by inducing an evolutionary process.

By integrating human-assisted acquisition or conversational contents, visual accumulation of conversational contents, and content-driven embodied conversational agents, we will be able to construct a conceptual framework for conversational knowledge circulation that will accelerate knowledge circulation based on conversational communication, as shown in Figure 24. Conversation quantization serves as a basic framework. Technologies from Web Intelligence will be used to navigate the user in the knowledge space and derive knowledge from data.

The major advantages of this framework is situatedness of information and knowledge to be circulated in a community, for the conversational communication style greatly encourages participants to communicate situated information and knowledge in the sense that the content of information and knowledge need to be grounded to the shared context for conversation in order for conversation to be successful. By providing a theory and mechanism for capturing and producing situated information and knowledge, Conversational Informatics will be able to make a contribution to Web Intelligence in remarkably improving the quality of information and knowledge to be communicated.

In the meanwhile, Brain Informatics is concerned with both brain mechanism and functions. Communication, conversation in particular, is among the most important brain functions. Uncovering the brain function underlying conversational behaviors is not only a challenging subject of Brain Informatics but also a great benefit to Conversational Informatics in the sense that it will permit us to obtain a more detailed account of the effect of conversational activities on the brain.

By integrating audio-visual measurement that characterizes what can be perceived by conversation participants with what is observed inside the brain and body, we will be able to not only build a more comprehensive theory of conversation grounded on the brain and biological activities but also invent more sophisticated artifacts that might be able to touch our heart.

One of the challenging issues at the intersection of Conversational Informatics and Brain Informatics is intention. In Brain Informatics, intention has been discussed in the context of consciousness [Becchio 2006]. In Conversational Informatics, classic models of intention have been highly symbolic and one-directional, and hence they are not considered to be adequate for accounting for conversational behaviors, nonverbal behaviors in particular. We need to model intentions from the signal level.

Intention formation and maintenance was a key issue in coordinating behaviors between human and robots [Nishida 2006]. Rather than formulating intention as something transferred from one agent to another, we employed the following hypothesis.

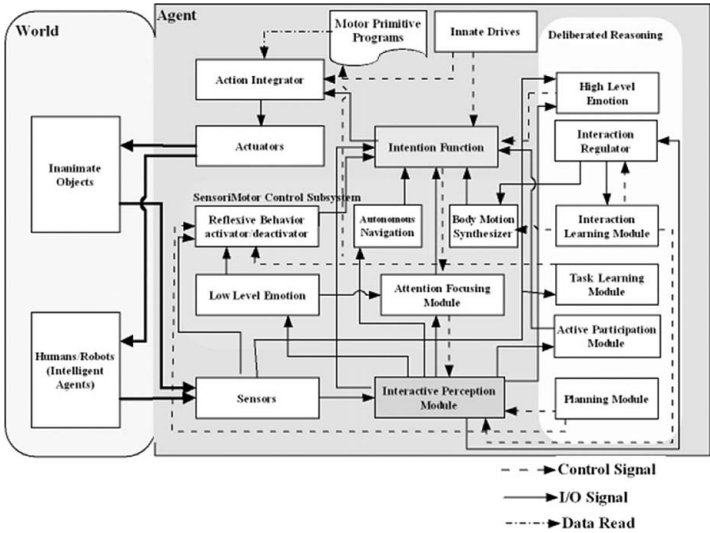


Fig. 25. Functional view of EICA architecture [Mohammad 2007b]

Intention through Interaction Hypothesis. Intention can be best modeled not as a fixed unknown value, but as a dynamically evolving function. Interaction between two agents couples their intention functions creating a single system that co-evolves as the interaction goes toward a mutual intention state.

Mohammad and Nishida introduced the notion of interactive perception, which is defined as an interaction schema allowing the perception module to intentionally affect the environment in order to align its actions with the feedback of the human, which should also be aligned with the system’s feedback [Mohammad 2007a]. The idea was implemented as a drawing program called NaturalDraw. The idea was further extended as the EICA architecture based on the following hypothesis:

Historical Embodiment Hypothesis. The precondition level of embodiment for achieving intelligent autonomous real world agents is the historical embodiment level as defined in [Ziemke 1999]. This level of embodiment emphasizes the role of the extended interaction between the agent and the environment.

It is a rather comprehensive covering from the low level emotion as an input summarizer to deliberate planning and intention function. By further integrating the insights from Brain Informatics, we envision a more robust and biologically grounded view of modeling intention.

In sum, we believe Conversational Informatics extends the intersection between Web Intelligence and Brain Informatics [Zhong 2006] from a communicative perspective, as shown in Figure 26.

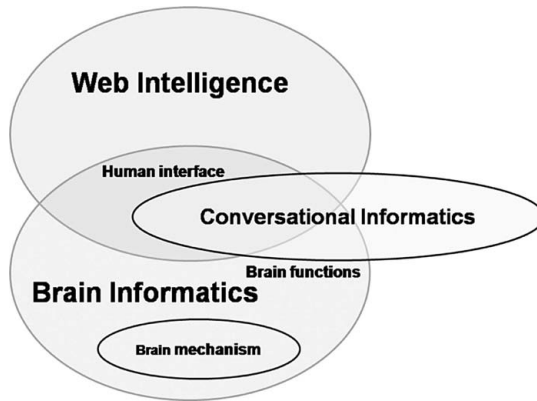


Fig. 26. A schematic view of Web Intelligence, Brain Intelligence, and Conversational Informatics

Application of Conversational Informatics at the intersection of Web Intelligence and Brain Informatics would be enormous. Conversational knowledge circulation in Figure 24 might be further extended so that it can be grounded on perceptual and cognitive activities on the brain. Such a framework will be extremely useful as a large-scale knowledge management system in the industrial and business world [Merckel 2007].

References

- [Becchio 2006] Becchio, C., Adenzato, M., Bara, B.G.: How the brain understands intention: Different neural circuits identify the componential features of motor and prior intentions. *Consciousness and Cognition* 15(1), 64–74 (2006)
- [Bono 2007] Bono, M., Sumi, Y., Nishida, T.: An experiment study of gesture-based human-robot interface. In: *International Conference on Complex Medical Engineering*, Beijing, China, pp. 475–479 (2007)
- [Hiramatsu 2006] Hiramatsu, T.: Listener robot that can acquire video contents from real world conversations, undergraduate thesis, Department of Computer Science, Faculty of Engineering, Kyoto University (in Japanese) (unpublished)
- [Huang 2006] Huang, H., Masuda, T., Cerekovic, A., Tarasenko, K., Pandzic, I., Nakano, Y., Nishida, T.: Toward a Universal Platform for Integrating Embodied Conversational Agent Components. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) *KES 2006. LNCS (LNAI)*, vol. 4252, pp. 220–226. Springer, Heidelberg (2006)
- [Kubota 2002] Kubota, H., Nishida, T.: EgoChat Agent: A Talking Virtualized Member for Supporting Community Knowledge Creation. In: Dautenhahn, K., Bond, A., Canamero, D., Edmonds, B. (eds.) *Socially Intelligent Agents - creating relationships with computers and robots*, IOS Press, Amsterdam (2002)

- [Kubota 2005a] Kubota, H., Takahashi, M., Satoh, K., Kawaguchi, Y., Nomura, S., Sumi, S., Nishida, T.: Conversation Quantization for Informal Information Circulation in a Community. In: The Fourth International Workshop on Social Intelligence Design (SID 2005) (2005)
- [Kubota 2005b] Kubota, H., Nomura, S., Sumi, Y., Nishida, T.: Sustainable Memory System Using Global and Conical Spaces. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 861–867. Springer, Heidelberg (2005)
- [Merckel 2007] Merckel, L., Nishida, T.: Solution of the Perspective-Three-Point Problem Calculation from Video Image by using Inclinometers. In: IEA/AIE 2007, Kyoto, Japan, pp. 324–333 (2007)
- [Mohammad 2007a] Mohammad, Y.F.O., Nishida, T.: NaturalDraw: Interactive Perception Based Drawing for Everyone. In: Proc. 2007 International Conference on Intelligent User Interfaces (IUI 2007), pp. 251–260 (2007)
- [Mohammad 2007b] Mohammad, Y.F.O., Nishida, T.: Intention through Interaction: Toward Mutual Intention in Real World Interactions. In: IEA/AIE 2007, Kyoto, Japan, pp. 115–125 (2007)
- [Nakano 2004] Nakano, Y.I., Murayama, T., Nishida, T.: Multimodal Story-based Communication: Integrating a Movie and a Conversational Agent. IEICE Trans. Information and Systems E87-D(6), 1338–1346 (2004)
- [Nishida 2005] Nishida, T.: Conversation Quantization for Conversational Knowledge Process, Special Invited Talk. In: Bhalla, S. (ed.) DNIS 2005. LNCS, vol. 3433, pp. 15–33. Springer, Heidelberg (2005)
- [Nishida 2006] Nishida, T., Terada, T., Tajima, T., Hatakeyama, M., Ogasawara, Y., Sumi, Y., Xu, Y., Mohammad, Y.F.O., Tarasenko, K., Ohya, T., Hiramatsu, T.: Towards Robots as an Embodied Knowledge Medium, Invited Paper, Special Section on Human Communication II. IEICE Trans. Information and Systems E89-D(6), 1768–1780 (2006)
- [Nishida 2007] Nishida, T.: Social Intelligence Design and Human Computing. In: Huang, T.S., et al. (eds.) Human Computing. LNCS (LNAI), vol. 4451, pp. 190–214 (2007)
- [Ohya 2007] Ohya, T.: Towards Robot as an Embodied Knowledge Medium – Development of a presenter robot that can efficiently convey knowledge to humans, master thesis, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University (in Japanese) (unpublished)
- [Okamoto 2005] Okamoto, M., Nakano, Y.I., Okamoto, K., Matsumura, K., Nishida, T.: Producing effective shot transitions in CG contents based on a cognitive model of User Involvement, IEICE Transactions of Information and Systems Special Issue of Life-like Agent and Its Communication. IEICE Trans. Information and Systems E88-D(11), 2532–2623 (2005)
- [Saito 2005] Saito, K., Kubota, H., Sumi, Y., Nishida, T.: Analysis of Conversation Quanta for Conversational Knowledge Circulation. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 296–302. Springer, Heidelberg (2005)

- [Sumi 2005] Sumi, Y., Mase, K., Mueller, C., Iwasawa, S., Ito, S., Takahashi, M., Kumagai, K., Otaka, Y., Tsuchikawa, M., Katagiri, Y., Nishida, T.: Collage of Video and Sound for Raising the Awareness of Situated Conversations. In: Bolc, L., Michalewicz, Z., Nishida, T. (eds.) IMTCI 2004. LNCS (LNAI), vol. 3490, pp. 185–194. Springer, Heidelberg (2005)
- [Xu 2007a] Xu, Y., Ueda, K., Komatsu, T., Okadome, T., Hattori, T., Sumi, Y., Nishida, T.: WOZ: experiments for understanding mutual adaptation, AI & Society (to appear)
- [Xu 2007b] Xu, Y., Guillemot, M., Nishida, T.: An experiment study of gesture-based human-robot interface. In: International Conference on Complex Medical Engineering, Beijing, China, pp. 458–464 (2007)
- [Ziemke 1999] Ziemke, T.: Does Representation need Reality. In: Rethinking Grounding, pp. 177–190. Plenum Academic / Kluwer Publishers
- [Zhong 2006] Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S.: WI meets BI: WICI Perspectives. In: The WICI International Workshop on Web Intelligence (WI) meets Brain Informatics (BI) (WImBI 2006), December 15–16, 2006 Beijing, China (2006)
- [Zhong 2007] Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies (iIT) from the Stand-Point of Web Intelligence (WI). Communications of the ACM 50(3), 89–94 (2007)

Intelligence for Upgrading Information

Setsuo Ohsuga

Professor Emeritus, University of Tokyo
ohsuga@s01.itscom.net

Abstract. Intelligence concerns many aspects of human mental activity and is considered difficult to be defined clearly. Apart from its relation to mental activity, however, it is possible to discuss intelligence formally based on information it deals with. This paper defines intelligence as the capability to upgrading information and notes that there have been four important phases in the progress of intelligence. These are: (1) language acquisition, (2) knowledge discovery, (3) conceptualization and (4) granulation. These phases are discussed in this paper.

1 Introduction

Today works that require high level intelligence are rapidly increasing such as developing software, robotics, intelligent Web and so on. Formalization and computerization of intelligence is required strongly. But it is considered difficult because intelligence is believed to concern many aspects of human mental activity. Apart from its relation to mental activity, however, it is possible to discuss intelligence formally based on information it deals with. The objective of this paper is to define intelligence purely on the basis of information.

There are number of words relating to intelligence such as intelligent, clever, wise, knowledge and so on. The meanings of these words other than knowledge are not yet made clear. Among them ‘clever’ and ‘wise’ represents different modes in revealing ‘intelligence’. These are of the same category. On the other hand ‘knowledge’ is clearly of the different concept from the former three. The word ‘knowledge’ represents something described in a language on the recognized object or the relation between the different objects. Today knowledge has been well formalized and is used widely for achieving the goal of subject’s activity. If intelligence is defined based on the concept of knowledge, then its formalization and computerization can also be achieved.

A word ‘knowledge’ extends widely to mean from very simple and naive knowledge to very high level knowledge. The latter has a complex meaning and can help subject for achieving very complex tasks. The higher is the level of knowledge however, the more difficult it becomes to be understood.

The nature of subject’s activities is different by each individual because of the level of knowledge they have. Here by the subject we mean to include ordinary creatures, computers as well as human being. The difference of intelligence comes from the quality of knowledge the subject has. A person as highly intelligent subject has

capability to understand and to use high level knowledge. High level knowledge is built up from a set of low level knowledge. That is, to understand high level knowledge is to understand the meaning of low level knowledge and also a manner of new concept being created by combining the low level knowledge. Furthermore, highly intelligent persons also have capability to create themselves not-yet-defined concept with a more complex form and meaning based on existing knowledge. These two aspects are very important. However the capability to create new concept implies to understand the meaning. Therefore in the sequel intelligence is simply defined as a capability to upgrade the information.

Formally this is to derive either of wider expression, deeper expression or higher level expression. The wider expression is such a knowledge expression that covers plural of other expressions. The deeper knowledge is an expression of such that includes the deeper meaning than any existing knowledge. Since, in many cases, wider expressions and deeper expressions are closely related to each other, these are called simply knowledge expression with larger meaning. Higher level expressions of knowledge are those that can represent a high level concept that has not been expressed before. These are closely related to a method of representation, i.e., language in the wide sense.

The upgrading of expressions relates both to upgrading the representation method or the rule of expressions (e.g. language grammar) and upgrading the way of making expression in this rule (e.g. way of making sentence using language). Today the rule of symbolic language has been established. Therefore the difference of intelligence appears exclusively in the way of making expressions. But if we have freedom to define a new expression rules too, then the difference in the level of expressions may extend much wider. In fact, in the history of development of information of human being, there was such a case that new rule have been made and it enabled ones to upgrade information greatly. A typical example was acquirement of symbolic language. In an early time human being might have only the biological information processing system like neural network similar to the other creatures and be almost in the similar intelligence level as them. But human being acquired language sometime in the history and became possible to represent explicitly various objects and their properties and mutual relations surrounding them. Acquirement of language was the greatest upgrading of information. How could human being acquire symbolic language? Let it be a first phase of intelligence. Though this was an event that arose in very old time, it relates also to very modern problems such as giving a robotic system the capability to adapt environment.

After symbolic language has been acquired, knowledge expression may have been made in this language. The characteristic of symbolic language is not merely in its expression rules but also in its inference capability that derives different expressions from the given expressions. It enables to expand the scope of expressions from the given small set of expressions to a new larger set that are implied by the given ones. Whether one can use inference capability correctly or not relates to the scope of activities based on information.

Today most persons can use inference through the use of language and perform a large scope of activities. Since the inference is itself a formal procedure and is realized by mechanical operations, mere to use the inference is not an intelligent activity. The scope of the extended expressions by inference is decided depending on the prior expressions. If, on the other hand, new knowledge that has not existed before is created and added to the old set, the scope expands considerably by the inference. To create new knowledge expressions relates intelligence. Creation of new knowledge, especially getting new language expressions from a set of observed data, is referred to the second aspect of intelligence in this paper. Attention is paid today to data-mining and knowledge discovery as the new knowledge generation methods. These are the activities to get (symbolic) language expression to represent observed (non-symbolic) data on the basis of given language rule.

A new knowledge expression has been made anytime it was possible. But there has been no idea on future upgrading of the knowledge expression. Later on some upgraded knowledge expressions have been discovered based on the existing knowledge. It enabled to enlarge further human activities

Upgrading knowledge is to create new conceptual knowledge based on the existing knowledge expressions. Among them knowledge on the way of using existing knowledge for achieving the intended goal is included. It is knowledge on knowledge. This knowledge needs meta-expression. Rule of expression in a language has to be expanded in order for representing it.

The process of making higher level expressions from the lower level expressions has been repeated from very long time before until today and a rich set of conceptual expressions has been formed. This process of making higher-level knowledge expression needs intelligence. There is no systematic guide line for enabling it. This is an important step of intelligence and is classified in this paper as the third aspect of intelligence.

These upgrading of information mentioned as above are made in a bottom-up manner. The upgraded information was useful for human being for managing complex activities in a more secure manner than ever. However it does not mean that the upgraded information is always satisfactory for achieving a goal or, in other word, for solving problem. To make decision in an optimal way using information is rephrased to solve problem. It is to find and represent explicitly one or more unknown route to a given goal starting from a given condition. Existing knowledge is used for the purpose. If the goal, i.e. solution, is near the start point, then problem is solved rather easily. Many problems that have arisen in the past were simple and have been solved without much difficulty by excellent persons using existing knowledge. Upgrading knowledge has been advantageous for the purpose of problem solving. But recently complex problems are increasing. The distance between wanted goal and existing knowledge in these problems becomes greater and greater. Bottom-up upgrading cannot catch up the increase of complexity. Problem solving became more and more difficult. A typical example is software development.

In general a solution of the given complex problem is represented by a structure of knowledge expressions. To look for a solution to the given problem is to look for such a structure of knowledge expressions that meets given condition. It is a kind of activity to upgrade information and needs intelligence. Different from the bottom-up step of the second and third phases this is a top-down activity. Let it be a fourth phase of intelligence.

As the conclusion, the following four phases of intelligence are considered important.

[A] Acquiring symbolic language (language acquisition),

[B] Generation of symbolic knowledge expression from raw data. (knowledge discovery)

[C] Making new concepts and meta-expression (conceptualization)

[D] Building knowledge structure to represent and solve given problem (granulation)

Intelligence is exhibited in any occasion of human activity but above four is selected as the largest upgrading steps among them. These four phases are substantially different to each other and appear mostly in the different times in the history of progress of information. It is illustrated in Figure 1.

Among these four phases each of the last three ([B], [C] and [D]) is a function defined on the basis of knowledge. That is, these are the functions that could be defined after a language has been acquired and more or less have been cognized and studied in the framework of language so far. On the other hand, the function of the first ([A]) must be defined without language. It is not for upgrading information directly like the other functions but is for creating the representation method. But any other upgrading ([B], [C] and [D]) could not be achieved without language and an intelligent function for enabling [A] played the greatest role in upgrading information. How could human being acquire language without using language? This is a very interesting problem. Few researches have been made on this subject [1]. Because of the difference in the nature, the intelligent function for creating knowledge ([B], [C] and [D] above) and that for creating language ([A]) are discussed separately in the sequel.

2 Non-symbolic Information and Symbolic Information

2.1 Information Representation and Information Processing Method

Information is useful when it is processed properly for achieving user's objective. The processing is to transform an input-information to an output that is suited for achieving the objective/ intension of the subject. In order to deal with the wide range of requirements, the processing method must be made as general as possible (not restricted to any specific objective/meaning) so that a representation suited for any specific objective is possible. That is, information representation must be considered together with processing mechanism. Currently a typical pairs of representation form and its processor are,

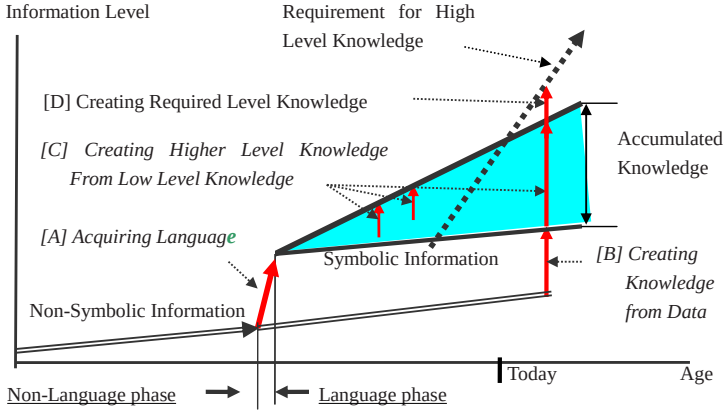


Fig. 1. Four phases of information upgrading

- (A) data that corresponds directly to physical quantity and its processor as its direct conversion, and
 (B) symbolic language and its processor.
 In the following these two are discussed.

2.2 Data and Language

For most of living things information is limited to a simple data. The living things have only such a simple processor as sensor to detect data input and biological network to transform it. The processor became the more complicated through the long history of evolution. It was however rather a gradual evolution without any abrupt change and its original form is kept even today in every creature including human being.

Another representation form is language. It is seen only in human being. In contrast to the fact that any simple data correspond directly to a physical object and its processor is realized by a physical circuit, language is an abstract expression and its processor is defined based on the abstract information. Language acquisition is rather a miraculously occurred event in the human history. In order to understand the process of human being having acquired language, the nature of non-symbolic and symbolic information is studied in more detail in the following.

After a language has been acquired, knowledge is created with it. Knowledge is defined as what is expressed in language on an object. In principle, knowledge is memorized and reused. Accumulation of knowledge enabled human being a variety of activities and urged them to acquire the further knowledge. It accelerated the further variations of activities. Every extension of intelligence after this time, such as creation of new concepts, representing the more complicated meaning of things by making structure of knowledge and so on became possible by the acquirement of language.

The term language refers a class of individual languages of the different types. For example, there are declarative type language like natural language for human being and procedural type language for computers. Substantially every symbolic language is similar in the sense that mutual conversion is possible mostly between every pair of them. Declarative language is considered among them in the following.

2.3 Visible Difference Between Non-symbolic and Symbolic Representation

There is a big difference between data processing (non-symbolic processing in the sequel) and symbolic processing. Non-symbolic processing uses a form of non-symbolic manifestation. It needs special equipment for each object to which a subject pays attention. It measures an interested aspect of the object. Even though numerical symbol is used for explanation it has no feature of symbolic information.

Symbolic expression on the other hand is not connected directly with any object. A world of symbols (language) composed of the combinations of finite number of symbols is made independent of objects. In general it has a layered structure composed of symbols, words as a simple combination of symbols, clauses and sentences composed of words. Object and its symbolic expression are linked via a mapping table (dictionary) indirectly. This linking is not fixed in advance but the table can be replaced by another one with the different object-symbol relations. Then the same language can represent the different world of objects.

A processing method is defined to each form of manifestation/expression. Since the form is made specifically to each object in case of non-symbolic processing, its processing method is also specific to each individual object. In many cases a different facets are measured and the manifestation forms a vector of measurements. Then a general way of non-symbolic processing is to generate a new vector (output) from the original (input) vector by conversion. Living thing performs biologically this operation. In many cases it has learning capability and adapts itself to environment. It is to modify some parameters of vector conversion with a feed-back path provided outside of the basic conversion mechanism. An example of its engineering model is neural-network model. To convert a vector of each measurement is a typical example of non-symbolic processing.

In symbol processing on the other hand, a symbolic world (language) is made in which different terms are made corresponding to objects in an object world composed of entities, their properties and changes, activities of human and the others, etc. This is a typical style of declarative language like natural language. Processing of this language is based on symbolic inference. Change in the object world corresponds to change of expressions in the symbolic world and is deduced by inference. What has been described in this language is knowledge on the object. Not-yet-known property and relation among the objects are derived from known knowledge by means of inference. Knowledge is used selectively at the inference and the different result is obtained depending on the selection.

Declarative language is assured modularity of expression and processing. Everything is represented using symbols of finite length. Because of this nature an explorative goal-search operation is possible. By the modularity of expression and flexibility of processing, declarative processing can accept any time a new statement. When a new entity is added to the object world a term corresponding to it can also be added to the language. Thus world of symbols can expand easily. Let it be called the expandability of declarative language.

There is a big gap between symbolic processing and non-symbolic processing. Because of this gap a smooth transition from non-symbolic information to symbolic information seems very difficult.

2.4 A Framework to Compare Non-symbolic and Symbolic Processing

In order to make this difference clearer, the internal structures of non-symbolic processing and symbolic processing are compared. A common framework to represent both of them is introduced [1,2]. Among possible symbolic languages, a typed logic is used in the sequel. In this logic every variable is assigned explicitly a domain as a set of instances. For example, instead of writing "man is mortal" like $(\forall x)[man(x) \rightarrow mortal(x)]$ as ordinary first order logic, it is written $(\forall x/MAN) mortal(x)$ where MAN is a domain set of the variable x. $(\forall x)$ denotes 'for all'. This representation is true for x in this domain, that is, $x \in MAN$.

The larger the domain is in a rule, the wider class of conclusions can be deduced. In this case the rule has larger scope of applicability. In acquiring, knowledge to generate the one with the larger scope is the more desirable because then the narrower scope knowledge can be deduced from it. Let assume a formula $(\forall x/D) predicate1(x)$ and its domain D be divided into a set $\{D1, D2, \dots, Dn\}$. Then $(\forall x/D) predicate1(x)$ implies every narrower domain formula $(\forall x/D) predicate1(x)$, $(Di \subset D)$, for all i, and can replace all the latter predicates.

Non-symbolic processing is represented as a conversion of input vector to output vector. Let the input and output vectors be represented mathematically as Pf and Pg respectively. Then non-symbolic processing is represented mathematically as $Pg = Pf \times T$ where T is a matrix to represent a conversion network.

On the other hand, predicate logic as symbolic processing represents a static relation of things. But an inference $(\forall x/D)F(X) \& (\forall x/D)[F(x) \rightarrow G(x)] \Rightarrow (\forall x/D)G(x)$ for deducing $(\forall x/D)G(x)$ from $(\forall x/D)F(X)$ and $(\forall x/D)[F(x) \rightarrow G(x)]$ corresponds to the above conversion. It is interpreted as a transition of an input $(\forall x/D)F(X)$ to an output $(\forall x/D)G(x)$ by a conversion rule $(\forall x/D)[F(x) \rightarrow G(x)]$. These two transition processes are compared. For this purpose it is attempted to represent a symbolic representation of 'imply' $(\forall x/D)[F(x) \rightarrow G(x)]$ as a transition matrix.

A symbolic representation is to describe a common property of all components in a group of individuals. $(\forall x/D)$ denotes it. It is difficult to represent such an abstract concept as set in the form of simple conversion matrix. Instead it is necessary to represent a set operation by a concurrent processing of plural

individuals. In the following a transition matrix to represent an input-output relation for a set of input-output data is looked for and a condition of the transition matrix to be the same as a logical predicate is found.

First a symbolic implicative typed-formula $(\forall x/D)[F(x) \rightarrow G(x)]$ is considered as a form to represent a behavior of an object. Here D is a set of individual objects, $D = (a, b, c, \dots, z)$, and x/D means $x \in D$. Let the predicates F (and G) be interpreted as a property of x in D , that is, ' $F(x)$; an element x in D has a property F '. This is the case when a property is represented by a logical predicate like ' $F(x)$; a patient x has a fever'. Even though in the real case a finer representation like 'the temperature of the patient is $38.8 < C$ ' may be used, the objective for the moment is to combine it to 'the patient is flu' and to derive such a logical expression like 'if a patient is fever, then the patient is flu'. For the purpose one's temperature is divided into 'high temperature' and 'low temperature'. Then these are represented by predicates. This is the case when the object is represented in a simple symbolic expression as $(\forall x/D)[F(x) \rightarrow G(x)]$. If the finer expression is necessary, then the more complex predicate must be used as will be discussed later. Based on this assumption the following quantities are derived.

First a concept of 'state of D ' is introduced. It is represented by a combination of $F(x)$ for all x/D . For example, ' $F(a)$:True', ' $F(b)$:False', ' $F(c)$:False', $\dots$, ' $F(z)$:True' forms a state, say SF_I , of D with respect to F . Namely, $SF_I = (F(a), -F(b), -F(c), \dots, F(z))$. There are $N = 2^n (= 2 * n)$ different states. Let ' $F(x)$:True' and ' $F(x)$:False' be represented by 1 and 0 respectively. Then SF_I as above is represented $(1, 0, 0, \dots, 1)$. Let the sequence be identified by a binary number $I = 100 \dots 1$ obtained by concatenating 0 or 1 in the order of arrangement. Let SF_I be I -th state in N states. By arranging all states in the increasing order of I , a state vector Sf is defined. That is, $Sf = (SF_0, SF_1, \dots, SF_{N-1})$. Among them, $Sf = \{(1, 1, \dots, 1)\} = (\forall x/D)F(X)$ and $Sf_{\exists} = \{Sf - (0, 0, \dots, 0)\}$ are the only vectors that are allowed to use in the ordinary first order predicate. $(\exists x)$ denotes 'for some x '.

If the truth or false of F for one of the elements in D changes, then the state of D changes accordingly. Let this change occurs probabilistically. A state probability PF_I is defined to a state SF_I as probability of D being in the state SF_I and a probability vector Pf are also defined as $Pf = (PF_0, PF_1, \dots, PF_{N-1})$.

There is a reason why a probability distribution is introduced. When property of an input-output data is not given in advance it is necessary to derive a transition matrix by learning so as to fulfill the input-output relation for the given data. That is, transition matrix must be obtained by sequential approximation, by changing elements in the matrix by every data. In this case the result changes every moment by the new data and only probabilistic treatment is possible. This is a subjective probability. Note that, since a state vector is used, a data occurrence induces change of plural elements in the matrix. If, for example, there is a pair $\{F(a), G(a)\}$ in the data, all states in Sf and Sg including ' $F(a)$; True' and ' $G(a)$; True' are such elements and increased by a small amount while the other elements are decreased in order for keeping every row sum equal one.

The mathematical form, $Pg = Pf \times T$, with the transition matrix thus obtained is equivalent to a logical inference $F \wedge [F \rightarrow G] \Rightarrow G$ if a transition matrix $T = | t_{IJ} |$ satisfies a special condition as is shown in Figure 2 as an example. This matrix is made as follows. Since $F \rightarrow G = \neg F \vee G$ by definition, if 'F(x); True' for some x in D, then 'G(x); True' for the x must be true. That is, there is no transition from a state SF_I including 'F(x); True' to a state SG_J of D in regard to G including 'G(x); False' and t_{IJ} for this pair is put zero. The other elements of the transition matrix can be any positive values less than one. The above form is similar to a stochastic process. Considering the convenience of learning from database, the correspondence of the logical process and stochastic process is made and a condition of the row sum of the matrix is made equal to one for every row. Let this matrix be called L-matrix.

It should be noted that many elements in this matrix are zero. In case of non-symbolic system there is no such restriction to the transition matrix. This is the substantial difference between symbolic system and non-symbolic system and there is a big gap between them. Conversion between symbolic and non-symbolic processing is possible when and only when input-output data is such ones that make L-matrix, namely the data is essentially the instance of a logical predicate.

Note that this is a mere example for showing the gap between symbolic and non-symbolic systems. It is not suited for being used for discovering symbolic expression for a practical purpose. The more suited method can be provided in each application [5,6].

In order to know the characteristic of L-matrix as shown in Figure 2 an attempt is made to perform a series of the same transitions to make a limit of product of matrix $\lim_{n \rightarrow \infty} T^n$. Then every element in the output vector except qf decreases monotonically and qf approaches to one. Since qf corresponds the last state of the state vector $Sf = (SF_0, SF_1, \dots, SF_{N-1})$, i.e., 'F(a): True', 'F(b): True', 'F(c): True', ..., 'F(z): True', it means $(\forall x/D)G(x)$. After the system reach the state both input and output vectors are $(0, 0, 0, 0, \dots, 1)$ and any element other than tff(= $t_{N-1, N-1}$) in the matrix has no effect. Then the transition matrix of Figure 2 represented by pi as a state probability of the set D can be replaced by a simpler matrix of occurrence probabilities of each element in D of which all elements are zero except the diagonal elements are 1. This is the characteristic of expressions in a logical form. It means that the occurrence probability of individual element is independent to each other and no cross coupling between elements can occur. This is the basis of modularity of logical expressions.

Symbolic processing deals only with this symbolized object from the beginning. It means that if predicates used there are represented in the matrix form, it is of the form as shown in Figure 2. The difference between non-symbolic and symbolic processing is that the former deals with information before it comes to the one as shown in this figure while the latter deals with after reaching this stage. As the representation of each transition becomes so simple that symbolic processing can apart from the concept of transition and be achieved based on

	q0	q1	q2	q3	q4	q5	q6	q7	q8	q9	qa	qb	qc	qd	qe	qf
p0	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
p1	0	x	0	x	0	x	0	x	0	x	0	x	0	x	0	x
p2	0	0	x	x	0	0	x	x	0	0	x	x	0	0	x	x
p3	0	0	0	x	0	0	0	x	0	0	0	x	0	0	0	x
p4	0	0	0	0	x	x	x	x	0	0	0	0	x	x	x	x
p5	0	0	0	0	0	x	0	x	0	0	0	0	0	x	0	x
p6	0	0	0	0	0	0	x	x	0	0	0	0	0	0	x	x
p7	0	0	0	0	0	0	0	x	0	0	0	0	0	0	0	x
p8	0	0	0	0	0	0	0	0	x	x	x	x	x	x	x	x
p9	0	0	0	0	0	0	0	0	0	x	0	x	0	x	0	x
pa	0	0	0	0	0	0	0	0	0	0	x	x	0	0	x	x
pb	0	0	0	0	0	0	0	0	0	0	0	x	0	0	0	x
pc	0	0	0	0	0	0	0	0	0	0	0	0	x	x	x	x
pd	0	0	0	0	0	0	0	0	0	0	0	0	0	x	0	x
pe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	x	x
pf	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

(Ax/d)[F(x)→G(x)], D = (a1, a2, a3, a4) x; non-negative value with row sum =1

Fig. 2. Transition matrix to represent logical expression

matching and replacement of symbols. It enables variety of operations for integration. In order for repetitive integration operations becoming possible, the form of representations before and after integration must be the same. On the other hand non-symbolic processing cannot leave the matrix representation. Let a non-symbolic processor of which the transition matrix is L-matrix be said to meet the symbolization condition.

2.5 Extending to the More General Predicate

It was explained that “if condition of Figure 2 is met, then input-output data set can be translated into predicate logic”. However this is only for convenience of explanation and is not correct. The L-matrix of Figure 2 correspond to a simplest form of logic $(\forall x/D)[F(x) \rightarrow G(x)]$. If the form of logical predicate is different from it, L-matrix must also be different correspondingly. For example, in the case multiple predicate appears in the premise like $(\forall x/D)[F1(x) \wedge F2(x) \rightarrow G(x)]$ or plural variables are included like $(\forall x/D)(\forall y/E)[F1(x) \wedge F2(x, y) \rightarrow G(y)]$, the different L-matrix must be obtained.

If the premise of an implicative formula includes two predicates with the same variable like $(\forall x/D)[F1(x) \wedge F2(x) \rightarrow G(x)]$, then two independent states $Sf1$ and $Sf2$ of D are made corresponding to $F1(x)$ and $F2(x)$ respectively. Then a compound state Sf such as $Sf = Sf1 \times Sf2$ is made as the Cartesian product. From its compound probability vector Pf , a probability vector Pg for the state Sg is derived in the same way as the one before. In this case the number of states

in Sf is $2^{**}(2^{**}n)$ and transition matrix T becomes $2^{**}(2^{**}n) \times 2^{**}n$ matrix. Figure 3 shows this example. For the ease of representation a set D is assumed composed of only two elements, that is, $D = (a, b)$.

Further extension is necessary for the case of more than two variables, e.g. $(\forall x/D)(\forall y/E)[F1(x) \wedge F2(x, y) \rightarrow G(y)]$. In this case a new variable z defined over the set $D \times E$ is introduced and a transition matrix is made. The following treaty is similar to the above case. In this way the set of logical implicative forms with the corresponding transition matrices are generated. As the matter of course, the more complex is a formula, the more complex becomes the structure of the corresponding L-matrix. It is noted that structure of the matrix is determined by the form of logical formula irrespective of the predicate symbols included therein. The computation $Qg_j = \sum_I Pf_I \times t_{IJ}$ for $Qg = Pf \times T$ is formally the same as that included in an ordinary non-symbolic operation for transforming inputs to outputs.

	q0	q1	q2	q3			
p0	x	x	x	x	p0 ; ~F1(a)~F1(b) ~F2(a) ~F2(b)	q0 ; ~G(a) ~G(b)	
p1	x	x	x	x	p1 ; ~F1(a) ~F2(a)~F1(b) F2(b)	q1 ; ~G(a) G(b)	
p2	x	x	x	x	p2 ; ~F1(a) F2(a) ~F1(b) ~F2(b)	q2 ; G(a) ~G(b)	
p3	x	x	x	x	p3 ; ~F1(a) F2(a) ~F1(b) F2(b)	q3 ; G(a) G(b)	
p4	x	x	x	x	p4 ; ~F1(a) ~F2(a) F1(b) ~F2(b)		
p5	0	x	0	x	p5 ; ~F1(a) ~F2(a)F1(b) F2(b)		
p6	x	x	x	x	p6 ; ~F1(a) F2(a) F1(b) ~F2(b)		
p7	0	x	0	x	p7 ; ~F1(a) F2(a) F1(b) F2(b)		
p8	x	x	x	x	p8 ; F1(a) ~F2(a)~F1(b) ~F2(b)		
p9	x	x	x	x	p9 ; F1(a) ~ F2(a) ~F1(b) F2(b)		
pa	0	0	x	x	pa ; F1(a) F2(a) ~F1(b) ~F2(b)		
pb	0	0	x	x	pb ; F1(a) F2(a)~F1(b) F2(b)		
pc	x	x	x	x	pc ; F1(a) ~F2(a) F1(b) ~F2(b)		
pd	0	x	0	x	pd ; F1(a) ~F2(a) F1(b) F2(b)		
pe	0	0	x	x	pe ; F1(a) F2(a) F1(b) ~F2(b)		
pf	0	0	0	1	pf ; F1(a) F2(a) F1(b) F2(b)		

Fig. 3. Transition matrix to represent logical expression $(\forall x/D)[F1(x) \wedge F2(x) \rightarrow G(x)]$, $D = (a1, a2)$

2.6 Including Ambiguity in Logical Expression

Pure logic based only on True/False for every variable is not suited for some applications but an expression to include quantitative measure is required. Syntax of predicate logic is expanded to include some quantitative value like probability of truth while preserving its advantages of modularity as far as possible. In this case the transition matrix is different from that of Figure 2 This discussion is important but is not concern directly the following issues. Those who have interests in this topic, refer [1].

2.7 Problems for Bridging the Gap Between Non-symbolic and Symbolic Processing

There are two problems concerning bridging this gap between non-symbolic processing and symbol processing. The first (Problem 1) asks if a non-symbolic processor can do symbolic processing. This is to bridge the gap from non-symbolic processing side, i.e. by using only non-symbolic processor. This leads us to the discussion on the way human being acquired language. In an early stage of language acquisition, there must be a period in which non-symbolic processing behaved like a symbolic processor, i.e., could do logical inference. The second (Problem 2) asks a general method to convert non-symbolic information into symbolic form using a symbolic processing technology. This is to bridge the gap from the symbolic processing side. Recently many researches are being made on the latter issue as discovery problem or data mining problem. A general form of this problem is, “(for a set of individual objects of the similar nature) given observations on the properties of individuals, to investigate whether there is a common causal relation between them, and represent it in symbolic form.” These problems correspond to [A] and [B] in the four aspects mentioned in Section 1, respectively.

3 Aspect [A] of Intelligence Evolution - Creating Symbolic Language

3.1 Symbolic Processing by Non-symbolic Processor

The first problem (Problem 1) above is to ask if non symbolic processor can do the same processing as a symbolic processing. Neural network is assumed here as the engineering model of the basic structural components of living thing (Figure 4).

In a biological system, a signal from a (or more) sensor becomes an input and its output is sent to an (or more) actuator. It is assumed that many biological processors are provided to work in parallel in a living thing and each processor, say P_k , is identified by their geometrical locations or in relation with sensor-actuator pair. Let an input and an output of P_k be kI and kO , respectively. The input-output conversion network of this processor, $kO = P_k(kI)$, is mathematically a matrix multiplication. If a network identified by P_k tends to be a L-matrix as mentioned before, then this processor produces the same result as logical inference.

It is assumed that the network system to represent the transition matrix is made by learning. It means that if information being processed is such one as logical representation being possible, then ordinary non-symbolic processor tends to have L-matrix. Figure 5 shows a result of a simulation of learning. It starts from a matrix of which all elements are given the same value. Elements in this matrix are modified by the occurrence of data. It is shown that the matrix converges to an L-matrix.

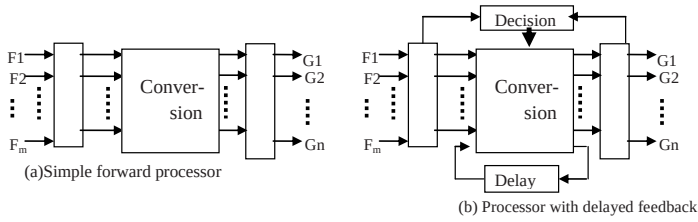


Fig. 4. Components of non-symbolic processing system

.078	.031	.004	.105	.021	.088	.073	.061	.048	.035	.084	.031	.043	.091	.122	.086
.000	.143	.000	.083	.000	.196	.000	.104	.000	.037	.000	.161	.000	.166	.000	.109
.000	.000	.090	.080	.000	.000	.167	.114	.000	.000	.067	.164	.000	.000	.136	.182
.000	.000	.000	.312	.000	.000	.000	.138	.000	.000	.000	.376	.000	.000	.000	.174
.000	.000	.000	.000	.146	.194	.071	.069	.000	.000	.000	.000	.195	.075	.164	.087
.000	.000	.000	.000	.000	.279	.000	.284	.000	.000	.000	.000	.000	.218	.000	.219
.000	.000	.000	.000	.000	.000	.333	.280	.000	.000	.000	.000	.000	.000	.209	.177
.000	.000	.000	.000	.000	.000	.000	.432	.000	.000	.000	.000	.000	.000	.000	.568
.000	.000	.000	.000	.000	.000	.000	.000	.156	.058	.161	.101	.066	.174	.144	.141
.000	.000	.000	.000	.000	.000	.000	.000	.000	.297	.000	.188	.000	.299	.000	.216
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.256	.314	.000	.000	.182	.248
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.566	.000	.000	.000	.434
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.363	.184	.255	.198
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.523	.000	.477
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.526	.474
.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.974

Fig. 5. A result of simulation

High level living thing that can do complex processing should have a complex biological structure. A feedback system with delay as is shown in Figure 4(b) is assumed instead of a simple forward processing as is shown in Figure 4(a). It represents a variety of cases by assigning properly parameters in it. For example, zero feedback-matrix means a processor without feedback. A processor that accepts two subsequent inputs is produced by making a proper feedback path with one time unit delay. More complex processor can also be produced. It is not clear yet whether such circuit is realizable biologically or what logical function can be achieved with it. It is considered here however as a basic circuit remaining a room for future extension of the theory. The decision part will be discussed in Section 3.3.

It is possible to connect different non-symbolic processors serially (lower part of Figure 6). If a coupling exists between elements, then it propagates to the succeeding processors. The forecasting the result of this chain is uneasy in general. If however the information through the chain is what is symbolized and every

processor has its own L-matrix, then modularity of processing is assured to each processor and the chain achieves the same operation as a deductive operation of symbolic processing.

3.2 Non-symbol Processor vs. Symbol Processor

This is what can happen to be in non-symbolic operation level but is not a true symbol processing. There must be an independent symbol processor to which a non-symbol processor satisfying the symbolization condition corresponds. The non-symbolic system must be translated into symbolic system over the syntactic difference between them. It is shown in Figure 6. In order to make clear this possibility, it is necessary to make clear the difference in their own syntax.

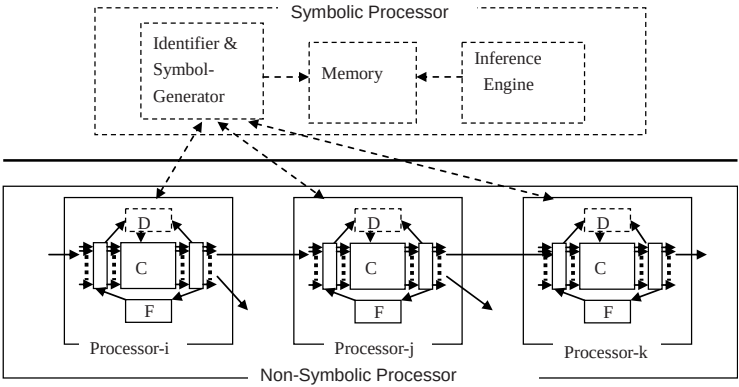


Fig. 6. A non-symbolic processor for enabling symbolic processing

Other than the fact that a symbolic processing uses symbols, there is another big difference between them. That is, a symbolic system processes information in a sequential manner while non-symbolic system processes information in parallel.

A set of functions is necessary for symbolic processing in whatever the way it may be achieved. These are; (1) to identify any symbolic expression, (2) to memorize any symbolic expression, (3) to make inference on symbolic expression. The term inference is used to mean a broad operation including a simple reference of a symbolic expression. Inference is performed in a sequential manner in the symbolic system with such a special set of functions as, (3-1) to memorize symbolic expression, (3-2) to search for matching different symbolic expressions and (3-3) to replace an expression with another expression [1].

First of all, how can a non-symbolic processor have the functions corresponding to them?

Let k-th processor be identified by P_k and last to exist after being produced. This is a kind of memorization. P_k is not explicitly represented but the processor is identified by its location ((1)). Memorization is to keep a generated processor un-erased ((2)). A processor meeting the symbolization condition can do

inference ((3)). Only such non-symbolic processor that meets the symbolization condition is translated into a symbolic processor. Then other than these, each non symbolic processor must have a decision function to examine whether the processor works as a symbolic processor. Mathematically it is to test for zero of predefined elements in the transition matrix. How it is realized by a biological system?

On the other hand, in order to prove the possibility of existence of symbol processor, it must be shown that (3-1), (3-2) and (3-3) are realized in a biological system. All these things are unknown yet and the matter to be made clear from now on.

3.3 Checking for Symbolization by Non-symbolic Processor

Symbolic processor deals with only such information that has no coupling between elements. To the contrary, non-symbolic processor has to deal with whatever the information. Let, in a processor as shown in Figure 4, a set of output $G_1(x)$, $G_2(x)$, ..., $G_n(x)$ be produced from a set of characteristics $F_1(x)$, $F_2(x)$, ..., $F_m(x)$ of an object as input. The decision part of Figure 4 tests for every G_i in the output for the equivalence to $(\forall x/D)[F_1(x) \wedge F_2(x) \wedge \dots, \wedge F_m(x) \rightarrow G_i(x)]$.

For ease of explanation, let there be two inputs $F_1(x)$ and $F_2(x)$ and one output $G(x)$. Its processing is shown in Figure 3. If in this case some non-zero value appears in the position that must be zero in the transition matrix as shown in Figure 3, then the process does not meet symbolization condition. The decision circuit checks for the condition. In a simpler way it can do checking the bottom right element. If it is nearly one the above decision is made approximately.

This decider forms a bypass to the main flow of information and has the higher level role than that. This kind of higher level operation must be included in every elemental processor for a whole system to grow to higher level.

This is a hypothetical model. It is designed to satisfy the requirement for explaining the process of human being acquiring language. There is no direct evidence to prove this hypothesis except an experiment by Y. Sekino and her group [3,4]. They found a phenomenon in Hippocampal Slice of rat a CA2 neuron showing a special signal pass. "The slices were prepared by oblique sectioning of the middle part of the hippocampus to preserve fiber connections. The mossy fibers were stimulated to include population spikes (PSs) and excitatory post synaptic potentials in the middle part of the CA1 region. Latencies of maximal PSs in CA1 varied widely among slices; they ranged from 7 to 13.5 ms, with two maxima at 9 and 11.5 ms. The fastest PSs probably are evoked by the Schaffer collaterals that connect the CA3 and CA1 regions in the well known trisynaptic circuit. However, the slower PSs suggest the existence of additional delayed input." Through the additional finer experiments, they conclude like "consideration of electrical signal propagation through the hippocampus should include flow through the CA2 region in addition to the traditional dentate gyrus-CA3-CA1 trisynaptic circuit." Sekino says that they don't know the meaning of this second path yet. But from the model of language acquisition point of view,

this result seems very important because it has clearly the same structure as detector and main signal processing as was shown in Figure 6. At the moment nothing can be said any further. What they found may or may not be the detector as mentioned above. There may be many similar kinds of higher level circuits having the different roles depending on the roles of the main flow of signal (like CA3 to CA1 above). It is left for the researches in the future. Collaboration of information science and neural science are necessary.

4 Aspect [B] of Intelligence Evolution-Knowledge Discovery from Data

After a language has been acquired, an effort to represent observed data in the form of language has began. This is still continuing as an important evolution of intelligence called as a discovery problem or a data mining. A general form of this problem is, “(for a set of individuals of the similar nature) given observations on the properties of individuals, to investigate whether there is a common causal relation between them, and represent it in symbolic form.” If a general method of finding knowledge is found, a big progress of intelligence is expected. This is to bridge the gap between non-symbolic and symbolic processing using symbolic processing technique or, in other words from the symbolic processing side.

Knowledge discovery and data mining form today a very important research area and many researches are being made [5,6]. It is not necessary to add something about it in this paper.

5 Aspect [C] of Intelligence Evolution-Conceptualization

5.1 Conceptualization

Language enabled accumulation of knowledge. Amount of knowledge increased and it accelerated the complexity of human activity. Furthermore an increase of knowledge urged the qualitative change of knowledge expression and made a new evolution of intelligence.

In the early stage, language syntax was very simple and knowledge expression was limited to direct description of observed object. The scope of knowledge expressions was limited to the objects in the neighbor of observers. Human activity was also simple and limited to ones connected directly to the object. It was decided by making use of this naive knowledge.

In principle, the scope of human activities is decided by the total amount of knowledge. When new knowledge is added the scope of activities expands and, consequently, the rate of achieving objectives may increase. Gradually human noted that a special meaning is hidden in a chunk of knowledge that was not derivable from each individual knowledge expression. If one can use it, then he/she can expand his/her scope of activities. In this way human recognized the existence of conceptual knowledge.

Conceptualization is a mental activity to think of knowledge as a whole. First such a chunk of knowledge that has a meaning as a whole (but was not seen in each of individual knowledge) is set. In general, such a meaning is found in relation to a subject's intention. Therefore there can be various conceptualizations depending on subject's view. But conceptualization as an activity 'to set a meaningful chunk of knowledge in a large whole of knowledge and finds its special meaning' is an important general step of intelligence evolution.

This conceptualization proceeded further. Many chunks of knowledge can be set in the whole knowledge and the same number of concepts is defined. In order to use these concepts their language expressions are made. These are the conceptual knowledge. Then in the same way as a concept has been made in the set of knowledge expressions on objects, it is possible to make a new concept with further special meaning based on this set of conceptual knowledge. This is the concept of concepts. Structurally an upper-level conceptualization proceeds in this way. The language expression of an upper level concept is knowledge on knowledge, that is, meta-knowledge. An expansion of language syntax is required accordingly.

Some new meta-knowledge can be knowledge on the way of using knowledge. It is possible to manage the method of performing activity based on knowledge. This kind of meta-knowledge is necessary for planning the more complex activity. For example, as knowledge increases then information generated by inference increases and necessary information can be embedded in the large amount of garbage knowledge. If the scope of knowledge to be used next is restricted properly by the control knowledge, then this problem is resolved.

People may say that, since an inductive inference produces knowledge that is not implied by the existing knowledge, it is to realize the above function in the form of logic. It is however not assured the completeness. Moreover as conceptualization is performed in relation with user's idea/intention, a variety of conceptualization must be possible. It is difficult to take this condition in inductive logic. Therefore, it is a very restricted form of a generally applicable method to conceptualization.

5.2 A Few Examples of Conceptualization

Knowledge represents some characteristic of an object (or objects) or relation between objects. For example, let there be relations $F(A, B)$ and $G(B, C)$ among objects A , B , and C . In this case A and C are linked indirectly via third object B . If the relations $F(A, B)$ and $G(B, C)$ are completely independent to each other in a user's view, the relation between A and C is very thin. On the other hand, if there is some commonality between $F(A, B)$ and $G(B, C)$ under a specific view of a user, the objects A and C are closely related (under this view) to each other via B . Similarly, if a relation $H(C, D)$ between C and D has commonality with $F(A, B)$ and $G(B, C)$ under the same view, then the set of objects A , B , C and D are bound in the same group under the view. Thus conceptualization is to discover a commonality among the relations between objects, make a set of object that can be bound in the same group, give a name to the common concept and express

it in a language. This discussion holds not only among binary relation like this but also holds among unary relation like $F(A)$, $G(B)$, $H(C)$ and $J(D)$. It holds in general among n -ary relations and if there is a commonality among these relations under a specific user's view, then a group of objects is formed.

There is no superficial commonality among F , G , H , $\dots$, in these explanations. But by introducing a user's view a commonality comes appear. In order to represent the concepts thus obtained explicitly, an expression of knowledge including this view is necessary. It needs high order logical expression. Depending on the view many specific concepts are made. Among them some common concepts survive as common knowledge.

Since conceptualization method depends on the subject's view, to find a general way of its realization is difficult. But if a view can be fixed to a specific one, then it is possible. Some interesting researches have been made in this way.

Y. Osawa proposes a method for discovering a chance of doing thing by defining a key graph composed of known objects and their relations [7,8]. First a key graph is made from, (1) a set of keywords relating to an intending event and appear frequently in an observation and (2) a set of links connecting the related keywords. Then a cluster of key words is made based on the strongly related links among them. This is the basis of conceptualization on intending event. Data mining and the other methods are used in this clustering. Osawa call this cluster an 'island'. A plural number of islands can be made in a key graph. The links that connects the different island are called bridges. A bridge is a knowledge representation to represent the relation between different concepts. Osawa showed that by making use of the bridge a chance of success for the intended event is found.

M. Akaishi [9,10] uses a similar technique for sentence retrieval. It is to retrieve information from sentence databases based on key sentence and generate an answer sentence. First every accumulated information is decomposed to a set of vocabularies, an attracting force is defined between vocabularies and a vocabulary chain graph is made. This is similar to the key graph of Osawa. By making use of vocabulary chain graphs of accumulated information and also of query sentence, a graph of answer is generated. From this graph a sentence or a story composed of plural sentences are created as an answer. In general the context of information when it is accumulated and when it is used is different. But this procedure enables to generate dynamically a new sentence based on the context of information use. The creation of sentence corresponds to concept formation. Since a view is fixed to sentence retrieval, an automatic formation of new concept becomes possible. Akaishi gives the methods of decomposing sentence information, of creating new sentence or story. In this case the components for creating new sentence are generated by decomposing given sentences. That is, low level concept is obtained from existing high level concept. Then this low level concept is used to create new high level concept.

Both of above methods enable conceptualization by fixing the views in advance. The method of creating upper level concept based on the connected set of base concepts is common. It may be possible to extend this idea in such a

way as to accept user's view in a formal way and to give a way of creating new concept on its basis. Akaishi's method, that is, to allow users to present a view in the form of sentence and to decompose it into a vocabulary graph, is one of such approaches.

6 Aspect [D] of Intelligence Evolution-Granular Computing

6.1 Problem and Problem Solving

Knowledge is significant for a person because it is used for achieving his/her objective. To find a way of achieving a goal of activity is problem solving. Problem appears in the different way by the case. For example in a case of system analysis, problem solving is to estimate unknown characteristic of the given system. In the case of engineering design as another example, problem solving is to find an artifact to meets a given condition/ constraint. In this cases problem solving is to find an object structure composed of many components satisfying the given performance requirement, safety requirement, cost requirement etc. Existing knowledge is used for problem solving. The higher the level of existing knowledge is, the more effectively it is used for problem solving.

Formally problem solving is to find an expression of a solution corresponding to a given expression of a problem. To express a problem is to describe unknown part of an object together with known part as well as the conditions these should satisfy as a whole. Problem solving is to insert a known expression in unknown part so that the whole expression becomes a known form. A known form is the form that does not include any unknown part. In addition its uniqueness is required in many cases. The more part is unknown in a problem expression, the more difficult is the problem to solve. A problem to which a practical method for leading a problem expression to a solution exists is a solvable problem. To find the path to solution by a computational method is automatic problem solving.

In a simple case such that a problem is expressed in a simple language, the solution is gotten by simple deductive inference and also represented in a simple language. The more complex problem cannot be represented by a few sentences of a language. A complex problem is such that includes many components that are related in a complicated way to each other. A representation of a solution of a complex problem may also be complex such as to be represented by a structure of mutually related components. An expression of problem must be such that an expression of the solution can be derived thereof. Solving a complex problem is difficult in general. But, if an expression of a problem is made properly, it is possible to find a method for solving it. That is, the major reason why solving complex problem is difficult is that representation of problem is difficult.

In general a problem representation is a mixture of known part and unknown part. In a sense to make such a problem representation is to create a new knowledge that is in a higher level than any existing knowledge. This is an intelligent task by definition and is the meaning of an aspect [D] of intelligence

evolution. Different from the aspect [A] through [C] that were all bottom up and to obtain general knowledge, this is a top down operation concerning a specific object.

6.2 Granular Computing

A problem becomes complex when it deals with a complex object. For this class of problems a concept of granular computing is attracting attentions. The term “Granular Computing” was first used by T.Y. Lin. Since then many researches have been made in the topic [11]. It was defined by L.A. Zadeh quoting Merriam-Webster’s Dictionary as “a small particle; especially, one of numerous particle forming a larger unit” [12]. Granulation involves a decomposition of whole into parts. Conversely, organization involves an integration of parts into whole. A granular thinking is necessary for solving a complex problem. The problem is represented in a multi-level structure. Then methods based on the granular computing are used [13,14]. Even though many theoretical researches are being made however, it is still a way of thinking rather than a concrete method today. In the next section a method of problem solving in this way of thinking is proposed.

6.3 Building Knowledge Structure to Represent and Solve Problems

In many cases a complex problem is made referring to an existing case. For instance a problem is made by changing some parts of a sample problem (Object 1) to a desirable object (Object 2). Both Object 1 and Object 2 are compounds (structure) of number of smaller objects, that is, granules. Object 2 is a prototype of goal representation. It involves some unknown terms as well as condition/constraint which should be met by the object. For example, when a certain performance is required to an artifact in an engineering design problem, the requirement is written into Object 2 and the unknown part of the object are decided to satisfy this requirement. If the requirement description is lacked, the right solution cannot be arrived. That is, the problem representation is not valid. To refer to existing cases or knowledge for making a rough goal (solution) structure is not a special approach but is almost only method of making a prototype of problem representation in relatively a short time. After then, a problem solving starts. It is to replace the unknown parts in goal by some known term meeting the given requirement (condition/ constraint).

It is observed that many real complex problems are represented in the similar way and human usually do the fixed granular thinking for complex problem. It means that formalizing a way of representing unknown goal object is possible. The resultant form must be such that autonomous operation for problem solving is possible.

A complex physical object is represented as composed of physical components. The characteristics of the compound are decided by the characteristics of components and their structural relations of components. A component may

be decomposed further to a set of the smaller components. Whole-component relation is represented in a structural form in which representation of characteristic/property/attribute represented in a language form is distributed. Some of characteristic/property/attribute as well as structural parts can be put unknown. The latter is the problem of building up a new structure such as the case of engineering design. In this case the prototype of goal structure is not made at the beginning but problem representation and problem solving proceeds step by step. An example is shown in Figure 7. It represents a hypothetical case of a new sea transportation system.

6.4 Problems Including Human Problem Solver

Many complex problems involve human and, since human capability and characteristics affect the solution, it must be solved taking human characteristics in account [15,16,17]. A problem is represented as a pair of human (or a subject of an activity in general, simply a subject in the sequel) and a physical object that the subject is responsible to deal with. This pair represents an object to be considered by a human problem solver. Its formal representation is a pair of entities; one to represent a subject and the other one to represent physical object. It is shown in the top of Figure 7. A subject labeled A represents a problem solver and the object being considered by A is shown in a double ellipse.

The object is designed as composed of a set of sub-systems as shown in a box labeled 2. At the time an important decision is made on the way of achieving the goal because there are many possibilities of realizing the goal with intended performance. This is the first step of problem solving. Figure 7 shows a case that the problem solver selected a hydro-foil craft. Whether this is the correct decision or not is not yet clear at this stage. It may be proved that this decision is not good after the many further steps of problem solving have been performed. Then problem solving should come back to this stage and retries another method. After the first selection has been made, the lower structure of the object is made depending on the decision. A new lower structure has to be made for each subsystem. And the same procedure goes on.

Instead of doing the same procedure downward further the top subject may assign some persons who are responsible to continue this procedure. These persons are shown as B, C, ..., D. Each subject is given a representation of ability/characteristics. Let it be called a sub-subject. After then each subsystem is dealt with by the person. Let it be called a sub-object in the whole object. In general the sub-object of a complex problem can be still complex and can be decomposed to a structure composed of a number of components. An attribute/property is given to every component.

The relation of a sub-subject to a sub-object is the similar as that of a subject and an object. This relation can continue to the lower level and a large problem is represented as of a multi-level structure of smaller problems. In this representation there can be the relations between components of physical objects and also between subjects involved in the whole structure. These relations are also represented explicitly.

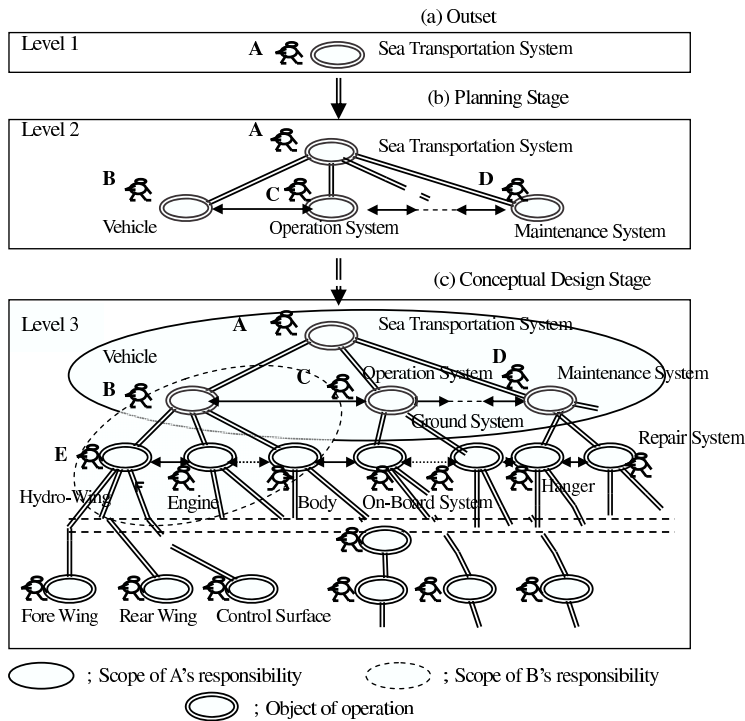


Fig. 7. Construction of problem structure

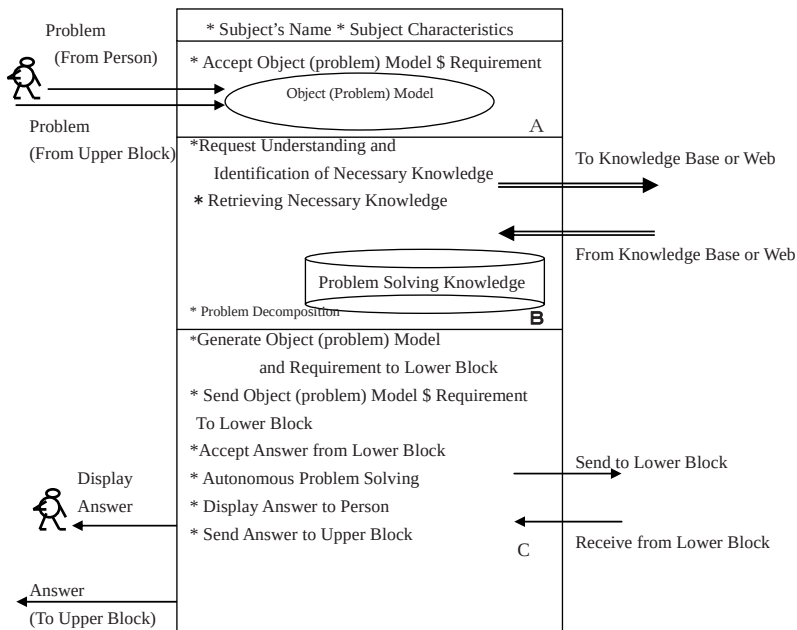
Let it be a formal granular representation of a problem. It can be a basis of granular computing. Human problem solver can represent a problem keeping this scheme in mind. In this process some parts are represented as unknown. Problem solving is to decide these unknown parts so that a reasonable representation of whole object is reached.

The final objective of problem solving is to derive or achieve a functionality of the top object in many cases such as performance, security, cost etc. In principle it is possible to remain every part unknown at the beginning. Then it is difficult to get solution. Usually as many part of the whole object as possible are made as known. That is, it is important to fix the form of the whole object as far as possible. This is necessary not only for making problem solving easier but also for enabling human problem solver understand the situation he/she is put and to represent a problem as correct as possible. Human can make clear his idea and arrive at understanding the matter he is considering by writing down his idea on an object in consideration. This is a substantial characteristic of human understanding. Computer must be able to support human in this process. In an early time when people did not understand this fact, problem solving was limited to only very simple problems.

6.5 Generation of Problem Solving Systems

After the problem is represented a method of solving this problem must be generated. Problem solving is first to organize such a system then to execute it. Since so many unknowns can be distributed over the problem representation and each of which requires the different method, the problem solving method itself must be a structure of these methods. It is shown that a multi-agent system is very suited for this purpose because it can be defined to fit the given problem representation. A subject-object pair in the problem representation forms a unit of processing. In Figure 7 a scope of problem solving by each subject is shown by an ellipse. Every scope has the same form. It is replaced by an agent. If the agent is so designed as to accept any kind of problem to occur in a scope, then it can be a common building block for constructing a global problem solving structure. In this case it is possible to make a multi-agent problem solving system automatically by assigning an agent to each subject-object pair in the problem representation. An agent has to work to decide the unknown part in the scope to which the agent assumes responsibility. It is shown in Figure 8. The relation between subjects and between objects described in the problem representation is implemented as the communication between agents.

In reality, every problem for a subject-object pair is very different. This difference can be absorbed by making the agent as a knowledge-based problem solving system. A problem solving system, A problem solving is,



A; Accept Problem/ Generate Problem Model B; Retrieve Knowledge C; Solve Problem

Fig. 8. Organization of building block

- (1) to accept problems from outside (human or upper block),
- (2) to identify the necessary knowledge,
- (3) to retrieve the knowledge,
- (4) to decompose the problem, define a problem to solve (make a model), to assign a lower subject (generate a new agent) if necessary and then to solve the problem using the knowledge,
- (5) to send back the answer.

It is assumed here that

- (a) Every agents is connected to a network so that communication between agents is possible.
- (b) A large knowledge base containing wide range of knowledge is provided so that every agent can retrieve any necessary chunk of knowledge for problem solving from it.

In general there is two types of operation in an agent. One is to find a lower level structure to satisfy the required functionality or characteristics and the other is to find the functionality characteristics based on the given structure.

As the problem representation is the same as the human image of an object in the real world, and problem solving system is built automatically based on the problem representation, it is comprehensive for persons and the error can be reduced. In a sense the procedure of problem solving is a simulation to ensure the correctness of problem representation, that is, human thinking.

The structure of the solution obtained in this way is a new concept on the object being considered for the user. It is derived from existing knowledge. But different from the conceptualization of [C], it is to make a concept that is some levels upper than existing knowledge by top-down operation and accordingly a completely different approach is necessary. It is the granular computing.

7 Conclusion

Intelligence has been discussed based on information. Today works that require high level intelligence are rapidly increasing such as developing software, robotics, intelligent Web and so on. Formalization and computerization of intelligence is required strongly. But these are considered difficult because intelligence is believed to concern many aspects of human mental activity. Apart from its relation to mental activity, however, it is possible to discuss intelligence formally based on information it deals with.

On the basis of this idea this paper defined intelligence as capability to upgrading information and noted that there are four important aspects of intelligence. These are: (1) language acquisition, (2) knowledge discovery, (3) conceptualization and (4) granulation. These aspects correspond to the different phases of progress of intelligence in the history of human being. By extending this discussion, it is expected that intelligence can be measured quantitatively. However there are many problems remained unsolved yet in each of the aspect. Efforts to solve these problems may lead us to the better understanding and computerization of intelligence.

Among these four aspects, the first is mostly in the cloud of darkness. It is mainly because it requires a view extending over the different disciplines such as information science, neuro-science and brain science. Collaboration of researchers in these research fields is strongly expected.

References

1. Ohsuga, S.: Bridging the Gap between Non-Symbolic and Symbolic Processing - How Could Human Being Acquire Language? - (to appear)
2. Ohsuga, S.: Symbol Processing by Non-Symbol Processor. In: Foo, N.Y., Göbel, R. (eds.) PRICAI 1996. LNCS, vol. 1114, Springer, Heidelberg (1996)
3. Sekino, Y., Obata, K., Tanifuji, M., Mizuno, M., Murayama, J.: Delayed Signal Propagation via CA2 in Rat Hippocampal Slices Revealed by Optical Recording. *Journal of The American Physiological Society*, 1662–1668 (1997)
4. Sekino, Y., Tanaka, S., et al.: Activation of N-Metyle-D-aspartate Receptor Induces a Shift of Drebrin Distribution: Disappearance from Dendritic Spines and Appearance in Dendritic Shafts. *Mol. Cell. Neuroscience* 31, 493–504 (2006)
5. Klossgen, W., Zytkow, J.M.: *Handbook of Data Mining and Knowledge Discovery*. Oxford Univ. Press, Oxford (2002)
6. Ohsuga, S.: Difference between Data Mining and Knowledge Discovery - A view to Discovery from Knowledge Processing. In: *Proc. of 2005 IEEE International Conference on Granular Computing*, pp. 7–12 (2005)
7. Ohsawa, Y., McBurney, P.: *Chance Discovery*. Springer, Heidelberg (2003)
8. Ohsawa, Y.: Data Crystallization: Chance Discovery with Unobservable Events. *New Mathematics and Natural Computation* 1(3), 373–392 (2005)
9. Akaishi, M., Satoh, K., Tanaka, Y.: An Associative Information Retrieval based on the Dependency of term Co-occurrence. In: Suzuki, E., Arikawa, S. (eds.) *DS 2004*. LNCS (LNAI), vol. 3245, pp. 195–206. Springer, Heidelberg (2004)
10. Akaishi, M.: A Dynamic Decomposition / Recomposition Framework for Documents based on Narrative Structure Model. *Journal of Japanese Society for Artificial Intelligence* 21(5), 428–438 (2006) (in Japanese)
11. Lin, T.Y.: *Granular Computing*. Announcement of the BISC Special Interest Group on Granular Computing (1997)
12. Zadeh, L.A.: Key roles of information granulation and Fuzzy logic in human reasoning, Concept formulation and computing with words. In: *Proc. of IEEE 5th International Fuzzy Systems*, p. 1 (1996)
13. Yao, Y.Y.: Perspectives of Granular Computing. In: *Proc. of, IEEE International Conference on Granular Computing*, pp. 85–90 (2005)
14. Yao, Y.Y.: Granular Computing. *Computer Science* 31(10.A), 1–5 (2004)
15. Ohsuga, S.: Toward Truly Intelligent Information Systems - From Expert Systems to Automatic Programming. *Knowledge Based Systems* 10 (1998)
16. Ohsuga, S.: How Can AI Systems Deal with Large and Complex Problems? *International Journal of Pattern Recognition and Artificial Intelligence* 15(3) (2001)
17. Ohsuga, S.: Significance of AI in Solving Problems That Can Not Be Foreseen Beforehand. *Journal of Japanese Society for Artificial Intelligence* 20 (2005) (in Japanese)

Toward Perception Based Computing: A Rough-Granular Perspective

Andrzej Jankowski¹ and Andrzej Skowron²

¹ Institute of Decision Processes Support
and AdgaM Solutions Sp. z o.o.
Wąwozowa 9 lok. 64, 02-796 Warsaw, Poland
andrzejj@adgam.com.pl

²Institute of Mathematics,
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl

Abstract. We discuss the Wisdom Granular Computing (WGC) as a basic methodology for Perception Based Computing (PBC). By wisdom, we understand an adaptive ability to make judgements correctly to a satisfactory degree (in particular, correct decisions) having in mind real-life constraints. We propose Rough-Granular Computing (RGC) as the basis for WGC.

Keywords: Granular computing, rough sets, rough-granular computing, perception based computing, ontology approximation, wistech.

1 Introduction

At the beginning, we would like to refer to Gottfried Wilhelm Leibniz. In [39] one can find the following sentence:

If controversies were to arise, there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, and say to each other: Let us calculate.

and in [40]:

Languages are the best mirror of the human mind, and that a precise analysis of the signification of words would tell us more than anything else about the operations of the understanding.

Hence, Gottfried Wilhelm Leibniz can be considered a precursor of Granular Computing (GC) understood as a calculus of human thoughts. Through centuries since then, mathematicians have been developing tools to deal with this issue. Unfortunately, the developed tools in crisp mathematics, in particular, in mathematical logic do not yet allow for understanding natural language used by human to express thoughts and reasoning about them, an understanding which will allow us to construct truly intelligent systems.

One of the reason is that humans, capable of solving efficiently many real-life problems, are able to express their thoughts by means of vague, uncertain,

imprecise concepts and reason about such concepts. Lotfi Zadeh (see, e.g., [89]) proposed to base the calculus of thoughts using fuzzy logic to move from computing with numbers to computing with words, and further from manipulations of measurements to manipulations of perceptions and Perception Based Computing (PBC). This idea has been developed by Lotfi Zadeh himself in a number of papers (see, e.g., [90,91]) and by other researchers, using also rough set methods (see, e.g., [56]).

In [86,87,88] Lotfi Zadeh proposed the term ‘information granule’: *An information granule is a clump of objects of some sort, drawn together on the basis of indistinguishability, similarity or functionality. In this definition, being general enough to comprise a large number of special cases, the stress is laid on the reasons for clustering objects into clumps, and three such motives are suggested: indistinguishability, similarity, and functionality.*

There are several papers on rough set theory in which an attempt has been made to develop methods for calculi of information granules (see, e.g., [56]).

In [32] Wisdom Technology (WisTech) is discussed as one of the main paradigms for development of new applications in intelligent systems. In this paper, it is emphasized that in developing more advanced applications, especially for PBC, we are moving from data granules to information granules, then from information granules to knowledge granules, and finally from knowledge granules to wisdom granules corresponding, e.g., to perception of compound situations.

In this paper, we discuss all these granules with a special emphasis on wisdom granules. Calculi of such advanced granules are very important for making progress in the development of intelligent systems.

Solving complex problems, e.g., by multi-agent systems requires new approximate reasoning methods based on new computing paradigms. One such recently emerging computing paradigm is Rough Granular Computing (RGC). Computations in RGC are performed on information granules representing often vague, partially specified, and compound concepts delivered by agents engaged in tasks such as knowledge representation, communication with other agents, and reasoning.

The research on the foundations on WGC and, in particular, on RGC is based on a continuation of approaches to computational models of approximate reasoning developed by Rasiowa (see [66]), Pawlak (see [57]) and their students. In some sense, it is a succession of ideas initiated by Leibniz, Boole and currently continued in a variety of forms. Of course, the Rasiowa - Pawlak school is also some kind of continuation of the Polish School of Mathematics and Logics which led to the development of the modern understanding of the basic computational aspects of logic, epistemology, ontology, foundations of mathematics and natural deduction. The two fundamental tools of the Rasiowa - Pawlak school [33] are: (i) *Computation models of logical concept (especially such concepts as deduction or algebraic many-valued models for classical, modal, and constructive mathematics)* - based on the method of treating the sets of logically equivalent statements (or formulas) as abstract algebras known as Lindebaum - Tarski algebras;

(ii) *Computation models of vague concept*- originally Łukasiewicz has proposed to treat uncertainty (or vague concepts) as concepts of many valued logic. The rough set concept, due to Pawlak [57], developed in the Rasiowa-Pawlak school, is based on classical two valued logic. The rough set approach has been developed to deal with uncertainty and vagueness. The approach makes it possible to reason precisely about approximations of vague concepts. These approximations are temporary, subjective, and are adaptively changing with changes in environments [8,71,74].

This paper is organized as follows. First, we outline WisTech in Section 2. In Section 3, we discuss a general concept of calculi of granules used in optimization processes in which the goals are achieved by performing computations on granules. Also a general definition of granule is discussed. Different kinds of granules are presented in Section 4. We start from data granules, next are presented information granules which are used to construct knowledge granules. The most advances granules are wisdom granules on which adaptive judgement is performed in solving real-life problems. Calculi of wisdom granules are discussed in Section 4.4. Examples of problems in developing of such calculi are discussed. In particular, algorithmic methods for developing efficient calculi of wisdom granules for solving real-life problems are one of the main challenge of GC.

2 Wisdom Technology (WisTech)

In this section, we give a short introduction to Wistech. For more details on WisTech the reader is referred to [32].

There are many indications that we are currently witnessing the onset of an era of radical changes depending on the further advancement of technology to acquire, represent, store, process, discover, communicate and learn wisdom. In this paper, we call this technology *wisdom technology* (or WisTech, for short). The term *wisdom* commonly means “judging rightly” [34]. This common notion can be refined. By *wisdom*, we understand an adaptive ability to make judgements correctly to a satisfactory degree (in particular, correct decisions) having in mind real-life constraints.

One of the basic objectives of the paper is to indicate the role of GC for the design and implementation of WisTech computation models. An important aspect of WisTech is that the complexity and uncertainty of real-life constraints mean that in practice we must reconcile ourselves to the fact that our judgements are based on non-crisp concepts and also do not take into account all the knowledge accumulated and available to us. This is why consequences of our judgements are usually imperfect. But as a consolation, we also learn to improve the quality of our judgements via observation and analysis of our experience during interaction with the environment. Satisfactory decision-making levels can be achieved as a result of improved judgements.

The intuitive nature of wisdom understood in this way can be expressed metaphorically as shown in (1).

$$wisdom = KSN + AJ + IP, \quad (1)$$

where *KSN*, *AJ*, *IP* denote *knowledge sources network*, *adaptive judgement*, and *interactive processes*, respectively. The combination of the technologies represented in (1) offers an intuitive starting point for a variety of approaches to designing and implementing computational models for WisTech. The equation (1) is called the *wisdom equation*. There are many ways to build WisTech computational models.

The issues discussed in this article are relevant for the current research directions (see, e.g., [15,16,23,29,35,43,67,79,84,92,93,94,95,96,97,98] and the literature cited in these articles and books).

Our approach to Wistech is based on RGC.

3 Calculi of Granules: General Comments

GC is needed to reach feasibility and efficiency in solving hard real-life problems not solvable by traditional methods. Such problems are related to distributed system of units (agents) interacting with each other and with dynamically changing environments. The information available to agents about objects and concepts is usually only partial and concepts are often vague. Moreover, this information is dynamically changing, which requires that the developed methods should be adaptive to these changes.

Calculi of granules create the core for GC. They consist of atomic granules and operations making it possible to construct more compound granules from already constructed granules. The challenge is how, for given problems or a class of problems, such calculi can be developed. Granular calculi for WisTech are compound and they require further study.

Computations in GC are performed on granules. The aim of computations is to reach the goal. Usually, this is an optimization process [4].

GC can deal with problems specified by means of vague concepts. The solution constructed for them should satisfy a given specification to a satisfactory degree. Certainly, this creates problems but also, as Lotfi Zadeh observed, e.g., in the foreword to [56] then ... *there is a tolerance for imprecision which can be exploited to achieve tractability, robustness and low solution cost*. Moreover, for many problems *the expressive power of words is higher than the expressive power of numbers* and/or *the available information is not precise enough to justify the use of numbers*.

Computation in GC are performed under partial and uncertain information about objects and concepts which are often vague. This requires special modeling techniques for building models of concepts. These techniques allow us to induce relational and syntactical structures used to represent objects and to express their properties. The relevant structures should be discovered to assure the satisfactory approximation of the considered concepts. Concept approximations are constructed over semantical and syntactical structures. Computations in GC should make it possible to reason from measurements by sensors to higher level

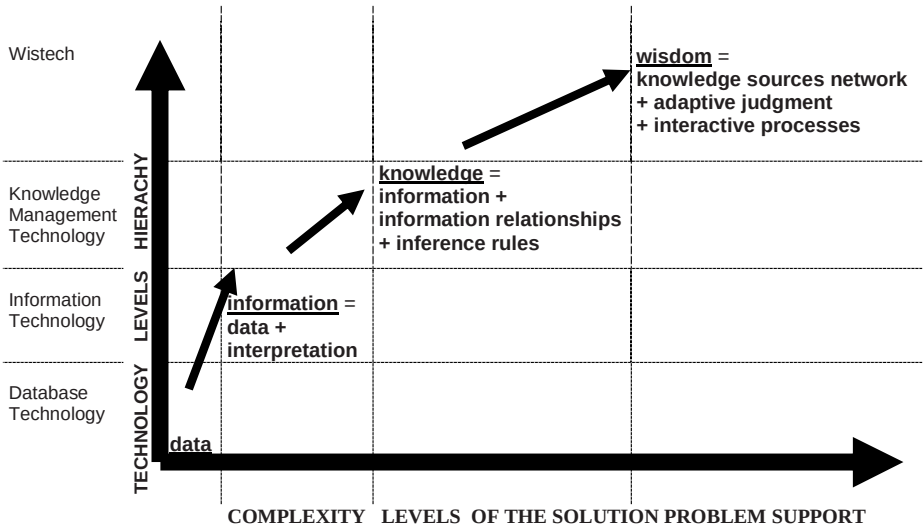


Fig. 1. Wisdom equation context

concepts representing perception (see, e.g., [56]). Due to the interaction of granules among themselves and with the environment the information about objects and concepts is changing. This requires developing adaptive methods for constructing new relevant granules for the current situation. Granules can be tuned in the optimization process. Usually, optimization is performed on parameters of granules [74].

4 From Data Granules to Wisdom Granules

In [75] several kinds of granules and operations on granules have been discussed. In this section we characterize granules from another perspective. We discuss granules corresponding to the diagram presented in [32] (see Figure 1) and discussed in the context of the wisdom equation (see (1)).

In this figure, there are four distinguished kinds of granules: data granules, information granules, knowledge granules, and wisdom granules. They correspond to technologies: Database Technology, Information Technology, Knowledge Management Technology, and WisTech, respectively. In [32] the main features of these technologies were discussed. Here, we discuss granules corresponding to these technologies.

4.1 Data Granules

In Figure 1 the term ‘data’ is understood as a stream of symbols without any interpretation of their meaning.

More formally, one can define data granules assuming that there is a given relational structure over a given set V_a of values of a given attribute a [57,58]

or over the Cartesian product of such sets for a given set of attributes B , i.e., $\prod_{a \in B} V_a$. Then by taking a language of formulas of signature defined by this relational structure, one can define subsets of $\prod_{a \in B} V_a$ equal to the semantic meaning of formulas with some free variable in the considered structure. These sets are data granules. For example, one can consider a linear order on V_a and define as data granules intervals. Having relational structures over V_a and V_b one can extend them on $V_a \times V_b$ by adding some constrain relations, e.g., in $V_a \times V_b$, representing for example the closeness of values from V_a and V_b if these sets are subsets of real numbers. Then formulas over the extended structure define new data granules.

4.2 Information Granules

Information granules are related to modeled objects. A typical example of information granules in rough sets are indiscernibility classes or similarity classes of objects defined by information systems or decision tables [57,58]. Usually, these granules are obtained as follows. For example, if $\alpha(x)$ is a formula defining a data granule with values for x in the set of attribute values vectors then an information granule can be defined assuming that an object y (from the considered set of objects, e.g., a set of patients) is satisfying a formula $\alpha^*(y)$ if and only if the B -signature of this object $Inf_B(y)$ satisfies α [57,58]. In this way, subsets of objects are defined which are interpreted as information granules. Certainly, one can consider also families of information granules an information granule. A typical example is a set of indiscernibility classes considered as an information granule.

4.3 Knowledge Granules

New knowledge granules are constructed by means of some operations from already established granules. At the beginning some atomic (primitive) granules should be distinguished. In constructing knowledge granules inductive reasoning is also used, e.g., in constructing granules corresponding to classifiers (see, e.g., [75]).

One can consider different relationships between information granules. Association rules [1,26] in a given information system can be treated as a standard example of relations between information granules. Another example are relationships defined by given two families of indiscernibility classes creating a partition of the universe of objects. The relationships between such granules can be defined using positive regions [57,58] or entropy [26,59,60].

One can consider different kinds of inference rules related to knowledge. These rules can be treated as schemes making it possible to derive properties of some knowledge granules from properties of other knowledge granules used for their construction. Dependencies between approximated concepts are examples of such rules. Another example is given by approximate reasoning schemes (AR schemes) (see, e.g., [73]) representing constructions of compound patterns from more elementary ones together with information how degrees of inclusion of these

elementary patterns in some input concepts are propagating to the degree of inclusion of the compound pattern in the target concept. Other examples of rules of inference on knowledge granules are considered in reasoning about knowledge [30].

It is worthwhile to mention that in GC there is no opposition between computational and dynamical approaches [49,31]. For example, one can consider granules corresponding to dynamical models of some processes with the syntactical structure represented by differential equation with initial conditions and the semantical structure representing solutions of equations. Moreover, one can consider interaction between such granules and other computational granules, e.g., representing discrete models of other processes. Results of such interactions are compositions of models of processes represented by interacting granules.

4.4 Wisdom Granules and Calculi of Wisdom Granules Based on RGC

Let us recall from [32] some general features of Wistech. From the perspective of the metaphor expressed in the wisdom equation (1), WisTech can be perceived as the integration of three technologies (corresponding to three components in the wisdom equation (1)). The first component is related to knowledge source networks:

1. *knowledge sources network* – by knowledge we traditionally understand every organized set of information along with the inference rules; in this context one can easily imagine the following examples illustrating the concept of knowledge sources network:
 - representations of states of reality perceived by our senses (or observed by the “receptors” of another observer) are integrated as a whole in our minds in a network of sources of knowledge and then stored in some part of our additional memory,
 - a network of knowledge levels represented by agents in some multi-agent system and the level of knowledge about the environment registered by means of receptors.

The second component is about interactive processes:

2. *interactive processes* – interaction understood as a sequence of stimuli and reactions over time; examples are:
 - the dialogue of two people,
 - a sequence of actions and reactions between an unmanned aircraft and the environment in which the flight takes place, or
 - a sequence of movements during some multi-player game.

Far more difficult conceptually seems to be the concept of adaptive judgement distinguishing wisdom from the general concept of problem solving. Intuitions behind this concept can be expressed as follows:

1. *adaptive judgement* – understood here as arriving at decisions resulting from the evaluation of patterns observed in sample objects. This form of judgement is made possible by mechanisms in a metalanguage (meta-reasoning)

which, on the basis of available knowledge sources' selection and on the basis of understanding interactive processes' history and their current status, enable us to perform the following activities under real-life constraints:

- identification and judgement of importance (for future judgement) of sample phenomena, available for observation, in the surrounding environment;
- planning current priorities for actions to be taken (in particular, on the basis of understanding interactive processes' history and their current status) toward making optimal judgements;
- selection of fragments of ordered knowledge (hierarchies of information and judgement strategies) satisfactory for making a decision at the planned time (a decision here is understood as a commencing interaction with the environment or as selecting the future course to make judgements);
- prediction of important consequences of the planned interaction of processes;
- adaptive learning and, in particular, reaching conclusions deduced from patterns observed in sample objects leading to adaptive improvement in the adaptive judgement process.

One of the main barriers hindering an acceleration in the development of wis-tech applications lies in developing satisfactory computation models implementing the functioning of “adaptive judgement”. This difficulty primarily consists in overcoming the complexity of the process of integrating the local assimilation and processing of changing non-crisp and incomplete concepts necessary to make correct judgements. In other words, we are only able to model tested phenomena using local (subjective) models and interactions between them. In practical applications, usually, we are not able to give global models of analyzed phenomena (see, e.g., [80,42,44,45,21,17]). However, we can only approximate global models by integrating the various incomplete perspectives of problem perception. One of the potential computation models for “adaptive judgement” might be the *rough-granular approach*. In particular, in PBC it is necessary to develop modeling methods for approximation of complex concepts being results of perception. Moreover, these approximations should be adaptively adjusted to the perceived changes in environment.

Granule calculi consist of some atomic granules and operations on granules used for new granule generation. Wisdom granules have compound structure and are generated by sophisticated operations. Both wisdom granules and operations should be discovered and adaptively learned.

Let us consider some examples of wisdom granules. Certainly, for their construction all other granules are used. However, the mentioned above tasks for WisTech require developing wisdom granules constructed by some sophisticated operations or strategies. Usually wisdom granules have a compound structure. Their different parts are responsible for different tasks such as interaction with other granules and environment, adaptive judgement, communication with knowledge networks or optimization of different granule parameters.

Let us start from examples of granules resulting by interaction between granules. The simplest example is a decision table representing the result of interaction between condition and decision attributes or two agents observing the same objects. The next example concerns the interaction of a given object and a decision rule. The result of such an interaction is a matching degree of this object to the rule. Continuing this example one can consider the results of interaction of rule with decision class, object with classifier, the interaction of measurements by sensors with ontology of concepts or behavioral patterns. One can also consider agents as compound granules and their interactions producing new granules. They can be related to actions, plans, agent cooperation or competition, coalition formation etc. Interactions in multiagent systems are intensively studied and are one of the main topics in multiagent systems.

In adaptive judgement we need wisdom granules representing degrees of satisfiability of concepts represented in the granule, granules for expressing the propagation of satisfiability degrees from sensor measurements to higher level concepts representing perception, granules used for representation of reasoning about changes in adaptive learning of action selection or plan execution. For generating such granules advanced strategies should be developed. Such strategies are parts of wisdom granules. Adaptive judgement requires strategies for adaptive learning of concepts represented in wisdom granules. Hence, in particular, granules representing the discovered by such strategies semantical and syntactical structures, as well as granules representing discovered new patterns and concepts constructed out of the patterns, are needed.

Advanced strategies are required to extract from compound granules their relevant parts. In particular, advanced strategies are needed to extract, for a given granule representing the goal and for given granules representing knowledge networks, some granules representing relevant fragments of such knowledge networks. Analogous strategies should be developed for relational and syntactical structures considered in the context of a given goal for given sensor measurements.

5 Rough Granular Computing

In this section we outline the Rough Granular Computing (RGC).

Developing methods for approximation of compound concepts expressing the result of perception belongs to the main challenges of Perception Based Computing (PBC). The perceived concepts are expressed in natural language. We discuss the rough-granular approach to approximation of such concepts from sensory data and domain knowledge. This additional knowledge, represented by ontology of concepts, is used to make it feasible the search for features (condition attributes) relevant for the approximation of concepts on different levels of the concept hierarchy defined by a given ontology. We report several experiments of the proposed methodology for approximation of compound concepts from sensory data and domain knowledge. The approach is illustrated by examples relative to interactions of agents, ontology approximation, adaptive hierarchical

learning of compound concepts and skills, behavioral pattern identification, planning, conflict analysis and negotiations, and perception-based reasoning. The presented results seem to justify the following claim of Lotfi A. Zadeh: "In coming years, granular computing is likely to play an increasingly important role in scientific theories-especially in human-centric theories in which human judgement, perception and emotions are of pivotal importance". The question how ontologies of concepts can be discovered from sensory data remains as one of the greatest challenges for many interdisciplinary projects on learning of concepts.

The concept approximation problem is the basic problem investigated in machine learning, pattern recognition and data mining [26]. It is necessary to induce approximations of concepts (models of concepts) consistent (or almost consistent) with some constraints. In the most typical case, the constraints are defined by a training sample. For more compound concepts, we consider constraints defined by domain ontology consisting of vague concepts and dependencies between them. Information about the classified objects and concepts is partial. In the most general case, the adaptive approximation of concepts is performed under interaction with dynamically changing environment. In all these cases, searching for sub-optimal models relative to the minimal length principle (MLP) is performed. Notice that in adaptive concept approximation one of the components of the model should be the adaptation strategy. Components involved in construction of concept approximation which are tuned in searching for sub-optimal models relative to MLP are called information granules. In rough granular computing (RGC), information granule calculi are used for construction of components of classifiers and classifiers themselves (see, e.g., [74]) satisfying given constraints. An important mechanism in RGC is related to generalization schemes making it possible to construct more compound patterns from less compound patterns. Generalization degrees of schemes are tuned using, e.g., some evolutionary strategies.

Rough set theory due to Zdzisław Pawlak [57,58,59,60,19], is a mathematical approach to imperfect knowledge. The problem of imperfect knowledge has been tackled for a long time by philosophers, logicians and mathematicians. Recently it became also a crucial issue for computer scientists, particularly in the area of artificial intelligence. There are many approaches to the problem of how to understand and manipulate imperfect knowledge. The most successful one is, no doubt, the fuzzy set theory proposed by Lotfi A. Zadeh [85]. Rough set theory presents still another attempt to solve this problem. It is based on an assumption that objects and concepts are perceived by partial information about them. Due to this some objects can be indiscernible. From this fact it follows that some sets can not be exactly described by available information about objects; they are rough not crisp. Any rough set is characterized by its (lower and upper) approximations. The difference between the upper and lower approximation of a given set is called its boundary. Rough set theory expresses vagueness by employing a boundary region of a set. If the boundary region of a set is empty it means that the set is crisp, otherwise the set is rough (inexact). A nonempty

boundary region of a set indicates that our knowledge about the set is not sufficient to define the set precisely. One can recognize that rough set theory is, in a sense, a formalization of the idea presented by Gotlob Frege [25].

One of the consequences of perceiving objects using only available information about them is that for some objects one cannot decide if they belong to a given set or not. However, one can estimate the degree to which objects belong to sets. This is another crucial observation in building the foundations for approximate reasoning. In dealing with imperfect knowledge one can only characterize satisfiability of relations between objects to a degree, not precisely. Among relations on objects the rough inclusion relation, which describes to what degree objects are parts of other objects, plays a special role. A rough mereological approach (see, e.g., [65,73,56]) is an extension of the Leśniewski mereology [41] and is based on the relation *to be a part to a degree*. It will be interesting to note here that Jan Łukasiewicz was the first who started to investigate the inclusion to a degree of concepts in his discussion on relationships between probability and logical calculi [46].

The very successful technique for rough set methods was Boolean reasoning [14]. The idea of Boolean reasoning is based on construction for a given problem P a corresponding Boolean function f_P with the following property: the solutions for the problem P can be decoded from prime implicants of the Boolean function f_P . It is worth to mention that to solve real-life problems it is necessary to deal with Boolean functions having a large number of variables.

A successful methodology based on the discernibility of objects and Boolean reasoning has been developed in rough set theory for computing of many key constructs like reducts and their approximations, decision rules, association rules, discretization of real value attributes, symbolic value grouping, searching for new features defined by oblique hyperplanes or higher order surfaces, pattern extraction from data as well as conflict resolution or negotiation [69,52,60]. Most of the problems involving the computation of these entities are NP-complete or NP-hard. However, we have been successful in developing efficient heuristics yielding sub-optimal solutions for these problems. The results of experiments on many data sets are very promising. They show very good quality of solutions generated by the heuristics in comparison with other methods reported in literature (e.g., with respect to the classification quality of unseen objects). Moreover, they are very time-efficient. It is important to note that the methodology makes it possible to construct heuristics having a very important approximation property. Namely, expressions generated by heuristics (i.e., implicants) close to prime implicants define approximate solutions for the problem (see, e.g., [2]).

The rough set approach offers tools for approximate reasoning in multiagent systems (MAS). The typical example is the approximation by one agent of concepts of another agent. The approximation of a concept is based on a decision table representing information about objects perceived by both agents.

The strategies for data models inducing developed so far are often not satisfactory for approximation of compound concepts that occur in the perception

process. Researchers from the different areas have recognized the necessity to work on new methods for concept approximation (see, e.g., [13,83]). The main reason for this is that these compound concepts are, in a sense, too far from measurements which makes the searching for relevant features infeasible in a very huge space. There are several research directions aiming at overcoming this difficulty. One of them is based on the interdisciplinary research where the knowledge pertaining to perception in psychology or neuroscience is used to help to deal with compound concepts (see, e.g., [48,24,23]). There is a great effort in neuroscience towards understanding the hierarchical structures of neural networks in living organisms [22,64,48]. Also mathematicians are recognizing problems of learning as the main problem of the current century [64]. These problems are closely related to complex system modeling as well. In such systems again the problem of concept approximation and its role in reasoning about perceptions is one of the challenges nowadays. One should take into account that modeling complex phenomena entails the use of local models (captured by local agents, if one would like to use the multi-agent terminology [45,80,21]) that should be fused afterwards. This process involves negotiations between agents [45,80,21] to resolve contradictions and conflicts in local modeling. This kind of modeling is becoming more and more important in dealing with complex real-life phenomena which we are unable to model using traditional analytical approaches. The latter approaches lead to exact models. However, the necessary assumptions used to develop them result in solutions that are too far from reality to be accepted. New methods or even a new science therefore should be developed for such modeling [27].

One of the possible approaches in developing methods for compound concept approximations can be based on the layered (hierarchical) learning [77,11]. Inducing concept approximation should be developed hierarchically starting from concepts that can be directly approximated using sensor measurements toward compound target concepts related to perception. This general idea can be realized using additional domain knowledge represented in natural language. For example, one can use some rules of behavior on the roads, expressed in natural language, to assess from recordings (made, e.g., by camera and other sensors) of actual traffic situations, if a particular situation is safe or not (see, e.g., [53,10,9,19]). The hierarchical learning has been also used for identification of risk patterns in medical data and extended for therapy planning (see, e.g. [7,6]). Another application of hierarchical learning for sunspot classification is reported in [54]. To deal with such problems one should develop methods for concept approximations together with methods aiming at approximation of reasoning schemes (over such concepts) expressed in natural language. The foundations of such an approach, creating a core of perception logic, are based on rough set theory [57,58,59,60,19] and its extension rough mereology [65,73,56]. The (approximate) Boolean reasoning methods can be scaled to the case of compound concept approximation.

In the following section, we discuss more examples.

6 Some Solutions Based on WGC and Challenges for WGC

The prediction of behavioral patterns of a compound object evaluated over time is usually based on some historical knowledge representation used to store information about changes in relevant features or parameters. This information is usually represented as a data set and has to be collected during long-term observation of a complex dynamic system. For example, in case of road traffic, we associate the object-vehicle parameters with the readouts of different measuring devices or technical equipment placed inside the vehicle or in the outside environment (e.g., alongside the road, in a helicopter observing the situation on the road, in a traffic patrol vehicle). Many monitoring devices serve as informative sensors such as GPS, laser scanners, thermometers, range finders, digital cameras, radar, image and sound converters (see, e.g. [81]). Hence, many vehicle features serve as models of physical sensors. Here are some exemplary sensors: location, speed, current acceleration or deceleration, visibility, humidity (slipperiness) of the road. By analogy to this example, many features of compound objects are often dubbed sensors. Some rough set tools been developed (see, e.g., [9]) for perception modelling that make it possible to recognize behavioral patterns of objects and their parts changing over time. More complex behavior of compound objects or groups of compound objects can be presented in the form of *behavioral graphs*. Any behavioral graph can be interpreted as a *behavioral pattern* and can be used as a complex classifier for recognition of complex behaviours. The complete approach to the perception of behavioral patterns, based on behavioral graphs and the dynamic elimination of behavioral patterns, is presented in [9]. The tools for dynamic elimination of behavioral patterns are used for switching-off in the *system attention* procedures searching for identification of some behavioral patterns. The developed rough set tools for perception modeling are used to model networks of classifiers. Such networks make it possible to recognize behavioral patterns of objects changing over time. They are constructed using an ontology of concepts provided by experts that engage in approximate reasoning on concepts embedded in such an ontology. Experiments on data from a vehicular traffic simulator [5] show that the developed methods are useful in the identification of behavioral patterns.

The following example concerns human computer-interfaces that allow for a dialog with experts to transfer to the system their knowledge about structurally compound objects. For pattern recognition systems [20], e.g., for Optical Character Recognition (OCR) systems it will be helpful to transfer to the system a certain knowledge about the expert view on border line cases. The central issue in such pattern recognition systems is the construction of classifiers within vast and poorly understood search spaces, which is a very difficult task. Nonetheless, this process can be greatly enhanced with knowledge about the investigated objects provided by a human expert. We developed a framework for the transfer of such knowledge from the expert and for incorporating it into the learning process of a recognition system using methods based on rough mereology (see, e.g., [55]).

Is is also demonstrated how this knowledge acquisition can be conducted in an interactive manner, with a large dataset of handwritten digits as an example.

The next two examples are related to approximation of compound concepts in reinforcement learning and planning.

In reinforcement learning [78,18,47,36,74,61,62,63], the main task is to learn the approximation of the function $Q(s, a)$, where s, a denotes a global state of the system and an action performed by an agent ag and, respectively and the real value of $Q(s, a)$ describes the reward for executing the action a in the state s . In approximation of the function $Q(s, a)$ probabilistic models are used. However, for compound real-life problems it may be hard to build such models for such a compound concept as $Q(s, a)$ [83]. We propose another approach to the approximation of $Q(s, a)$ based on ontology approximation. The approach is based on the assumption that in a dialog with experts an additional knowledge can be acquired, making it possible to create a ranking of values $Q(s, a)$ for different actions a in a given state s . In the explanation given by expert about possible values of $Q(s, a)$, concepts from a special ontology are used. Then, using this ontology one can follow hierarchical learning methods to learn the approximations of concepts from ontology. Such concepts can have a temporal character too. This means that the ranking of actions may depend not only on the actual action and the state but also on actions performed in the past and changes caused by these actions.

In [7,6] a computer tool based on rough sets for supporting automated planning of the medical treatment (see, e.g., [28,82]) is discussed. In this approach, a given patient is treated as an investigated complex dynamical system, whilst diseases of this patient (RDS, PDA, sepsis, Ureaplasma and respiratory failure) are treated as compound objects changing and interacting over time. As a measure of planning success (or failure) in experiments, we use a special hierarchical classifier that can predict the similarity between two plans as a number between 0.0 and 1.0. This classifier has been constructed on the basis of the special ontology specified by human experts and data sets. It is important to mention that besides the ontology, experts provided the exemplary data (values of attributes) for the purpose of concepts approximation from the ontology. The methods of construction such classifiers are based on approximate reasoning schemes (AR schemes, for short) and were described, e.g., in [10,53,10,9]. We applied this method for approximation of similarity between plans generated in automated planning and plans proposed by human experts during the realistic clinical treatment.

One of the WGC, and in particular, RGC challenges is to develop approximate reasoning techniques for reasoning about dynamics of distributed systems of judges, i.e., agents judging rightly. These techniques should be based on systems of evolving local perception logics rather than on a global logic [70,72]. The approximate reasoning about global behavior of judge's system is infeasible without methods for approximation of compound vague concepts and approximate reasoning about them. One can observe here an analogy to phenomena related to the emergent patterns in complex adaptive systems [17]. Let us observe that judges can be organized into a hierarchical structure, i.e., one judge can represent a coalition

of judges in interaction with other agents existing in the environment [3,37,42]. Such judges representing coalitions play an important role in hierarchical reasoning about behavior of judges populations. Strategies for coalition formation and cooperation [3,42,44] are of critical importance in designing systems of judges with dynamics satisfying to a satisfactory degree the given specification. Developing strategies for discovery of information granules representing relevant coalitions and cooperation protocols is another challenge for RGC.

All these problems can be treated as problems of searching for information granules satisfying vague requirements. The strategies for construction of information granules should be adaptive. It means that the adaptive strategies should make it possible to construct information granules satisfying constraints under dynamically changing environment. This requires reconstruction or tuning of already constructed information granules which are used as components of data models, e.g., classifiers. In the adaptive process, the construction of information granules generalizing some constructed so far information granules plays a special role. The mechanism for relevant generalization here is crucial. One can imagine for this task many different strategies, e.g., based on adaptive feedback control for tuning the generalization. Cooperation with specialists from different areas such as neuroscience (see, e.g., [48] for visual objects recognition), psychology (see, e.g., [64] for discovery of mechanisms for hierarchical perception), biology (see, e.g., [12] for cooperation based on swarm intelligence) or social science (see, e.g., [42] for modeling of agents behavior) can help to discover such adaptive strategies for extracting sub-optimal (relative to the minimal length principle) data models satisfying soft constraints. This research may also help us to develop strategies for discovery of ontologies relevant for compound concept approximation.

Let us consider one more challenge for WGC. There is given a distributed system of locally interacting agents (parts). These interactions lead to the global pattern represented by means of so called *emergent* behaviors. One can model such a system as a special kind of game in which each agent can only interact with agents in a local neighborhood of this agent. The strategy of each agent in the game is defined by rules of interaction of this agent with neighbors. The emergent patterns are very hard to predict. There are many real-life problems of this kind, e.g., ecological systems, immune systems, economies, global climate, ant colonies. There are well known emergent patterns of such systems [50,68]. The challenge we consider is related to developing strategies for learning local interactions among agents so that a given emergent pattern will appear. The evolutionary strategies are here the candidates (see, e.g., [50,51,68]). We would like to emphasize the role of WGC in learning such strategies. One can easily observe that the learning process should be organized in such a way that gradually learned interactions lead to granules of agents represented by their coalitions rather than by particular agents. The behavior of each coalition is determined by specific interactions of agents in coalition and interaction of the coalition as the whole with other agents or coalitions. Learning a hierarchy of coalition is hard but this can be made feasible by using a domain knowledge analogously to hierarchical learning (see, e.g., [9,6]). Domain knowledge can help to discover

languages for expressing behavioral patterns on each level of coalitions and relevant behavioral patterns. Without such “hints” the learning seems to be infeasible for real-life problems. Using this approach, the evolutionary strategies should only make it possible to learn interactions on each level of hierarchy between existing on this level coalitions leading to discovery on the next level relevant new coalitions. Note that each coalition can be treated as a concurrent system. Let us observe that the discovered coalition by relevant granulation of lower level coalitions will usually satisfy the specification for this higher level coalition only to a degree. The discovered coalitions, i.e., new granules can be treated as approximate solutions on a given level of hierarchy. It is worthwhile to mention that another possibility is to learn relevant hierarchical coalitions for gradually changing tasks for the whole system from simple to more compound.

Finally, let us observe that the following four principles of adaptive information processing in decentralized systems presented in [51] are central for WGC and RGC in searching for relevant granules by using evolutionary techniques:

1. Global information is encoded as statistics and dynamics of patterns over the system’s components.
2. Randomness and probabilities are essential.
3. The system carries out a fine-grained, parallel search of possibilities.
4. The system exhibits a continual interplay of bottom-up and top-down processes.

7 Conclusions

We presented WGC as a basic methodology in WisTech for PBC. Different kinds of granules, from data granules to the most compound wisdom granules were discussed. Several important features of granules were distinguished such as size, diversity, structure with parts, ability to interact with other granules, and adaptiveness. The important role of RGC for solving problems related to WisTech was discussed. Conclusions from the current projects for WGC and RGC were reported and some challenges for WGC and RGC were included.

We would like to extend our current studies to fragments of natural language in which the results of perception and solutions of simple tasks (e.g., performed by robots) will be expressed. This will require developing more advanced than so far methods for expressing semantics of expressions from such fragments in terms of sensor measurements. The link between these two very distant semantical levels, i.e. the sensor level and the perception level will be built hierarchically using methods based on RGC developed in PBC and WisTech.

Acknowledgments

The research of Andrzej Jankowski was supported by Institute of Decision Process Support. The research of Andrzej Skowron has been supported by the grant from Ministry of Scientific Research and Information Technology of the Republic of Poland.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C, pp. 207–216. ACM Press, New York (1993)
2. Rough Set Exploration System (RSES). Available at: <http://logic.mimuw.edu.pl/~rses>
3. Axelrod, R.M.: *The Complexity of Cooperation*. Princeton University Press, Princeton, NJ (1997)
4. Bargiela, A., Pedrycz, W.: *Granular Computing: An Introduction*. Kluwer Academic Publishers, Dordrecht (2003)
5. Bazan, J.: The Road simulator. Available at: <http://logic.mimuw.edu.pl/~bazan/simulator>
6. Bazan, J., Kruczek, P., Bazan-Socha, S., Skowron, A., Pietrzyk, J.J.: Automatic planning of treatment of infants with respiratory failure through rough set modeling. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) *RSCTC 2006. LNCS (LNAI)*, vol. 4259, pp. 418–427. Springer, Heidelberg (2006)
7. Bazan, J., Kruczek, P., Bazan-Socha, S., Skowron, A., Pietrzyk, J.J.: Risk pattern identification in the treatment of infants with respiratory failure through rough set modeling. In: *Proceedings of IPMU 2006, Éditions E.D.K., Paris, July 2-7, 2006*, pp. 2650–2657 (2006)
8. Bazan, J., Skowron, A., Swiniarski, R.: Rough sets and vague concept approximation: From sample approximation to adaptive learning. In: Peters, J.F., Skowron, A. (eds.) *Transactions on Rough Sets V. LNCS*, vol. 4100, pp. 39–62. Springer, Heidelberg (2006)
9. Bazan, J.G., Peters, J.F., Skowron, A.: Behavioral pattern identification through rough set modelling. In: Ślęzak, et al. (eds.), pp. 688–697 [76]
10. Bazan, J.G., Skowron, A.: Classifiers based on approximate reasoning schemes. In: Dunin-Kępicz, et al. (eds.), pp. 191–202 [21]
11. Behnke, S.: *Hierarchical Neural Networks for Image Interpretation. LNCS*, vol. 2766. Springer, Heidelberg (2003)
12. Bonabeau, E., Dorigo, M., Theraulaz, G.: *Swarm Intelligence. From Natural to Artificial Systems*. Oxford University Press, Oxford, UK (1999)
13. Breiman, L.: Statistical modeling: The two cultures. *Statistical Science* 16(3), 199–231 (2001)
14. Brown, F.: *Boolean Reasoning*. Kluwer Academic Publishers, Dordrecht (1990)
15. Cassimatis, N.L.: A cognitive substrate for achieving human-level intelligence. *AI Magazine* 27, 45–56 (2006)
16. Cassimatis, N.L., Mueller, E.T., Winston, P.H.: Achieving human-level intelligence through integrated systems and research. *AI Magazine* 27, 12–14 (2006)
17. Desai, A.: Adaptive complex enterprises. *Comm. ACM* 48, 32–35 (2005)
18. Dietterich, T.G.: Hierarchical reinforcement learning with the MAXQ value function decomposition. *Artificial Intelligence* 13(5), 227–303 (2000)
19. Doherty, P., Łukaszewicz, W., Skowron, A., Szalas, A.: Knowledge Representation Techniques: A Rough Set Approach. In: *Studies in Fuzziness and Soft Computing* 202, Springer, Heidelberg (2006)
20. Duda, R., Hart, P., Stork, R.: *Pattern Classification*. John Wiley & Sons, New York (2002)

21. Dunin-Kępicz, B., Jankowski, A., Skowron, A., Szczuka, M. (eds.): Monitoring, Security, and Rescue Tasks in Multiagent Systems (MSRAS'2004). *Advances in Soft Computing*. Springer, Heidelberg (2005)
22. Fahle, M., Poggio, T.: *Perceptual Learning*. MIT Press, Cambridge (2002)
23. Forbus, K.D., Hinrichs, T.R.: Companion cognitive systems: A step toward human-level ai. *AI Magazine* 27, 83–95 (2006)
24. Forbus, K.D., Hinrichs, T.R.: Engines of the brain: The computational instruction set of human cognition. *AI Magazine* 27, 15–31 (2006)
25. Frege, G.: *Grundgesetzen der Arithmetik*, 2. Verlag von Hermann Pohle, Jena (1903)
26. Friedman, J., Hastie, T., Tibshirani, R.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Heidelberg (2001)
27. Gell-Mann, M.: *The Quark and the Jaguar - Adventures in the Simple and the Complex*. Brown and Co., London (1994)
28. Ghallab, M., Nau, D., Traverso, P.: *Automated Planning: Theory and Practice*. Elsevier, Morgan Kaufmann, CA (2004)
29. Granger, R.: Engines of the brain: The computational instruction set of human cognition. *AI Magazine* 27(2), 15–31 (2006)
30. Halpern, J.Y., Fagin, R., Moses, Y., Vardi, M.Y.: *Reasoning about Knowledge*. MIT Press, Cambridge (1995)
31. Ivancevic, V.G., Ivancevic, T.T.: *Geometrical Dynamics of Complex Systems. A Unified Modelling Approach to Physics, Control, Biomechanics, Neurodynamics and Psycho-Socio-Economical Dynamics*. Springer, Dordrecht (2006)
32. Jankowski, A., Skowron, A.: A wistech paradigm for intelligent systems. In: Peters, J.F., Skowron, A., Düntsch, I., Grzymała-Busse, J., Orłowska, E., Polkowski, L. (eds.) *Transactions on Rough Sets VI: Journal Subline*, vol. 4374, pp. 94–132. Springer, Heidelberg (2006)
33. Jankowski, A., Skowron, A.: Logic for artificial intelligence: A Rasiowa–Pawlak school perspective. In: Ehrenfeucht, A., Marek, W., Srebrny, M. (eds.) *Seventy Years of Foundational Studies*, IOS Press, Amsterdam (2007)
34. Johnson, S.: *Dictionary of the English Language in Which the Words are Deduced from Their Originals, and Illustrated in their Different Significations by Examples from the Best Writers*, 2 volumes, F.C. and J. Rivington, London (1816)
35. Jones, R.M., Wray, R.E.: Comparative analysis of frameworks for knowledge-intensive intelligent agents. *AI Magazine* 27(2), 57–70 (2006)
36. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4, 227–303 (1996)
37. Kraus, S.: *Strategic Negotiations in Multiagent Environments*. MIT Press, Massachusetts (2001)
38. Langley, P.: Cognitive architectures and general intelligent systems. *AI Magazine* 27, 33–44 (2006)
39. Leibniz, G.W.: *Dissertio de Arte Combinatoria*, Leipzig (1666)
40. Leibniz, G.W.: *New Essays on Human Understanding (1705)* Translated and edited by Peter Remnant and Jonathan Bennett, Cambridge UP, Cambridge (1982)
41. Leśniewski, S.: Grunzüge eines neuen Systems der Grundlagen der Mathematik. *Fundamenta Mathematicae* 14, 1–81 (1929)
42. Liu, J.: *Autonomous Agents and Multi-Agent Systems: Explorations in Learning, Self-Organization and Adaptive Computation*. World Scientific Publishing, Singapore (2001)

43. Liu, J.: Web Intelligence (WI): What makes Wisdom Web? In: Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003) pp. 1596–1601 (2003)
44. Liu, J., Jin, X., Tsui, K.C.: *Autonomy Oriented Computing: From Problem Solving to Complex Systems Modeling*. Kluwer/Springer, Heidelberg (2005)
45. Luck, M., McBurney, P., Preist, C.: *Agent Technology. Enabling Next Generation Computing: A Roadmap for Agent Based Computing*. AgentLink (2003)
46. Łukasiewicz, J.: Die logischen Grundlagen der Wahrscheinlichkeitsrechnung, Kraków 1913. In: Borkowski, L. (ed.) *Jan Łukasiewicz - Selected Works*, pp. 16–63. North Holland & Polish Scientific Publishers, Amsterdam, London, Warsaw (1970)
47. McGovern, A.: *Autonomous Discovery of Temporal Abstractions from Interaction with an Environment*. PhD thesis, University of Massachusetts, Amherst (2002)
48. Mikkiläinen, R., Bednar, J.A., Choe, Y., Sirosh, J.: *Computational Maps in the Visual Cortex*. Springer, Heidelberg (2005)
49. Mitchell, M.: A complex-systems perspective on the "Computation vs. Dynamics" debate in cognitive science. In: Gernsbacher, M.A., Derry, S.J. (eds.) *Proceedings of the 20th Annual Conference of the Cognitive Science Society (COGSCI 1998)*, pp. 710–715 (1998)
50. Mitchell, M., Newman, M.: *Complex systems theory and evolution*. In: Pagel, M. (ed.) *Encyclopedia of Evolution*, Oxford University Press, New York (2002)
51. Mitchell, M.: *Complex systems: Network thinking*. *Artificial Intelligence* 170(18), 1194–1212 (2006)
52. Nguyen, H.S.: *Approximate boolean reasoning: Foundations and applications in data mining*. In: Peters, J.F., Skowron, A. (eds.) *Transactions on Rough Sets V*. LNCS, vol. 4100, pp. 344–523. Springer, Heidelberg (2006)
53. Nguyen, H.S., Bazan, J., Skowron, A., Nguyen, S.H.: *Layered learning for concept synthesis*. In: Peters, J.F., Skowron, A., Świniarski, R.W., Szczuka, M. (eds.) *Transactions on Rough Sets I*. LNCS, vol. 3100, Springer, Heidelberg (2004)
54. Nguyen, S.H., Nguyen, T.T., Nguyen, H.S.: *Rough set approach to sunspot classification*. In: Ślęzak, et al. (eds.), pp. 263–272 [76]
55. Nguyen, T.T., Skowron, A.: *Rough set approach to domain knowledge approximation*. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) *RSFDGrC 2003*. LNCS (LNAI), vol. 2639, pp. 221–228. Springer, Heidelberg (2003)
56. Pal, S.K., Polkowski, L., Skowron, A. (eds.): *Rough-Neural Computing: Techniques for Computing with Words*. Cognitive Technologies. Springer, Heidelberg (2004)
57. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*. System Theory, Knowledge Engineering and Problem Solving 9. Kluwer Academic Publishers, Dordrecht, The Netherlands (1991)
58. Pawlak, Z., Skowron, A.: *Rudiments of rough sets*. *Information Sciences* 177(1), 3–27 (2007)
59. Pawlak, Z., Skowron, A.: *Rough sets: Some extensions*. *Information Sciences* 177(1), 28–40 (2007)
60. Pawlak, Z., Skowron, A.: *Rough sets and boolean reasoning*. *Information Sciences* 177(1), 41–73 (2007)
61. Peters, J.F.: *Approximation spaces for hierarchical intelligent behavioural system models*. In: D-Keplić, B., Jankowski, A., Skowron, A., Szczuka, M. (eds.) *Monitoring, Security and Rescue Techniques in Multiagent Systems*. *Advances in Soft Computing*, pp. 13–30. Physica-Verlag, Heidelberg (2004)

62. Peters, J.F.: Rough ethology: Towards a biologically-inspired study of collective behaviour in intelligent systems with approximation spaces. In: Peters, J.F., Skowron, A. (eds.) *Transactions on Rough Sets III*. LNCS, vol. 3400, pp. 153–174. Springer, Heidelberg (2005)
63. Peters, J.F., Henry, C.: Reinforcement learning with approximation spaces. *Fundamenta Informaticae* 71(2-3), 323–349 (2006)
64. Poggio, T., Smale, S.: The mathematics of learning: Dealing with data. *Notices of the AMS* 50(5), 537–544 (2003)
65. Polkowski, L., Skowron, A.: Rough mereology: A new paradigm for approximate reasoning. *International Journal of Approximate Reasoning* 15(4), 333–365 (1996)
66. Rasiowa, H.: *Algebraic models of logics*. Warsaw University, Warsaw (2001)
67. Schlenoff, C., Albus, J., Messina, E., Barbera, A.J., Madhavan, R., Balakirsky, S.: Using 4d/rcs to address ai knowledge integration. *AI Magazine* 27, 71–81 (2006)
68. Segel, L.A., Cohen, I.R. (eds.): *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Oxford University Press, New York (2001)
69. Skowron, A.: Rough sets in KDD (plenary talk). In: Shi, Z., Faltings, B., Musen, M. (eds.) *16-th World Computer Congress (IFIP 2000) Proceedings of Conference on Intelligent Information Processing (IIP 2000)*, Publishing House of Electronic Industry, Beijing, pp 1-14 (2000)
70. Skowron, A.: Perception logic in intelligent systems. In: Blair, S., et al. (eds.) *Proceedings of the 8th Joint Conference on Information Sciences (JCIS 2005)*, Salt Lake City, Utah, USA, July 21-26, 2005, X-CD Technologies: A Conference & Management Company, Toronto, Ontario, Canada pp. 1-5 (2005)
71. Skowron, A.: Rough sets and vague concepts. *Fundamenta Informaticae* 64(1-4), 417–431 (2005)
72. Skowron, A.: Rough sets in perception-based computing (keynote talk). In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) *PReMI 2005*. LNCS, vol. 3776, pp. 21–29. Springer, Heidelberg (2005)
73. Skowron, A., Stepaniuk, J.: Information granules and rough-neural computing. In: Pal, et al. (eds.), [56] pp. 43–84.
74. Skowron, A., Stepaniuk, J., Peters, J.F., Swiniarski, R.: Calculi of approximation spaces. *Fundamenta Informaticae* 72(1-3), 363–378 (2006)
75. Skowron, A., Stepaniuk, J.: Rough sets and granular computing: Toward rough-granular computing. In: Pedrycz, W., Skowron, A., Kreinovich, V. (eds.) *Handbook of Granular Computing*, Wiley, New York (in preparation, 2007)
76. Ślęzak, D., Yao, J., Peters, J.F., Ziarko, W., Hu, X. (eds.): *RSFDGrC 2005*. LNCS (LNAI), vol. 3642. Springer, Heidelberg (2005)
77. Stone, P.: *Layered Learning in Multi-Agent Systems: A Winning Approach to Robotic Soccer*. MIT Press, Cambridge (2000)
78. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press, Cambridge (1998)
79. Swartout, W., Gratch, J., Hill, R.W., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Towards virtual humans. *AI Magazine* (27), 96–108 (2006)
80. Sycara, K.: Multiagent systems. *AI Magazine* 19(2), 79–92 (1998)
81. Urmson, C., Anhalt, J., Clark, M., Galatali, T., Gonzalez, J.P., Gowdy, J., Gutierrez, A., Harbaugh, S., Johnson-Roberson, M., Kato, H., Koon, P.L., Peterson, K., Smith, B.K., Spiker, S., Tryzelaar, E., Whittaker, W.R.L.: High speed navigation of unrehearsed terrain: Red team technology for grand challenge 2004. Technical Report CMU-RI-TR-04-37, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (June 2004)

82. Van Wezel, W., Jorna, R., Meystel, A.: Planning in Intelligent Systems: Aspects, Motivations, and Methods. John Wiley & Sons, Hoboken, New Jersey (2006)
83. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
84. Yao, Y.Y., Zhong, N., Liu, J., Ohsuga, S.: Web Intelligence (WI): Research challenges and trends in the new information age. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) WI 2001. LNCS (LNAI), vol. 2198, pp. 1–17. Springer, Heidelberg (2001)
85. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)
86. Zadeh, L.A.: Outline of a new approach to the analysis of complex system and decision processes. *IEEE Trans. on Systems, Man, and Cybernetics* 3, 28–44 (1973)
87. Zadeh, L.A.: Fuzzy sets and information granularity. In: Gupta, M., Ragade, R., Yager, R. (eds.) *Advances in Fuzzy Set Theory and Applications*, pp. 3–18. North-Holland, Amsterdam (1979)
88. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* 90, 111–127 (1997)
89. Zadeh, L.A.: From computing with numbers to computing with words – From manipulation of measurements to manipulation of perceptions. *IEEE Transactions on Circuits and Systems* 45, 105–119 (1999)
90. Zadeh, L.A.: A new direction in AI: Toward a computational theory of perceptions. *AI Magazine* 22(1), 73–84 (2001)
91. Zadeh, L.A.: Toward a generalized theory of uncertainty (GTU) - An outline. *Information Sciences* 171, 1–40 (2005)
92. Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies (iIT) from the Stand-Point of Web Intelligence (WI). *Communications of the ACM* 50(3), 89–94 (2007)
93. Zhong, N.: Impending Brain Informatics (BI) research from Web Intelligence (WI) perspective. *International Journal of Information Technology and Decision Making* 5(4), 713–727 (2006)
94. Zhong, N., Liu, J. (eds.): *Intelligent Technologies for Information Analysis*. Springer, Berlin (2004)
95. Zhong, N., Liu, J., Yao, Y.Y. (eds.): *Web Intelligence*. Springer, Berlin (2003)
96. Zhong, N., Liu, J., Yao, Y.Y.: In search of the Wisdom Web. *IEEE Computer* 35(11), 27–31 (2002)
97. Zhong, N., Yao, Y.Y., Liu, J., Ohsuga, S.: Web Intelligence: Research and Development. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) WI 2001. LNCS (LNAI), vol. 2198, Springer, Heidelberg (2001)
98. Zhong, N., Liu, J., Yao, Y.Y., Ohsuga, S.: Web Intelligence (WI). In: *Proc. the 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC 2000)*, pp. 469–470. IEEE Computer Society Press, Los Alamitos (2000)

Granular Computing: Modeling Human Thoughts in the Web by Polyhedron

Tsau Young (T. Y.) Lin and Mong-Hang Vo

Department of Computer Science
San Jose State University
San Jose, CA 95192-0249
Berkeley Initiative in Soft Computing
University of California, Berkeley
tylin@cs.sjsu.edu

Abstract. In this chapter, granular computing is used to process the Web information. The main result is that the human thinking inside a document set can be represented by a polyhedron; more general cases are discussed. A polyhedron is a subset of Euclidean space that supports a combinatorial structure, called simplicial complex. From the granular computing point of view, a simplicial complex is a special form of the granular model, in which the granular structure satisfies the closed condition (geometric concept). It is important to note that it is equivalent to the Apriori principle (data mining concept). A point in such a polyhedron represents a human thought. Each simplex represents a concept. A maximal simplex is a primitive concept. A connected component represents a complete concept. The totality of these concepts is called the basic knowledge.

1 Introduction

In this chapter, we will model the human thinking in Web documents. The backbone technology is the granular computing (GrC), the underlying intuition is the pragmatic view of knowledge theory.

The heart of GrC is the collection of granules that has been called granular structure. The intuition is that a granule represents a basic unit of knowledge or lack of knowledge (uncertainty). Each such basic unit has been called concept [10,11]. These intuitive concepts can be regarded as names of these granules. The collection of those names has been called the concept/knowledge/linguistic structure. We will use basic knowledge or knowledge structure here.

The main result of this chapter is that the human thinking, as a set of human thoughts, can be modeled by a pseudo-polyhedron. A polyhedron is a subset of Euclidean space that has a combinatorial structure, called simplicial complex [8].

The combinatorial structure is derived from the keywords and co-occurring keywordsets. In this paper, a keyword will be defined by TFIDF, and a co-occurring keywordset will be defined by certain variations of TFIDF. We will use U to denote the set of keywords, and β the collection of co-occurring keywordsets. The pair (U, β) is a granular model. By taking a geometric view, namely,

we interpret keywords as vertices, and then keywordsets naturally become simplexes. The consequence is the granular model becomes a simplicial complex; Thanks to the facts that the Apriori principle is the closed condition of simplicial complex. The simplicial complex naturally induces a polyhedron, this is our goal.

2 An Overview

In this section, we will explain the construction of the Polyhedron of concepts. A concept is carried by a simplex that represents a frequent co-occurring keywordset; recall the analogous term, frequent itemset in association rules.

The major steps are:

- Step 1.* Construct the simplicial complex of linear texts. The main task is to convert the knowledge representations of human thinking (authors' idea) by linear texts (documents) into a new representation by a simplicial-complex. Though the latter one is quite foreign to computer scientists, however, it is an old notion in combinatorial topology (cf. Section 4).
- Step 2.* Construct the language of concepts.
- Step 3.* Polyhedrons – the latent semantic model of the language.

We have an interesting corollary:

Corollary 1. *If A and B are the same document expressed in different languages, then their polyhedra are also the same.*

We may schematically summarize our approach as follows:

Note that $\Rightarrow$ denotes the “information flow”, which is not a mathematical mapping. The goal of this chapter is to have a mathematical model of authors' idea, so that the “information flow”, becomes a mathematical mapping:

$$\begin{array}{c}
 \left[\begin{array}{l} \textit{Human thoughts} \\ \textit{(Will be)Polyhedron} \end{array} \right] \Rightarrow \left[\begin{array}{l} \textit{linear} \\ \textit{text} \end{array} \right] \\
 \Downarrow (\text{summarize}) \\
 \left[\begin{array}{l} \textit{Human thoughts} \\ \textit{(Will be)Polyhedron} \end{array} \right] \Rightarrow \left[\begin{array}{l} \textit{Simplicial} \\ \textit{Complex} \end{array} \right] \\
 \Downarrow (\text{summarize}) \\
 \left[\textit{Human thoughts} \right] \Longleftrightarrow \left[\textit{Polyhedron} \right]
 \end{array}$$

2.1 Topology of Linear Texts

First we need to step the topology of linear text. Let us first recall the notion of Apriori condition [1].

Definition 1. *Apriori condition: Any q -subset of a n -keywordset is a q -keyword set for $q \leq n$.*

where a set containing exactly q elements is abbreviated as a q -subset. If we regard a keyword as a vertex, and a keywordset as a simplex, then the Apriori condition is the closed condition of the simplicial complex (cf. Section 4). With this observation, we have the following theorem:

Theorem 1. *The pair (V_{text}, S_{text}) is an Abstract Simplicial Complex, where*

1. V_{text} is the set of keywords, and is regarded as a set of abstract vertices, called keyword-vertices.
2. S_{text} is the set of keywordsets (associations) and is regarded as a set of abstract simplexes, called keyword-simplexes.

This simplicial complex is called a Keyword Simplicial Complex (KSC). Note that this KSC is a combinatorial structure that consists, except the vertices, of “invisible objects” hidden in the document set (cf. Section 6.3).

To see its strength, let us recall a striking theorem: Using the geometry, we can determine if two sets of documents written in different languages are similar, even without translation.

Corollary 2. *Let A and B be two document sets, where B is a translation of A into another language then the simplicial complexes of A and the simplicial complexes of B are isomorphic.*

3 Granular Computing and Models

What is Granular Computing? To date there is no well accepted formal definition yet. To get some traces of the intuitions, let us recall the event. In the fall 1996, Professor Lotfi Zadeh suggested granular mathematics to be my research area during my sabbatical leave at Berkeley. To limit the scope, the term granular computing (GrC) was used to label the area of research [27,28]. Therefore, initially, it was the computable portion of the granular mathematics (GrM). Informally, it is a new mathematics, in which the notion of “points” in the classical mathematics is replaced by that of granules. The well-known example is the notion of infinitesimals that led to the invention of calculus. During that time, the granules of infinitesimal were visualized as follows: Each real number was surrounded by a granule of infinitesimal quantities. The notion was not precise, each school had his own vague idea.

There are two formal solutions; one is the non-standard analysis in non-standard world. Intuitively speaking, each real number x is expanded to a granule

$N(x)$ that consists of $x +$ infinitesimals. The other one is the concept of topology. Therefore, roughly, an infinitesimal granule in non-standard world is equivalent to a topological neighborhood system in standard world.

Totally from other directions, in fact, approximate retrieval and reasoning, in 1988, we have generalized the topological neighborhood system to the neighborhood system (NS) [13], simply by dropping the requirements, the axioms of topology. The NS was the first mathematical model for GrC [10]. In fact, we have generalized it to two spaces, namely, a neighborhood system (NS), which is a mapping (function), $NS : V \longrightarrow 2^U$, that for each point p in V , we associate it with a family of subsets, $NS(p) \subseteq U$ of another universe U ; note this family can be an empty family at certain points. For simplicity, we define a granular model using single space.

Definition 2. A pair (U, β) , where U is a classical set and β is an NS, is called *Local Granular Model*.

If we restrict NS to be a singleton, in other words, to each point p in V , associated with a unique subset $B(p) \subseteq U$ (that can be empty), then, we have called this a binary neighborhood system (BNS). It can easily be shown that such a BNS induced a binary relation $B = \bigcup \{B(p) \mid p \in V\}$ [10]. Again, we will restrict ourselves to single space case.

Definition 3. A pair (U, β) , where U is a classical set and β is a BNS, is called *Binary Granular Model*.

The notion of topology can also be defined as a pair of objects that consists of one classical set and one family of open sets that satisfies topological axioms. Taking the notion of subbase, we define:

Definition 4. A pair (U, β) , where U is a classical set and β is a family of subsets, is a (global) granular model.

Classical rough setters call this (partial) covering.

Definition 5. A granular model, whose granular structure satisfies the closed condition, is called a simplicial complex.

In GrC, U is often referred to as a universe (of discourse) and β as the granular structure.

So far, these models are set based. From previous analysis, a neighborhood system (NS) is an infinitesimal granule. As a real valued function has been interpreted as a fuzzy set, we will go further for the notion of granules. For example, a fuzzy subset (a membership function), a function, a set function (measure, probability, belief functions), a generalized function (e.g., Dirac functions) and so forth are all possible candidates of granules.

Note that ‘how to solve it’ [18] has not been any part of formal mathematics, and hence not GrM. However, yet ‘how to compute it’ is an integral part of computing. So ‘granulate and conquer’ should be an integral part of GrC, hence a granule could be a subproblem, a subtask, a sub-Turing machine and etc. This

view introduced the problem of information integration theory. Intuitively, we have been saying that human has granulated daily things to subthings, such as human body to head, neck and etc. We will formulate in the most general abstract form.

Mathematicians formalize their domains of interests by Category Theory. For example, the domain for set theory is the Category of Sets. In this chapter, the language of category theory will be used to describe various universes of discourse for GrC. Roughly, we give the following definition:

Definition 6. *A category consist of*

- (1) *a class of objects, and*
- (2) *a set $Mor(X, Y)$ of morphisms for every ordered pair of objects X and Y , which satisfies certain properties. For this chapter, the formal details are not important; we only need the language loosely.*

Here are some examples:

1. The Category of Sets: The objects are classical sets. The morphisms are the maps.
2. The Category of Sets and Binary Relations: This category is very similar to this one. The objects $X, Y \dots$ are classical sets. However, the morphisms are the binary relations, or equivalently BNS. In the entity-relationship model, we use this category.
3. The Category of Power Sets: The object U_X is the power set of a classical set X . Let U_Y be another object, where Y is another classical set. The morphisms are the maps, $P(f) : U_X \longrightarrow U_Y$, which are induced by maps $f : X \longrightarrow Y$.

Let Γ be an abstract category, whose objects are Γ_i , where i runs through an index set.

Definition 7. *A pair (U, β) , where U is a family of Cartesian product of finite objects and β is a family of subjects of those Cartesian objects, is a (categorical) granular model.*

When U is a single object, this categorical granular model is formalized. The intuitive idea: thing is granulated into subthings. If U is a family of objects that are classical sets, this category defines the relational granular model. If U is a single set, this (categorical) granular model is a (global) granular model.

4 Combinatorial Topology

This section is purely for reference purposes. Let us introduce and define some basic notions in combinatorial topology. The central notion is n -simplex.

Definition 8. *A n -simplex is a set of independent abstract vertices $[v_0, \dots, v_{n+1}]$. A r -face of a n -simplex $[v_0, \dots, v_{n+1}]$ is a r -simplex $[v_{j_0}, \dots, v_{j_{r+1}}]$ whose vertices are a subset of $\{v_0, \dots, v_{n+1}\}$ with cardinality $r + 1$.*

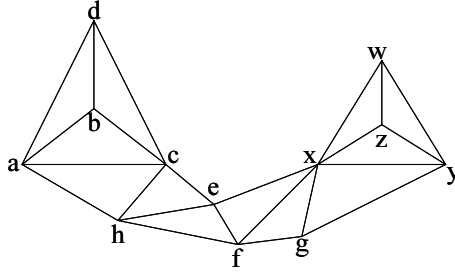


Fig. 1. A complex with twelve vertexes

Geometrically 0-simplex is a vertex; 1-simplex is an open segment (v_0, v_1) that does not include its end points; 2-simplex is an open triangle (v_0, v_1, v_2) that does not include its edges and vertices; 3-simplex is an open tetrahedron (v_0, v_1, v_2, v_3) that does not include all the boundaries.

Definition 9. A simplicial complex C is a finite set of simplexes that satisfies the following two conditions:

- Any set consisting of one vertex is a simplex.
- Any face of a simplex from a complex is also in this complex.

The vertices of the complex $v_0, v_1, \dots, v_n$ is the union of all vertices of those simplexes ([22], pp. 108).

If the maximal dimension of the constituting simplexes is n then the complex is called n -complex.

Note that, any set of $n + 1$ objects can be viewed as a set of abstract vertices, to stress this abstractness, some times we refer to such a simplex a combinatorial n -simplex. The corresponding notion of combinatorial n -complex can be defined by (combinatorial) r -simplexes.

In Figure 1, we have a simplicial complex that consists of twelve vertices that are organized in the form of 3-complex, denoted by S^3 .

Let us enumerate every simplex of S^3 :

1. The maximal 3-simplex $\Delta(a, b, c, d)$, and
 - (a) Its four 2-simplex faces $\Delta(a, b, c)$, $\Delta(a, b, d)$, $\Delta(a, c, d)$, $\Delta(b, c, d)$, and
 - (b) Its six 1-simplex faces $\Delta(a, b)$, $\Delta(a, c)$, $\Delta(a, d)$, $\Delta(b, c)$, $\Delta(b, d)$, $\Delta(c, d)$.
2. The maximal 3-simplex $\Delta(w, x, y, z)$, and
 - (a) Its four 2-simplex faces $\Delta(w, x, y)$, $\Delta(w, x, z)$, $\Delta(w, y, z)$, and $\Delta(x, y, z)$, and
 - (b) Its six 1-simplex faces $\Delta(w, x)$, $\Delta(w, y)$, $\Delta(w, z)$, $\Delta(x, y)$, $\Delta(x, z)$, $\Delta(y, z)$.
3. The maximal 2-simplexes lying between two 3-simplexes: $\Delta(a, c, h)$, $\Delta(c, h, e)$, $\Delta(e, h, f)$, $\Delta(e, f, x)$, $\Delta(f, g, x)$, $\Delta(g, x, y)$.
4. Some of their 1-simplex faces $\Delta(a, h)$, $\Delta(c, h)$, $\Delta(c, e)$, $\Delta(h, e)$, $\Delta(e, f)$, $\Delta(h, f)$, $\Delta(e, x)$, $\Delta(f, x)$, $\Delta(f, g)$, $\Delta(g, x)$, $\Delta(g, y)$; non of them are maximal.
5. Their 0-simplex faces (certices) $a, b, c, d, e, f, g, h, w, x, y, z$.

4.1 Combinatorial Topology of Pairs

Let the simplicial complex pair (X, Y) consists of a simplicial complex X and a subcomplex Y . A r -skeleton (denoted by S^r) of n -complex is a n -complex, in which all k -simplexes ($k \geq r$) have been removed. We often interested in the pair (S^n, S^r) , where $n \geq r$.

Two simplexes in a complex are said to be *directly connected* if the intersection of them is a nonempty face. Two simplexes in a complex are said to be *connected* if there is a finite sequence of directly connected simplexes connecting them. For any non-empty two simplexes A, B are said to be r -connected if there exists a sequence of k -simplexes $A = S_0, S_1, \dots, S_m = B$ such that S_j and S_{j+1} has an h -common face for $j = 0, 1, 2, \dots, m-1$; where $r \leq h \leq k \leq n$.

Let us consider the $(3, 1)$ -skeleton pair, in which X consists of all simplex of S^3 , and Y those of S^1 . The simplex in X , but not in Y are:

1. The maximal 3-simplex $\Delta(a, b, c, d)$, and
 - (a) Its four 2-simplex faces $\Delta(a, b, c)$, $\Delta(a, b, d)$, $\Delta(a, c, d)$, $\Delta(b, c, d)$, and
 - (b) Its six 1-simplex faces $\Delta(a, b)$, $\Delta(a, c)$, $\Delta(a, d)$, $\Delta(b, c)$, $\Delta(b, d)$, $\Delta(c, d)$.
2. The maximal 3-simplex $\Delta(w, x, y, z)$, and
 - (a) Its four 2-simplex faces $\Delta(w, x, y)$, $\Delta(w, x, z)$, $\Delta(w, y, z)$, and $\Delta(x, y, z)$, and
 - (b) Its six 1-simplex faces $\Delta(w, x)$, $\Delta(w, y)$, $\Delta(w, z)$, $\Delta(x, y)$, $\Delta(x, z)$, $\Delta(y, z)$.
3. The maximal 2-simplexes lying between two 3-simplexes: $\Delta(a, c, h)$, $\Delta(c, h, e)$, $\Delta(e, h, f)$, $\Delta(e, f, x)$, $\Delta(f, g, x)$, $\Delta(g, x, y)$.
4. Some of their 1-simplex faces $\Delta(a, h)$, $\Delta(c, h)$, $\Delta(c, e)$, $\Delta(h, e)$, $\Delta(e, f)$, $\Delta(h, f)$, $\Delta(e, x)$, $\Delta(f, x)$, $\Delta(f, g)$, $\Delta(g, x)$, $\Delta(g, y)$; non of them are maximal.

For simplicity, we explain the simplicial complexes. For Web clustering we need a simplicial complex pair (X, Y) . A pair consists of a simplicial complex X and a subcomplex Y . We will refer to them as a simplicial complex pair. A r -skeleton (denoted by S^r) of n -complex is a n -complex, in which all k -simplexes ($k \geq r$) have been removed. We are often interested in the pair (S^n, S^r) , where $n \geq r$.

A simplex δ is said to be in a simplicial complex pair (X, Y) means δ is a simplex in X , but not in Y . Two simplexes are said to be *directly connected* if the intersection of them is a nonempty face. Two simplexes in (X, Y) are said to be *connected* if there is a finite sequence of directly connected simplexes connecting them.

Let us consider the $(3, 1)$ -skeleton pair, in which X consists of all simplex of S^3 , and Y those of S^1 . The simplex in X , but not in Y is:

1. The maximal 3-simplex $\Delta(a, b, c, d)$, and
 - (a) its four 2-simplex faces $\Delta(a, b, c)$, $\Delta(a, b, d)$, $\Delta(a, c, d)$, $\Delta(b, c, d)$, and
 - (b) its six 1-simplex faces $\Delta(a, b)$, $\Delta(a, c)$, $\Delta(a, d)$, $\Delta(b, c)$, $\Delta(b, d)$, $\Delta(c, d)$.

2. The maximal 3-simplex $\Delta(w, x, y, z)$, and
 - (a) its four 2-simplex faces $\Delta(w, x, y)$, $\Delta(w, x, z)$, $\Delta(w, y, z)$, and $\Delta(x, y, z)$, and
 - (b) its six 1-simplex faces $\Delta(w, x)$, $\Delta(w, y)$, $\Delta(w, z)$, $\Delta(x, y)$, $\Delta(x, z)$, $\Delta(y, z)$.
3. The maximal 2-simplexes lying between two 3-simplexes: $\Delta(a, c, h)$, $\Delta(c, h, e)$, $\Delta(e, h, f)$, $\Delta(e, f, x)$, $\Delta(f, g, x)$, $\Delta(g, x, y)$, and
4. some of their 1-simplex faces $\Delta(a, h)$, $\Delta(c, h)$, $\Delta(c, e)$, $\Delta(h, e)$, $\Delta(e, f)$, $\Delta(h, f)$, $\Delta(e, x)$, $\Delta(f, x)$, $\Delta(f, g)$, $\Delta(g, x)$, $\Delta(g, y)$; non of them are maximal.

Let us consider pair (A, B) , in which A consists of all simplex of S^3 , and B is the closed maximal 2-simplexes lying between two 3-simplexes. Geometrically, the pair consists of two closed tetrahedrons with two missing closed edges; the two edges connected to Y .

1. The maximal 3-simplex $\Delta(a, b, c, d)$, and
 - (a) its four 2-simplex faces $\Delta(a, b, c)$, $\Delta(a, b, d)$, $\Delta(a, c, d)$, $\Delta(b, c, d)$, and
 - (b) among its six 1-simplex faces $\Delta(a, c)$ will be missing.
2. The maximal 3-simplex $\Delta(w, x, y, z)$, and
 - (a) its four 2-simplex faces $\Delta(w, x, y)$, $\Delta(w, x, z)$, $\Delta(w, y, z)$, $\Delta(x, y, z)$, and
 - (b) among its six 1-simplex faces $\Delta(x, y)$.
3. Its 0-simplex faces $\Delta(w)$, $\Delta(z)$, $\Delta(b)$, $\Delta(d)$.

5 Keywords and TFIDF

Though the best way to select a set of keywords for representing a document is manually, there can be an infinite number of documents in a set, therefore, automation is necessary. We will use TFIDF [19,20] and SUPPORT as the measure of the significance of tokens. Briefly, a token is regarded as a keyword if it has high TFIDF and SUPPORT values.

Definition 10. Let T_r denote the total number of documents in the collection. The significance of a token t_i in a document d_j in T_r is its TFIDF value. It is calculated as

$$\text{TFIDF}(t_i, d_j) = \text{tf}(t_i, d_j) \log \frac{T_r}{\text{df}(t_i)}$$

where $\text{df}(t_i)$ denotes the number of documents in T_r in which t_i occurs at least once, and

$$\text{tf}(t_i, d_j) = \begin{cases} 1 + \log(N(t_i, d_j)) & \text{if } N(t_i, d_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $N(t_i, d_j)$ denotes the frequency of token t_i occurs in document d_j by counting all its nonstop words.

Roughly, the TFIDF is $\text{tf} \times \text{idf}$ indexing [19,20], where tf denotes token frequency as it appears in the document and idf denotes the inverse document frequency where document frequency (df) is the number of distinct documents that contain the token. It follows that stop words will have value zero, since the log value of $T_r/\text{df}(\text{stop word}) = 1$ is 0.

Note that TFIDF only reflects the importance of a token in one particular document; it does not measure its overall significance in the set of documents. $\text{df}(t_i)$ is at its highest value if the token appears in only one document. the SUPPORT is introduced to reflect the notion of “frequency”.

Definition 11. *The SUPPORT of a keyword or a keywordset in a document set is the number of distinct documents that contain keywords and keywordset within d tokens respectively. We say that the SUPPORT is high if it is greater than a threshold.*

6 The Simplicial Structure in Documents

In this section, we build the Simplicial Complex Structure for Documents. The major step is to convert linear-text-representations (documents) into a simplicial-complex-representation. The schematic diagram has been shown in Section 2.

6.1 Association (Rule) Mining

A formal definition of association rule can be found in [1]. The two measurements in question are SUPPORT and CONFIDENCE. In this paper, we only use the SUPPORT value. The SUPPORT of a set of a nearby keywords is the number of documents in the document set that contain the keywordsets. An association is a set of keywords $K = \{k_1, k_2, \dots, k_n\}$ in a set of document $D = \{d_1, d_2, \dots, d_i\}$ where $d_i = \{k_{i1}, k_{i2}, \dots, k_{ij}\}$ and $k_{ij} \in K$, which SUPPORT is greater than a threshold value. Such an association is also called a *frequent itemset* in the literature. Our primary interest is in the interactions among the keywordsets.

6.2 Topology of Linear Texts

Let us first recall the notion of Apriori condition [1].

Definition 12. *Apriori condition: Any q -subset of a n -keywordset is a q -keyword set for $q \leq n$.*

where a set containing exactly q elements is abbreviated as a q -subset. If we regard a keyword as a vertex, and a keywordset as a simplex, then the Apriori condition is the closed condition of the simplicial complex (cf. Section 4). With this observation, we have the following theorem:

Theorem 2. *The pair $(V_{\text{text}}, S_{\text{text}})$ is an Abstract Simplicial Complex, where*

1. V_{text} is the set of keywords, and is regarded as a set of abstract vertices, called keyword-vertices.

2. S_{text} is the set of keywordsets (associations) and is regarded as a set of abstract simplexes, called keyword-simplexes.

This simplicial complex is called a Keyword Simplicial Complex (KSC). Note that this KSC is a combinatorial structure that consists, except the vertices, of “invisible objects” hidden in the document set (cf. Section 6.3).

To see its strength, let us recall a striking theorem: Using the geometry, we can determine if two sets of documents written in different languages are similar, even without translation.

Corollary 3. *Let A and B be two document sets, where B is a translation of A into another language then the simplicial complexes of A and the simplicial complexes of B are isomorphic.*

6.3 Capturing the Hidden Semantics

Let us start intuitively. Each document describes some idea in author’s mind, which may consist of many levels and wide ranges of concepts. We will explore some of it through high frequency keywords and keywordsets. Following the terminology in association mining, we treat an association and a keywordset as synonyms.

Roughly, association is stressed more on the semantics of a keywordset, while the latter one is a phenomena of the document set. They will be used interchangeably throughout this chapter. First, let us recall some examples [8].

1. The keywordset “Wall Street” represents a concept that has nothing to do with “Wall” and “Street”.
2. The keywordset “White House” represents an object that has very little to do with “White” and “House”.

These examples indicate that the strength of this approach is the ability to capture the notion that is defined implicitly by the keywordset, in plain words, the capability of “reading between lines”.

7 Polyhedrons - The Latent Semantic Model

7.1 The Language of Concepts in Documents

The simplicial complex is a structure of concepts in the document set. For concept analysis, we need a set of symbols to various level of concepts such as keywords, keywordsets (associations), human idea, and all the “interactions” that are invisible in the text.

- L1)** IDEA is a notation to denote the totality of common thoughts or concepts among all authors of the documents under consideration; it is the universe of discourse.

L2) THOUGHT is a notation to denote some important basic ingredient, called element, in IDEA.

Please note that IDEA, at this point, is not a mathematical object yet. Instead, we merely view it as an element of an informal language. Human THOUGHT is not a mathematical concept either. The goal here is to establish IDEA as a set of THOUGHTs. Therefore, both are mathematical objects. Loosely speaking, the main goal of this chapter is to discover/construct a reasonable mathematical model for IDEA. Therefore, the language (a set of notations) will be interpreted into a mathematical model. This model will be the base of our concept analysis.

L3) B-concept(k), read as a basic concept, is a notation to denote the MEANING of the keyword k .

L4) I-concept(Δ), read as an intermediate concept, is a notation to denote the MEANING of a q -simplex Δ .

L5) P-concept, read as a primitive concept, is a notation to denote the I-concept of a maximal simplex.

L6) C-concept, read as a complete concept, is a notation to denote the MEANING of a connected component.

Thus, we have sufficient notations to conduct concept analysis. However, we do not have adequate facilities in the language, such as a deductive system, etc. to reason about these concepts and understand the interrelationships among them. Therefore, we will set up, in Tarski's style semantics, namely, a mathematical model to interpret this language.

7.2 Polyhedrons

We may map V_{text} as the unit points of an Euclidean space (cf. Section 4). By doing so, each keywordset in S_{text} can be interpreted as an open simplex in an Euclidean space [22]. The union of all these open simplexes is the Polyhedron $|S_{text}|$ of the linear texts.

The Polyhedron is a closed set (due to the closed condition) in the Euclidean space. Next, we will introduce an important notion: star(Δ) [22]. We will denote it by $S^*(\Delta)$.

Definition 13. $S^*(\Delta) = \cup\{ \text{all open simplexes that have } \Delta \text{ as a face} \}$.

In the case when $\Delta = k$ is a keyword, $S^*(k)$ is an open set in $|S_{text}|$. It is the minimal open neighborhood of k that is definable as simplexes only. It is referred to as the basic neighborhood or basic granule of k . Consider it as the first building block of our simplicial complex. For example, we have

Proposition 1. $S^*(\Delta) = \cap\{S^*(k_i) \mid k_i \in \Delta\}$.
 $S^*(\Delta_{max}) = \Delta + max$ for a maximal simplex Δ_{max} .

7.3 Latent Semantic Model

Now, we are ready to introduce the model: Let $P = |S_{text}|$ be the polyhedron of KSC. The Latent Semantic Model consists of a KSC and its Polyhedron:

- The IDEA is modeled by the polyhedron P .
- A human THOUGHT is an element in the IDEA. P is referred to as the Latent Semantic Space.
- The mapping $S^* : KSC \longrightarrow P$ realizes the interpretations, for example, a B-concept is mapped to an open set.

With this interpretation, all language elements, such as concepts, are granules (subsets) in P . In particular,

- B-concept(k) = $S^*(k)$.
- I-concept(Δ) = $S^*(\Delta) = \cap \{S^*(k_i) \mid k_i \in \Delta\}$.
- P-concept is the I-concept of a maximal simplex.
- C-concept is a connected component.

It should be clear that the set of I-concepts forms a semigroup generated by $S^*(k)$ via intersections. This semigroup is a partial order set obtained using the “inclusion” of the set theory. On the other hand, the I-concept(Δ) is also a partial order set, but this time, due to the face-relation of simplexes Δ .

Proposition 2. *The collection of I-concepts has two partial orderings: One by KSC, and another by the set inclusion in the polyhedron. Of importance, these two partial orders do agree.*

We leave it to the reader to verify that the required properties are verified by these sets. This collection is indexed by the KSC. We will call it the Knowledge Complex and denote it by β . It is far more complex than a concept hierarchy. In particular, the pair (P, β) , where $\beta = \text{KSC}$, is a granular model in the sense of [6,7].

8 High Dimensions (HD)

8.1 An Overview

It came to our attention early on that the notion of TFIDF as described in Section 5 can naturally be extended to high dimension tokens. In this section, we conduct a case study on the use of several relevance ranking methods. Latent Semantic Indexing (LSI) addresses relevance ranking based on how significant a search word is in each document. Traditionally, the LSI is one-dimensional. The term frequency-inverse document frequency (TFIDF) is calculated based on how significant a single word is in a document. A common way to compute TFIDF is to use the following formula based on the works by Salton and McGill [19].

Definition 14. $TFIDF(term_i, document_j) = tf(t_i; d_j) * \log|Tr|/|Tr(t_i)|$

where $Tr(t_i)$ is the number of documents in Tr in which t_i occurs at least once, and $tf(t_i; d_j) = 1 + \log(N(t_i; d_j))$ if $N(t_i; d_j) > 0$ ($= 0$ otherwise), where $N(t_i, d_j)$ is the number of times that token t_i occurs in document d_j .

However, since queries often consist of more than one word, it is reasonable to question whether an one-dimensional LSI is always adequate. For example, the query “wall street” consists of two words, and the semantics of the query extend beyond the concepts of the individual words “wall” and “street”. The authors explored four methods of generalizing LSI to higher dimensions. The notation LSI n is a short hand for n -dimensional LSI. Because a major goal of LSI is to produce a good relevance ranking of search results, the authors examine the precision (rather than the recall) in each method explored.

8.2 Three Factors Examined

Among the four methods of LSI generalization studied, we examine three factors that can potentially affect precision. These factors are:

- the threshold for filtering data from one dimension to a higher dimension.
- the normalization technique for the co-occurrence matrix $N(t_i, d_j)$.
- the effect of stop-word removal.

Threshold. To understand the need for a threshold for filtering data from single to higher dimension, let us look at the computation of two-dimensional TFIDF. The goal is to find out how significant is each pair of tokens in each document. In theory, the input is the self product of the inverted table. From a reasonable corpus [23], the authors generated an inverted table of about 5 million rows. The Cartesian product would have yielded more than 2.5×10^{13} rows. The resource that is required to process such a large amount of data is beyond the scope of this study, but there are several ways to reduce the size of the input to multi-dimensional TFIDF computation, and the authors examined two of them. One way to limit the input size is to use one-dimensional TFIDF as a threshold. The size of the inverted table is reduced by removing the terms whose TFIDF values are lower than a certain threshold. Then, the Cartesian product of the reduced inverted table is used as the input of the multi-dimensional TFIDF computation. Another way to limit the input size is to use the document frequency (DF) as the threshold. The DF of a term is the number of documents in which the term appears at least once.

Normalization. When TFIDF is computed using the formula above, long documents (those with many tokens) usually get higher values because TF s are high. A way to normalize $N(t_i, d_j)$ is to divide it by the total number of all types of tokens in document d_j . For TFIDF to meaningfully indicate the significance of a term in a document, both the TF and the IDF parts must have the same sign. For example, if TF was negative and IDF was positive, the product of TF and IDF would be meaningless. Since IDF is always non-negative, TF should

be made non-negative too. With the denominator introduced in the calculation of $N(t_i, d_j)$, TF (being $1 + \log(N(t_i, d_j))$) may be negative in some case. To solve this problem, a large enough constant coefficient is also introduced in the calculation of $N(t_i, d_j)$. In this case, a coefficient of 40,000 is valid because it is larger than the total number of tokens of the longest document.

Stopword Removal. In a given document set, some tokens cannot be used to distinguish a document over another using a ranking algorithm. Those words that carry a vague, general meaning provide little or no help in ranking, and are called *stopwords*. For the purpose of this case study, stopwords are specified in a stopwords list, which is a text file downloaded from the WordNet website [24]. The purpose of removing stopwords is to reduce the size of the inverted table for LSI1 computation. Stopwords, such as “a” and “the”, often occur frequently in many documents, but they rarely contribute to the content or the meaning. Therefore, it is often believed that removing stopwords does not affect significantly the result of statistical content analysis while gaining some performance in the execution time. The authors chose the stop analyzer of Lucene for removing stopwords.

8.3 Overview of Four Methods of HD TF-IDF Computation

The three factors form an algorithm design space, within which the authors chose four points to study the dimensional trends in terms of precision. Table 1 shows the positions of the four chosen points in this study.

In the case when the co-occurrence matrix $N(t_i, d_j)$ is not normalized, it is the number of times that token t_i occurs in document d_j . Each of methods was implemented in Java. Each implementation processed the entire collection of 16,330 documents from University of California, Irvine Knowledge Discovery in Databases (UCI-KDD) in each case [23].

Table 1. Four Methods of HD TF-IDF Computation

Methods	Threshold	Co-occurrence Matrix	Stopwords
Method 1	TFIDF	Not Normalized	Included
Method 2	TFIDF	Normalized	Included
Method 3	DF	Normalized	Included
Method 4	TFIDF	Normalized	Excluded

8.4 Software Architecture

Figure 2 shows the software architecture of the Java implementation that consists of the following major components:

- The document preprocessor.
- The computation unit of LSI for one keyword (LSI1).
- The computation unit of LSI for two keywords (LSI2).

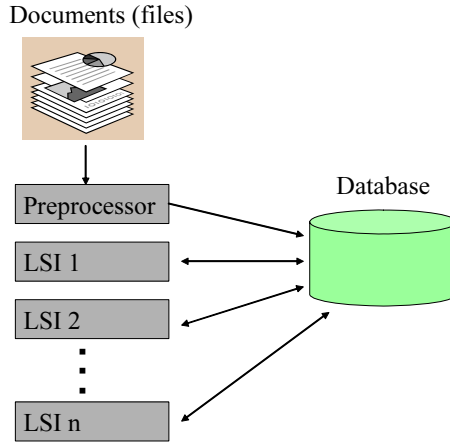


Fig. 2. Overview (I) of the Design

- The computation units of higher dimensional LSI for n keywords (LSIn) work similarly.

The object-oriented framework is designed to facilitate the extension of LSI to any dimension. The authors have implemented the computation units up to dimension four in this study.

The document preprocessor reads the entire set of documents from the file system and stores the resulting inverted files as tables in a database. LSI1 takes the inverted tables from the database to calculate the LSI and stores the LSI into a different table in the same database. LSI2 takes as its input the Cartesian product of the inverted tables filtered with LSI to reduce its size. It then computes the LSI for each pair of terms and stores them into a different table. LSI3 and higher dimensional LSI modules work in a way similar to LSI2.

Document Preprocessor. The document preprocessor reads text documents from a file system. It preprocesses each document by extending the analyzer of a popular open-source software package, Apache Lucene. The analysis involved in document preprocessing consists of the following steps:

1. Tokenization is the preprocessing step for dividing a document into terms or words. The step is performed by the LetterTokenizer of Lucene, which uses the Java built-in method `Character.isLetter(char)` to determine whether a character is a letter or not. Any non-letter character is regarded as a separator between terms.
2. Canonicalization is the preprocessing step for reducing different forms of the same term into a single representation for accurate comparison between terms. Canonicalization involves lowercasing and stemming.
 - (a). Lowercasing converts each letter of each term into lower case.
 - (b). Stemming removes inflectional morphemes from each term. For example, “apples” is converted to the same representation as “apple” so that

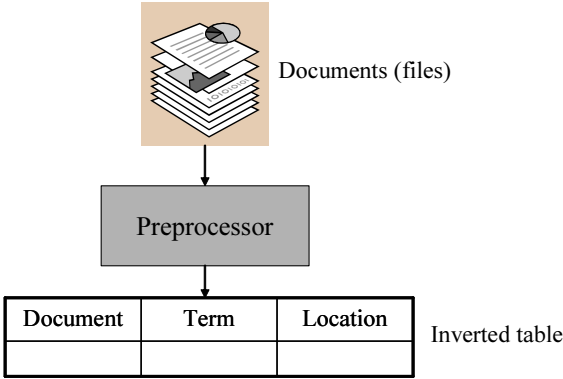


Fig. 3. Overview (II) of the Design

the system regards both “apples” and “apple” as being the same term regardless of whether the term is in its plural or singular form. Lucene uses Porter’s algorithm for stemming.

For each term in each document, as shown in Figure 3, the document preprocessor canonicalizes the term and associates it with the document in which the term appears by inserting a row into inverted tables in an Oracle database. In the same row, the document preprocessor also records the location or position in which the term appears in the document.

8.5 Method 1

In Method 1, the TFIDF is used as a threshold for limiting the size of the input of the multi-dimensional computation. No normalization is performed on the co-occurrence matrix. No stopword is removed from the input. As shown in Figure 4, Method 1 appears to produce very good precision, especially in higher dimensions. Moreover, its preciseness appears to be independent of the dimension. The method performs well no matter whether short documents are included in the analysis or not. A potential drawback of Method 1 is that it seems to favor long documents. This is not an issue if the long document is full of content in the form of unstructured text because the latent semantic indexing works especially well when it has enough content to “crunch”.

This long-document effect becomes an issue only when the document is somewhat “content-less”. In the UCI-KDD corpus, a common example of such a “content-less” long document is a uuencoded binary file. The corpus consists of Usenet newsgroup articles from the 80s and early 90s, during which it was a common practice to post uuencoded binary files on the Internet. Although these files have many tokens when tokenized by Lucene, the LSI algorithm performs poorly in capturing their semantics or contents. Because (by the design of uuencoding scheme) uuencoded binary files appear to be ASCII to any algorithm, it poses a challenge to remove them from the corpus automatically. Fortunately,

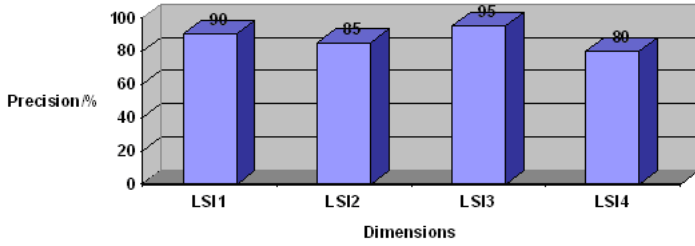


Fig. 4. Dimensional Trends of Method 1

most tokens produced from uuencoded files are four-character or shorter. The authors use this four-character threshold to produce meaningful results from this experiment practically reducing the effect from uuencoded files. While uuencoding has become less popular nowadays, another type of long documents will likely to continue to pose a challenge to LSI: *semi-structured documents*, which contain many repeated fields (name-value pairs with different values) embedded in ASCII files in non-standard ways.

An example is as follows: *Prices: \$75.00; Latest Version: 1.5.3.*

In this example, because the combination “price latest version” happens to occur many times in the documents, the TFIDF ends up to be very high from the LSI3 algorithm. However, “price latest version” is not a significant phrase in the document. The document was a list of platform-specific *X* servers. It remains a challenge to identify semi-structured documents or to improve the precision of LSI on them.

8.6 Method 2

In Method 2, the TFIDF is used as a threshold for limiting the size of the input to multi-dimensional computation. Normalization is performed on the co-occurrence matrix. No stopword is removed from the input.

As shown in Figure 5, in LSI1 and LSI2, the precision of the algorithm is impaired by the fact that the introduction of the denominator in the calculation of $N(ti, dj)$ favors short documents. Nevertheless, an interesting observation is that as the dimension goes higher than the previous method, reducing the adverse short-document effect. In shorter documents, the algorithm is more likely to be misled by tokens in email addresses and email header paths, which do not usually contribute to the main content of the document.

The results presented here came from the analysis of the long documents exclusively (those with more than 200 tokens).

8.7 Method 3

In Method 3, the DF is used as a threshold for limiting the size of the input to multi-dimensional computation. Normalization is performed on the

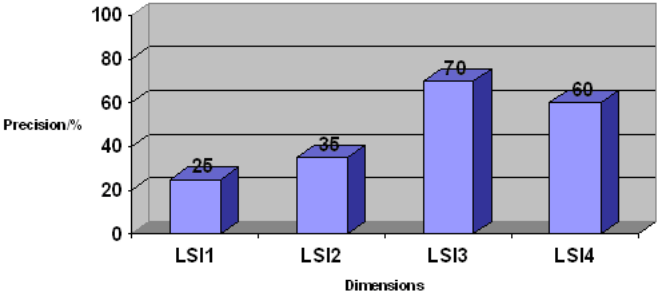


Fig. 5. Dimensional Trends of Method 2

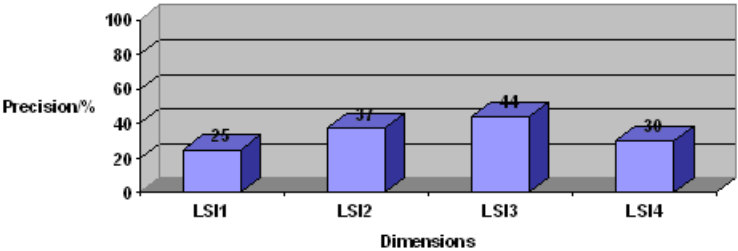


Fig. 6. Dimensional Trends of Method 3

co-occurrence matrix. No stopword is removed from the input. Since it is not possible to put a threshold to reduce the size of the input of the LSI1 calculation, Method 3 is the same as Method 2 for LSI1. From LSI2, the DF is used as the threshold instead of TFIDF for the purpose of reducing the size of the input to the algorithms. As shown in Figure 6, the precision does not seem to improve in the multidimensional cases.

Among the many cases in which the algorithms made mistakes in the multidimensional cases, most multi-token phrases are pertinent to the authors, such as their name, email, and organization. These phrases are usually not significant in their documents in terms of content. The DF is a good indicator of how prolific an individual or an organization is because the more documents that they appear in, the higher the value of DF. However, the TFIDF is a better indicator of the significance of a phrase in the documents where it occurs. This explains why TFIDF may be a better threshold than DF for reducing the size of the input in the multidimensional cases, as we compare the precision results of Method 3 with those of Methods 1 and 2.

8.8 Method 4

In Method 4, the TFIDF is used as a threshold for limiting the size of the input to multi-dimensional computation. Normalization is performed on the co-occurrence matrix. Stopwords are removed from the input.

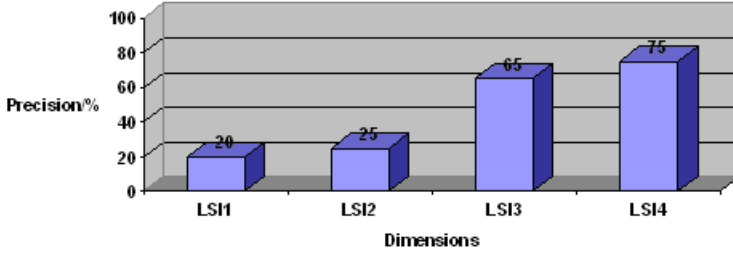


Fig. 7. Dimensional Trends of Method 4

As shown in Figure 7, the results are similar to those from Method 2. The removal of stopwords does not significantly affect the precision of LSI.

8.9 Summary of Four Methods

From the results of a straightforward computation of latent semantic indexing (in Method 1), the authors discovered that the prospect of useful and meaningful extension of LSI to higher dimensions is promising. A challenge is posed by long documents whose content cannot be captured by the LSI algorithm. An important example is the case of semi-structured documents. To explore the possibility and practicality of normalizing the LSI computation against the length of documents, the authors explored that idea of introducing the total number of tokens in a document as the denominator when calculating $N(t_i, d_j)$ (Method 2). Although the results were disappointing for one and two dimensions, some prospect was shown in higher dimensions. Method 4 is a variant of Method 2 in which stopwords were removed during document preprocessing. The effect on precision is not significant. The authors also explored the method of using document frequency (DF) instead of TFIDF as a threshold to limit the size of the input to HD-LSI computation. The precision gets worse as the number of dimensions gets higher. This is probably because TFIDF is a better significance indicator than DF alone. The central question is whether to extend LSI to higher dimensions. The analysis of the research reveals the strengths and weaknesses of each approach to make the computation of HD-LSI tractable.

9 Conclusions

In this chapter, we focus on simplicial complex, which is a special form of the (set based) granular model. This is because, we use frequency as the criterion for keywordset. If we use a generalized TFIDF, the set of simplexes may not have the closed condition. Therefore, the resulting model is a non-closed subset of the simplicial complex. In this case, simplicial complexes are too restrictive, we need the full blown set based a granular model.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. 20th VLDB Conference (1994)
2. Bairamian, S.: Goal Search in Relational Databases, Thesis, California State University at Northridge (1989)
3. Brewer, D.C., Nash, M.J.: The Chinese Wall Security Policy. In: IEEE Symposium on Security and Privacy, Oakland, pp. 206–214 (May 1988)
4. Chu, W., Chen, Q.: Neighborhood and Associative Query Answering. *Journal of Intelligent Information Systems* 1, 355–382 (1992)
5. Dubois, D., Prade, H.: Putting Rough Sets and Fuzzy Sets Together. In: Slowinski, R. (ed.) *Decision Support by Experience - Application of the Rough Sets Theory*, pp. 203–232. Kluwer Academic Publishers, Dordrecht (1992)
6. Lin, T.Y.: Granular Computing II: Infrastructures for AI-Engineering. In: *Proceedings of 2006 International Conference on Granular Computing*, May 10–12, Atlanta, Georgia, USA (2006)
7. Lin, T.Y.: Granular Computing: Examples, Intuitions and Modeling. In: *Proceedings of 2005 IEEE International Conference on Granular Computing*, July 25–27, Beijing China, pp. 40–44 (2005)
8. Lin, T.Y., Chiang, I.-J.: A Simplicial Complex, a Hypergraph, Structure in the Latent Semantic Space of Document Clustering. *International Journal of Approximate Reasoning* 44, 55–80 (2005)
9. Lin, T.Y.: Granular Computing: Structures, Representations, Applications and Future Directions. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) *RSFDGrC 2003. LNCS (LNAI)*, vol. 2639, pp. 16–24. Springer, Heidelberg (2003)
10. Lin, T.Y.: Granular Computing on Binary Relations I: Data Mining and Neighborhood Systems. In: Skowron, A., Polkowski, L. (eds.) *Rough Sets In Knowledge Discovery*, pp. 107–120. Springer, Heidelberg (1998)
11. Lin, T.Y.: Granular Computing on Binary Relations II: Rough Set Representations and Belief Functions. In: Skowron, A., Polkowski, L. (eds.) *Rough Sets In Knowledge Discovery*, pp. 121–140. Springer, Heidelberg (1998)
12. Lin, T.Y.: Topological and Fuzzy Rough Sets. In: Slowinski, R. (ed.) *Decision Support by Experience - Application of the Rough Sets Theory*, pp. 287–304. Kluwer Academic Publishers, Dordrecht (1992)
13. Lin, T.Y., Huang, K.J., Liu, Q., Chen, W.: Rough Sets, Neighborhood Systems and Approximation. In: *Proceedings of the Fifth International Symposium on Methodologies of Intelligent Systems*, October 25–27, Knoxville, Tennessee, pp. 130–141 (1990)
14. Lin, T.Y.: Chinese Wall Security Policy – An Aggressive Model. In: *Proceedings of the Fifth Aerospace Computer Security Application Conference*, Tuscon, Arizona, pp. 282–289 (December 1989)
15. Lin, T.Y.: Neighborhood Systems and Relational Database. In: *Proceedings of 1988 ACM Sixteen Annual Computer Science Conference*, 725 (abstract) (1988)
16. Munkres, J.: *Topology*, 2nd edn. Prentice-Hall, Englewood Cliffs (2000)
17. Pawlak, Z.: *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht (1991)
18. Polya, G.: *How to Solve It*, 2nd edn. Princeton University Press (1957)
19. Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York (1986)

20. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5), 513–523 (1988)
21. Sierpinski, W., Krieger, C.: *General Topology*. University of Toronto Press (1952)
22. Spanier, E.: *Algebraic Topology*. McGraw-Hill, New York (1966)
23. University of California, Irvine, Knowledge Discovery in Databases Archive. [WWW Document] (2006), <http://kdd.ics.uci.edu/>
24. WordNet Similarity 1.02 Stoplist. [WWW Document] (visited August 10, 2005), <http://search.cpan.org/src/TPEDERSE/WordNet-Similarity-1.02/samples/stoplist.txt>
25. Zadeh, L.A.: Fuzzy Sets and Information Granularity. In: Gupta, M., Ragade, R., Yager, R. (eds.) *Advances in Fuzzy Set Theory and Applications*, pp. 3–18. North-Holland, Amsterdam (1979)
26. Zadeh, L.A.: Towards a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic. *Fuzzy Sets and Systems* 19, 111–127 (1997)
27. Zadeh, L.A.: Some Reflections on Soft Computing, Granular Computing and Their Roles in the Conception, Design and Utilization of Information/Intelligent Systems. *Soft Computing* 2, 23–25 (1998)
28. Zadeh, L.A.: Some Reflections on Information Granulation and Its Centrality in Granular Computing, Computing with Words, the Computational Theory of Perceptions and Precisiated Natural Language. In: Lin, T.Y., Yao, Y.Y., Zadeh, L.A. (eds.) *Data Mining, Rough Sets, and Granular Computing*, pp. 3–20 (2002)

Biophysical Models of Neural Computation: Max and Tuning Circuits

Ulf Knoblich, Jake Bouvrie, and Tomaso Poggio

McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract. Pooling under a softmax operation and Gaussian-like tuning in the form of a normalized dot-product were proposed as the key operations in a recent model of object recognition in the ventral stream of visual cortex. We investigate how these two operations might be implemented by plausible circuits of a few hundred neurons in cortex. We consider two different sets of circuits whose different properties may correspond to the conditions in visual and barrel cortices, respectively. They constitute a plausibility proof that stringent timing and accuracy constraints imposed by the neuroscience of object recognition can be satisfied with standard spiking and synaptic mechanisms. We provide simulations illustrating the performance of the circuits, and discuss the relevance of our work to neurophysiology as well as what bearing it may have on the search for maximum and tuning circuits in cortex.

1 Introduction

A recent theory of object recognition in the feed-forward pathway of the ventral stream in primate visual cortex [1, 2] is based on a hierarchical model with two main operations iterated several times throughout the hierarchy. The two basic operations are a bell-shaped tuning mechanism to provide selectivity of units – learned in an unsupervised way – along the pathway and a maximum-like operation to achieve invariance for position, scale and clutter while maintaining selectivity. To support the biological plausibility of the theory it is important to show that these two basic operations can be implemented using well-established, plausible properties of neurons and synapses. The primary goal here is to describe realistic circuits for the maximum and tuning operations utilized by this model. There are several possible circuit designs that can theoretically provide the level of robustness required by model simulations, and yet, little is known about how such circuits might be implemented in the brain. The circuits described herein therefore represent a plausibility proof and may also be seen as specific hypotheses that neurophysiologists may try to prove or disprove using, for example, extracellular and intracellular recordings.

Several groups have reported neurons in visual cortex that respond rapidly and sub-linearly to the combined presentation of two simple stimuli in their receptive field [3, 4], i.e. the response to the combined stimulus is significantly smaller than the sum of the responses to the single stimuli. It has been proposed that, instead

of summing the inputs, these neurons compute either the maximum or the average of the inputs. Normalization circuits were suggested for gain control in [5] and for the biased competition model in [6] (see also [7, 8] and for older work on related dynamics of shunting inhibition [9]). Another possible mechanism for gain control relies on synaptic depression [10]. Several possible circuits for computing the maximum operation have been proposed on an abstract level [7, 11], but were not investigated with respect to their biophysical plausibility. A recent study presented a framework to unify both computations, softmax and tuning, providing a single equation that will produce either of the two behaviors depending on a small set of parameters [12]. The circuits presented here share some of the general architectural features presented in the above research. Additionally, we emphasize biological plausibility to a novel extent.

First, we introduce quite restrictive timing constraints on the circuits performing both computations imposed by physiological data and discuss how and where such constraints arise. We then present two different coding schemes in different sensory cortices and their implications, leading to the development of two sets of circuit models matched to the conditions prevalent in the respective areas. In section 2 we describe the architecture, dynamics, parameters and results of the Population Packet Code models, followed by the Poisson Code models in section 3 and a discussion of the results and their relevance.

1.1 Time, Computational, and Bandwidth Constraints

A plausible circuit must satisfy several constraints imposed by the physiological mechanisms underlying synaptic transmission: we must perform the above computations assuming that information leaves and arrives in discrete spikes. Furthermore, object recognition simulations with a computational model [2] suggest that the connections between visual areas must be able to relay at least 2 bits of information (within about 10 ms – to be discussed later), *i. e.*, represent approximately 4 different levels of spike activity at the population level during time bins of as little as around 10-20 ms. It has been found that the model is robust to perturbations of the max and tuning operations. In particular, the input to the highest level can be binarized without significantly affecting performance in multi-class recognition tasks. The entire visual recognition model, encompassing many maximum and tuning sub-circuits, must be able to perform the basic computation within an amount of time that is faithful to the known physiology. Recordings from inferotemporal cortex (IT) [13] reveal that information containing sufficient accuracy for good categorization arrives in anterior IT about 100-120 ms after stimulus onset and about 20 ms after onset of activity in AIT (which is about 80 ms after stimulus onset). In addition, [13] found that neural activity in a population of 256 neurons during a time bin of just 12.5 ms at around 20 ms after onset of the AIT response contain significant information for good categorization accuracy. This implies that each stage in the recognition model – roughly corresponding to a visual area (V1, V2, V4, PIT and AIT) – has about 10 to 20 ms for processing and representing the information. In the mammalian brain, the firing rates of

excitatory cells in visual cortex rarely exceed 100Hz, and we can expect at most 1-2 spikes per neuron within the prescribed time window. Thus, a single neuron cannot possibly provide enough dynamic range given the timing constraints imposed by available physiological data.

To overcome this insufficiency, we replace each unit with an ensemble of n cells. All cells in an ensemble have similar, but not identical, parameters and receive highly correlated inputs from the other units in the circuit. Instead of 0 or 1 spike, the postsynaptic cell can now receive up to $2n$ spikes, *i.e.*, the dynamic range of the input is multiplied by n , as shown in Fig. 1¹. In early stages of the visual pathway, a large dynamic range of the inputs is needed, whereas at the other extreme in IT, only the binary presence or absence of each feature has to be conveyed.² In the spiking circuits presented below, we therefore consider ensembles of spiking neurons for each unit in the computational model presented in [2]. In particular, we will also consider redundant parallel copies of circuit inputs, outputs, and networks themselves in order to overcome the above constraints placed on the timing and resolution of the two computations of interest. This is the essence of our contribution: we use ensembles of redundant spiking cells to achieve fast, “non-binary” computation.

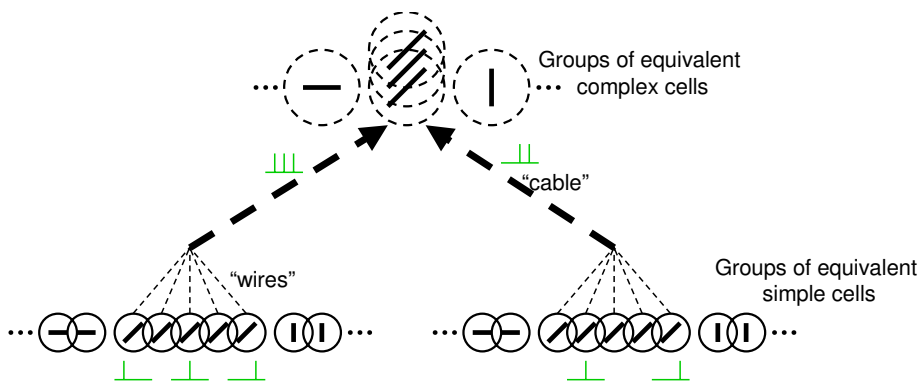


Fig. 1. Signal propagation between two somata with spiking neurons and “cables”. Each of the ensemble cells on the bottom, which receive highly correlated input, spikes independently. Each spike is propagated independently in its own “wire” in the “cable” and then triggers neurotransmitter release at one synapse for each of the ensemble postsynaptic neurons. This process is used for each synapse in the model.

¹ It is thought that the number of cells per ensemble n decreases along the visual hierarchy from V1 to IT.

² Contrast invariance data provide some indirect support for the idea that the *cables get thinner along the hierarchy*. [14] showed that the steepness of the contrast-response functions of neurons increases from LGN through V1, V2 to MT and that “cells become, in the contrast domain, progressively more like switches, being either on or off” [15].

1.2 Different Coding Regimes in Sensory Cortices

Neurons in the visual cortex of cats and primates are thought to have spontaneous and evoked firing rates on the order of tens of Hertz. At these rates, Poisson spike trains are a fair approximation for the observed firing patterns. Each presynaptic neuron can contribute up to 3-4 spikes within the first 20-30ms of processing in the postsynaptic cell and dynamic circuit models can assume that the underlying rate of the Poisson spike train inputs is constant.

However, there are other sensory cortices in which the coding is very different. Rodent somatosensory “barrel” cortex is an example for extremely sparse representation of sensory stimuli. Both the spontaneous and evoked firing rates in barrel cortex have been estimated to be below 1Hz by several groups. In other words, a single neuron responds to a given stimulus with on average less than one spike per trial, *i. e.*, either zero or one spike. These extremely low rates in conjunction with precise spike timing call for circuits that perform computations based on a single volley or “packet” of incoming spikes from an ensemble of presynaptic neurons. We call the corresponding coding scheme *Population Packet Code*.

Rat primary somatosensory cortex (SI) is an ideal model system to address questions about circuitry and coding, not only because of its sparse coding but also due to its regular columnar architecture, the barrel columns [16] and its externalized receptors, the vibrissae. These properties allow exquisite control over the stimulation, recording and analysis of discrete channels of information processing which are impossible with visual stimulation, since every visual stimulus will excite a large number of photoreceptors while the deflection of a single whisker only activates a single follicle. In addition, SI shares several basic properties of its circuitry and cells with cat or primate primary visual cortex (V1) and the two systems exhibit similar generalization and learning capabilities, likely because both are high-resolution sensory systems.

2 Population Packet Code Models

Transferring analog graded information with digital spikes poses the problem of how to encode analog values with spikes. Three common proposals to overcome this problem are to either use the population firing rate, a purely temporal code such as time-to-spike or representations based on synchrony. Our proposed coding scheme lies somewhat in between continuous population rate coding and synfire chains, two more widely studied coding schemes [17]. Given the constraints outlined above, we assume here that information is coded by the number of spikes in a brief packet of spikes from a presynaptic population, *i. e.*, by a short-time “packet” population rate code. In this coding scheme, when a single cell integrates spikes from several converging inputs, the dynamic range of the signal is strongly compressed. Either the combination of inputs from the ensemble of presynaptic neurons provides enough excitatory conductance to drive the postsynaptic membrane potential above spike threshold or not. Considering an ensemble of identical postsynaptic neurons all receiving identical input from

the presynaptic ensemble, the responses of all cells are identical. Thus, there will be no postsynaptic spikes for any combination of inputs below a certain threshold. Above this threshold, all postsynaptic cells will spike together. The dynamic range of the signal is compressed into a binary signal. However, neurons in cortex show diversity in their morphology and physiological parameters and receive ongoing “background”, noise-like input from many different presynaptic cells that are overlapping but not identical for a given set of postsynaptic neurons. It has been shown that this synaptic background input can decrease the slope of the spike probability function dependent on input strength [18]. In an ensemble of neurons, the synaptic activity together with variable intrinsic properties of the cells effectively linearizes the transfer function of the ensemble under the proposed coding scheme (see also [19] for a general argument).

2.1 Network Architecture for Softmax and Gaussian-Like Tuning

The population packet code model is a simplified canonical cortical microcircuit [20] using ensembles of integrate-and-fire neurons. Importantly, *the same canonical circuit perform either a max or a tuning operation depending on just different synaptic conductance values*. In the proposed coding scheme, computation has to be quasi-instantaneous, *i. e.*, carried out over in a short period of time during which each neuron can only spike at most once or twice. Thus the notion of a steady state firing rate is not appropriate for this case. Unlike the circuit presented in [21], which operates on a timescale of hundreds of milliseconds, our model is entirely feed-forward since the need for very fast and inherently transient computation implies that recurrent connections would most likely be too slow to contribute.

The most salient feature of our model architecture is the combination of monosynaptic excitation ($x \rightarrow y$) with disynaptic inhibition ($x \rightarrow p \rightarrow y$), which is strongly supported by the observation that strong excitation in form of an excitatory postsynaptic potential (EPSP) is generally followed by an inhibitory

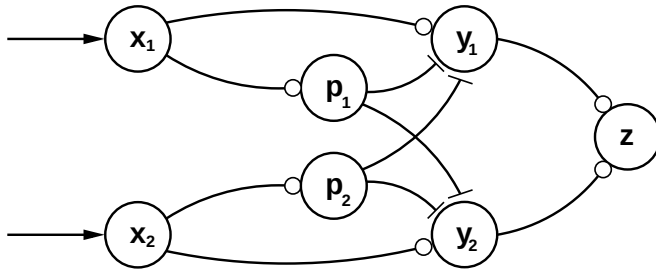


Fig. 2. Population packet code network architecture. Small circles indicate excitatory synapses, bars represent inhibitory connections. Note that the *same circuit is used to approximate a max as well as a tuning operation with different conductances for some of the synapses*.

postsynaptic potential (IPSP) after about 4ms, which has been made during intracellular recordings by several labs [22, 23, 24, 25]. Following the analogy with barrel cortex, the x units correspond to thalamic cells while the p , y and z units correspond to cortical cells. The inhibitory interneurons p synapse onto y units of the same channel as well as the other channel. The operating regime of these interneurons will be the determining factor for which computation (max or tuning) is performed by the microcircuit. Both y units synapse onto the output z unit, which relays the output. Each unit depicted in Fig. 2 is implemented as an ensemble of 100 cells modeled as a variant of integrate-and-fire neurons as described below.

Sparse connectivity. Although commonly used in models and simulations, there has been no anatomical evidence for ensembles of hundreds of neurons exhibiting all-to-all connectivity. Instead, connectivity patterns of cortical neurons depend on spatial distance. Although more careful analysis reveals a dependence on specific cell types [26, 27], nearby neurons are generally more interconnected and show a high correlation in their membrane potential [22]. Because not all cells in an ensemble are driven by exactly the same primary input, their output is more variable, increasing the dynamic range of the ensemble. We model this kind of connectivity by using a noisy Gaussian synaptic strength pattern between a given presynaptic cell and all its potential targets in the postsynaptic ensemble. For a presynaptic cell with index m and a postsynaptic cell with index n , the synaptic scaling factor is

$$\tilde{g}^{mn} = \exp\left(\frac{-(m-n)^2}{2\sigma_d^2}\right) \cdot \mathcal{N}(1, \sigma_g) \quad (1)$$

Consistent with experimental observations [28], inhibitory interneurons receive less specific inputs, *i. e.*, σ_d is larger for interneurons. The parameters used for these simulations are $\sigma_{de} = 50$ and $\sigma_{di} = 100$.

For the max computation, the synaptic conductances were $\hat{g}_{xy} = 1.2nS$, $\hat{g}_{xp} = 0.16nS$, $\hat{g}_{py} = 0$, $\hat{g}_{pyl} = 0.8nS$ and $\hat{g}_{yz} = 0.5nS$. The input conductance to the inhibitory interneurons p is smaller than the inputs to the excitatory neurons because of the different morphology, physiology and connectivity of these cells. The, in comparison to the tuning, relatively high $\hat{g}_{xp}$ causes the p units to be fairly active even for moderate input levels, causing strong cross-inhibition via the cross-channel $p \rightarrow y$ synapse.

For the tuning computation, the synaptic conductances were $\hat{g}_{xy} = 1.2nS$, $\hat{g}_{xp} = 0.09nS$, $\hat{g}_{py} = 4nS$, $\hat{g}_{pyl} = 4nS$ and $\hat{g}_{yz} = 0.72nS$. The tuning point, *i. e.*, the input eliciting the maximal output, is governed by the balance of excitatory and inhibitory input to the y unit in combination with the transfer function of the p unit. The lower $\hat{g}_{xp}$ in this configuration causes the p units to be activated only for stronger inputs, causing the decreasing output for those high input values because the inhibitory transfer function is steeper than the excitatory one, compatible with high-threshold, high-gain inhibitory interneurons found in cortex such as Chandelier cells [29].

2.2 Neural Dynamics

Each cell is modeled as a conductance-based single compartment leaky integrate-and-fire neuron with a simple kinetic model of neurotransmitter receptors. Variables and constants that are unique to each of the n_w instances in a group are marked with upper indices such as V_i^m and similarly for terms specific to an instance of a synapse such as g_{ij}^{mn} , the conductance of the synapse between cell m of group i and cell n of group j . If the upper indices are missing, the value is the same for all instances.

Integrate-and-Fire Dynamics. The membrane potential of a postsynaptic neuron is calculated according to the membrane equation

$$C_j^n \frac{dV_j^n}{dt} = \sum_{i,m} g_{ij}^{mn} (E_{ij} - V_j^n) + g_{Lj} (E_{Lj} - V_j^n) \quad (2)$$

The capacitive current charging the cell membrane is equal to the sum of the synaptic and leak currents. In order to better capture the firing properties of cortical neurons, we do not reset the membrane potential once it reaches the spike threshold V_θ but instead activate a strong hyper-polarizing current with a time constant of a few milliseconds which will bring the cell's potential back to its "resting" state.

Intrinsic cell parameters. Neurons in cortex have very different morphologies and channel distributions. While we cannot model these directly in our single compartment models, we impose some variability on the resulting physiological parameters such as membrane capacitance, input resistance and spike threshold. In accordance with several physiological studies, interneurons are assumed to be smaller, *i. e.*, having smaller membrane capacitance. The interneurons in this model are of the high-threshold, high-gain class discussed in numerous studies, *i. e.*, their spike threshold is higher than for the excitatory neurons ($V_{\theta_e} = -50mV$, $V_{\theta_i} = -45mV$). The membrane capacitance of a cell is a truncated normally distributed random but fixed variable of the form: $C_j^n = \hat{C}_{[e,i]} \cdot \tilde{C}_j^n$ with $\tilde{C}_j^n = \mathcal{N}(1, \sigma_C) \in [0.5, 1.5]$. The reversal potential of the leak current is set to $E_L = -70mV$ and its conductance is $g_L = 10nS$.

Synaptic Input. Release of neurotransmitter, the first step of synaptic transmission, is modeled as an all-or-none release of one "vesicle" of $1mM$ transmitter into the synaptic cleft which is present for $1ms$.

To model ionotropic receptors such as AMPA/Kainate and GABA_A receptors, we assume simple two-state kinetics where the receptor is either open or closed. In the closed state it can bind neurotransmitter which results in an transition to the open state. In the open state, the channel will close as a result of dissociation.



These simple kinetics can be described by the following first-order kinetic equation:

$$\frac{dr_{ij}^{mn}}{dt} = \alpha_{ij} T_{ij}^m (1 - r_{ij}^{mn}) - \beta_{ij} r_{ij}^{mn} \quad (4)$$

The rate constants α and β , are set to the values reported in [30] for either AMPA and GABA_A, respectively. The synaptic input conductance is calculated from the fraction of open channels calculated in (4), adding a normally distributed random background component (see below):

$$g_{ij}^{mn} = [r_{ij}^{mn} \hat{g}_{ij} \tilde{g}_{ij}^{mn} + \mathcal{N}(\bar{g}_{ij}, \sigma_{g_{ij}})]^+ \quad (5)$$

The current results are based on AMPA as excitatory and GABA_A as inhibitory neurotransmitters and their respective receptors. The parameters are $E_{AMPA} = 0mV$, $E_{GABA_A} = -80mV$, $\alpha_{AMPA} = 1.1 \cdot 10^6 M^{-1}s^{-1}$, $\alpha_{GABA_A} = 5 \cdot 10^6 M^{-1}s^{-1}$, $\beta_{AMPA} = 190s^{-1}$, $\beta_{GABA_A} = 180s^{-1}$.

Noise-like background synaptic input. Background synaptic noise, which is likely to represent unknown inputs, helps to recover the dynamic range of a population of neurons by linearizing the spike probability function [18]. If different neurons in an ensemble have slightly different synaptic inputs, their membrane potential will slightly vary. Thus, the same incoming postsynaptic potential (PSP) might drive some neurons with higher membrane potentials above the spike threshold, initiating a spike, while others with lower potential will stay below threshold and not spike. Adding synaptic background inputs also introduces a variability in the output spike timing. Compatible with data from *in vivo* and *in vitro* recordings, spike time jitter decreases with increasing strength of the input. The parameters used for the conductances (5) are $\bar{g}_e = 2nS$, $\bar{g}_i = 1nS$, $\sigma_{g_e} = 1nS$, $\sigma_{g_i} = 1nS$.

External Input. As input to the circuit, the x units receive a variable number of spikes proportional to the input strength. Each spike is assumed to originate at one out of 100 presynaptic cells. The connectivity pattern is the same Gaussian shaped synaptic strength profile as for the other excitatory units in the model and the spike arrival times are normally distributed as $\mathcal{N}(\bar{t}_x, \sigma_{t_x})$ with $\bar{t}_x = 30ms$ and $\sigma_{t_x} = 5ms$.

2.3 Population Packet Code Simulations

Maximum. Figure 3 (left) shows the mean output over 50 runs of the circuit depicted in Fig. 2 in maximum configuration for all possible combinations of four levels of inputs (0, 50, 75, 100 spikes in a packet). For input combinations including the maximum (100 spikes) input, the circuit output is actually below the desired maximum value. Interestingly, similar effects can be seen *in vivo*, *e. g.*, in rat barrel cortex, where subthreshold neural responses tend to be more suppressive for stronger inputs. The histograms on the right show the distribution of outputs for the three possible desired output values (50, 75 and 100 spikes from top to bottom). As mentioned above, the circuit underestimates the desired output for inputs that include full activity (100 spikes) for at least one of the two channels. For all three cases, about 20-25% of the runs result in an output that does not match the desired output value. While this behavior is consistent

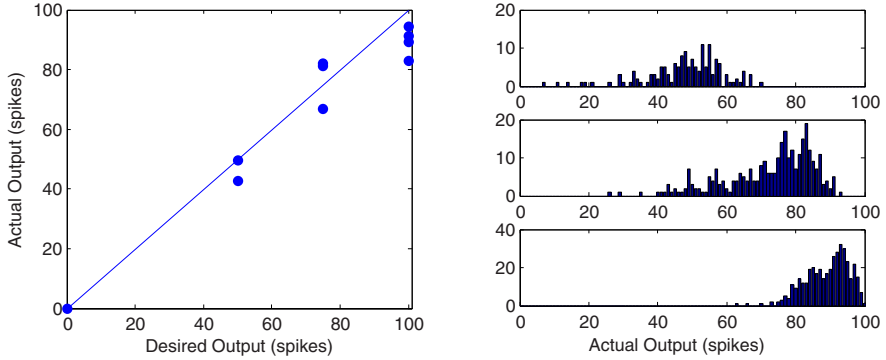


Fig. 3. Mean response of max circuit depicted in Fig. 2 over 50 runs for all possible combinations of 0, 50, 75 and 100 spikes per input packet, plotted against the desired (true) maximum of the inputs (left). Histogram of all outputs (spike count in output packet) for three cases (right). The true maximum of the inputs is 50, 75 and 100 spikes, respectively (top to bottom).

with physiological data, it has to be investigated how well a large-scale model of object recognition [2] will perform under these conditions.

Tuning. In the tuning configuration, the circuit shown in Fig. 2 effectively performs a one-dimensional tuning for each of the input channels, *i. e.*, each y unit will be maximally active for the input level it is tuned to and activity will decrease for lower or higher input values. The y unit activity is then combined to yield the output which is tuned in the multidimensional space. An example for the activity of a y unit can be seen in Fig. 4 (left). In this case, the circuit is tuned for an input of 50 spikes. Output activity is highest for a 50 spike packet

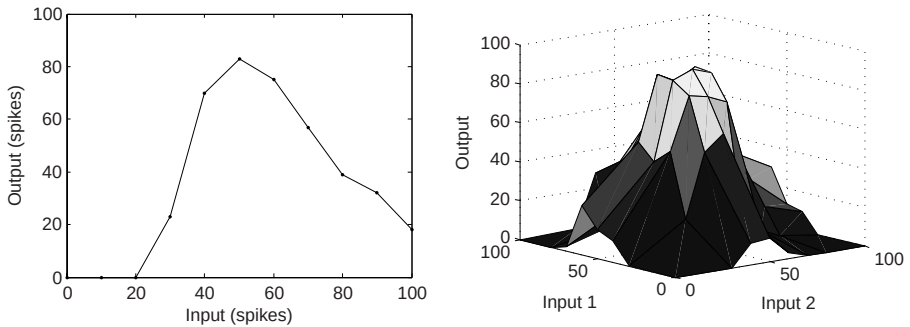


Fig. 4. Output (spike count in output packet) of a one-dimensional Gaussian-like tuning circuit tuned to 50 a spike packet input (left). Output (spike count in output packet) of the two-dimensional tuning circuit depicted in Fig. 2 tuned to the combination of two 50 spike packet inputs (right).

and falls off on both sides. The tuning curve is not completely symmetric but approximates a Gaussian tuning curve well.

The output of a circuit tuned to the combination of two 50 spike packets as inputs is shown on the right of Fig. 4. The peak output activity is located at the (50,50) input pair and the response falls off in a roughly Gaussian (bell-shaped) fashion for any other input configuration.

2.4 Problems and Future Goals

- The main issue with the circuit of Fig. 2 is how well it can be extended to a significantly higher number of inputs and dimensions.
- The performance in the softmax configuration should be improved in order to meet the likely requirements of the recognition model (which are not stringent however).
- Learning the parameters that determine the Gaussian-like “centers” – that is the optimal stimulus for the tuning circuit – is an open problem.
- More careful measurements of the time required by the computation and of the bandwidth of the output activity are needed.

3 Poisson Spike Code Models

In this section we present independent spiking models for the maximum operation performed by complex V1 cells, and for normalized tuning observed in simple V1 cells. In both cases, and in the simulations which follow, we assume that the networks are driven by Poisson distributed spike inputs. For the maximum circuit, our goal is to output a sequence of spikes that encodes the maximal level of activity present at the circuit’s inputs. In the case of the tuning operation, we wish to output a sequence of spikes which encodes some notion of the similarity between the input activity pattern and a preferred stimulus. We describe a circuit for which the similarity peaks when the input activity vector is collinear with the preferred stimulus, and falls off to zero along orthogonal directions in a Gaussian-like fashion. In the sections that follow, we will make these ideas more precise.

3.1 Max Circuit Architecture

Neural Dynamics. The maximum circuit design incorporates leaky integrate-and-fire models of neurons augmented with plausible synapse dynamics and an absolute refractory period. The choice of a first order linear model for the membrane dynamics was made in order to make large scale simulations involving many circuits computationally tractable. The maximum operation itself can be seen as a particular instance of K-winner-take-all behavior, and has been explored in networks of continuous nonlinear dynamical elements [31, 32, 33], and to a lesser extent, in networks of “spiking” elements [34, 35]. The use of dynamics which include reset rules, such as integrate-and-fire neurons, makes detailed

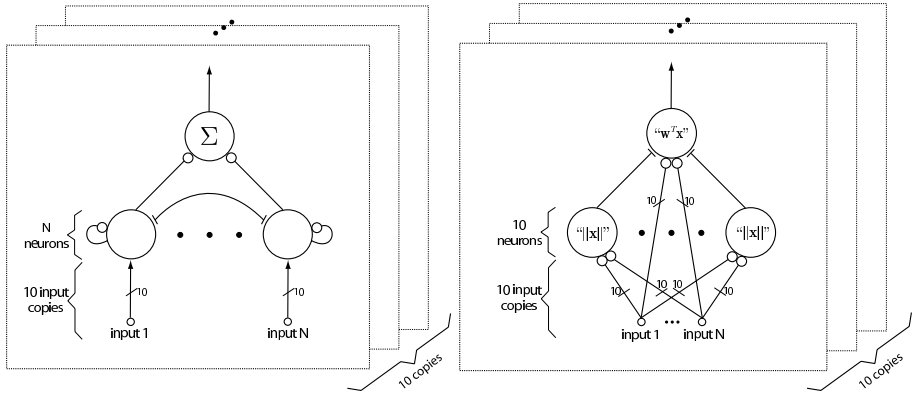


Fig. 5. Maximum circuit architecture (left), and tuning circuit architecture (right), assuming Poisson distributed inputs. Small circles and bars denote excitatory and inhibitory connections respectively.

analytical explorations of the behavior of interconnected networks of such elements difficult, and we do not carry out such an analysis here. Instead, we provide a description of the membrane potential of a single unit as it evolves in time, explain how individual elements should be connected together and tuned in order to compute the maximum over the input activities, and finally, show how multiple copies of circuits can be arranged to match the timing observed in cortex.

The particular winner-take-all design at the core of the maximum implementation exploits a balance between self-excitation and, in this case, all-to-all inhibitory dynamics. The connections are illustrated graphically in Figure 5 (left), where we denote excitatory feedback connections with small circles and inhibitory connections with bars. For additional computational simplicity, we do not include sign-changing interneurons in the inhibitory feedback paths.

We denote by V_L the neuron's resting potential, g_L the leakage conductance, and by g_j the synaptic conductances. With these definitions, the dynamics of each cell's subthreshold membrane potential $V_i(t)$ can be described by:

$$C_m \frac{dV_i}{dt} = g_L(V_L - V_i) + \sum_{j=1}^N g_j(V_j^r - V_i) \quad i = 1, \dots, N. \quad (6)$$

The $g_j = g_j(t)$ are time-varying conductances that depend on activity at the incoming synapses, while C_m , g_L and V_L , are constants. The synaptic reversal potentials V_j^r take on one of two constant values $V_j^r \in \{V_{ex}, V_{inh}\}$ depending on whether the synapse is excitatory ($V_{ex} > V_\theta$) or inhibitory ($V_{inh} < V_\theta$, $V_{inh} \neq V_L$). The subthreshold dynamics (6) are combined with a nonlinear reset rule which stipulates that when the membrane potential $V_i(t)$ reaches the threshold V_θ , an action potential is fired and the potential is then immediately reset to the

reset voltage $V_0 < V_\theta$. Following reset, an absolute refractory period is simulated by further suspending the subthreshold dynamics (6) for τ_{abs} seconds. Integration is then resumed with the initial condition $V_i(0) = V_0$.

We can rearrange the subthreshold dynamics (6) into the form

$$\tau_i \frac{dV_i}{dt} = V_i^\infty - V_i \quad (7)$$

where

$$\tau_i = \frac{C_m}{g_L + \sum_j g_j}, \quad \text{and} \quad V_i^\infty = \frac{g_L V_L + \sum_j g_j V_j^r}{g_L + \sum_j g_j}. \quad (8)$$

When viewed in this way, it is clear that the presence of time-varying synaptic inputs, in contrast to electrical feedback connections, implies that the total synaptic conductance as well as the membrane time constants τ_i , must depend on time via the input activity.

Synapse Dynamics. As an added degree of realism, the synapse conductances $\{g_i\}$ corresponding to excitatory feed-forward inputs are not scaled versions of the input spike trains, but are instead modeled as filtered (averaged) approximations. Given a discrete spike train $p_i(t)$, a post-synaptic current profile $\alpha(t)$, and a constant input conductance multiplier $\bar{g}_{in}$, synapse integration dynamics can be modeled as

$$g_i(t) = \bar{g}_{in}(\alpha * p_i)(t). \quad (9)$$

The change in conductance $\alpha(t)$ in response to an incoming spike is modeled as a decaying exponential with finite initial rise time. While this choice of post-synaptic conductance response is indeed more realistic than a simple decaying exponential of the form $e^{-t/\tau_{syn}}$, the finite rise time exponential was ultimately chosen because it closely approximates the detailed dynamics (4). In the simulations presented in section 3.3, we simply pre-compute the response and perform a table lookup, rather than simulate additional dynamics online. The particular parametrization we have chosen models the time course of the conductance as a difference of two exponentials [36]

$$\alpha(t) = \frac{1}{\tau_d - \tau_r} \left(e^{-t/\tau_d} - e^{-t/\tau_r} \right) \quad (10)$$

where the rise and decay time constants are denoted τ_r and τ_d respectively. Given this description of the conductance, the free parameters were tuned so that the shape of (10) matched as closely as possible the dynamics (4) given the parameters introduced in section 2. We find that $\tau_r = 0.4\text{ms}$ and $\tau_d = 4\text{ms}$ fits the dynamics for the AMPA-based excitatory response, and $\tau_r = 0.2\text{ms}$ with $\tau_d = 5.4\text{ms}$ fits the dynamics for the GABA_A-driven inhibitory response. In the absence of multiple spikes arriving in a short time interval, the channel dynamics (4) are well approximated by this more familiar “alpha function”³.

³ We will refer Equation (10) as an alpha-function, even though it is not, strictly speaking, of the form $te^{-t/\tau}$.

Given the firing rate constraints described above, it is unlikely that multiple spikes will arrive within a short (e.g. 12.5ms) window.

In the interest of computational expediency, conductance changes for the excitatory inputs can be computed by convolving the kernel (10) with the input spike trains prior to simulation. Convolution amounts to summation of responses in the case where a spike arrives before the previous spike's alpha-function has decayed to zero. In the case of inhibitory synapses in the feedback paths of the circuit, however, it is often more convenient to follow a slightly different convention: If a new spike arrives before the previous spike's alpha function has decayed to zero, the two need not be added. Instead, the alpha function is "reset" to its initial value for that synapse. Resetting can be justified by making the assumption that an incoming spike causes most of the channels in the vicinity of the synapse to open, and that if another spike arrives, more than the maximum number of channels cannot open. Once again, because firing rates are limited to approximately 100Hz, we can assume that both choices will produce roughly equivalent behaviors.

With the input conductances computed beforehand, numerical integration of the system (7) is accomplished by discretizing time into finite steps δt , and applying an Euler update rule. The conductance response $\alpha(t)$ for inhibitory spikes that occur during the simulation can be incorporated into the integration by simply retrieving from a table the particular conductance along the curve (10) corresponding to the amount of time since the last spike arrived for the synapse of interest.

Circuit Organization & Simulation. As the stacked "planes" in Figure 5 (left) suggest, we combine multiple redundant copies of the circuit in order to (1) reliably estimate input activity and (2) compute the maximum within a short time window. Each individual "copy" of the circuit we have described is defined to have identical integrate-and-fire parameters, conductances, and architectures. The external inputs applied to each copy, however, will differ on a spike-by-spike basis and are not assumed to be synchronized, but will still have identical mean activities. Thus the outputs of the circuits will also not be synchronized, but will have identical average activities and mean times to the first spike. The collection of outputs taken from a group of circuits can then be fed into subsequent stages so as to enforce the required timing constraints. One can think of a group of such circuits as a larger meta-circuit capable of delivering an answer in a short amount of time. In this case, the larger circuit requires $K(2N+1)$ neurons, for K circuit copies and N inputs. One complex unit in the model [2] thus corresponds to this same quantity of integrate-and-fire neurons.

The physical cell parameters for the max circuit described in this section, and simulated in section 3.3, were chosen as follows: leakage conductance $g_L = 25\text{nS}$, feedback (self-)excitatory and inhibitory conductances $g_{ex} = 10g_L$ and $g_{inh} = 15g_L$, external (excitatory) input conductance $g_{in} = 4g_L$, refractory delay $\tau_{ref} = 8.5\text{ms}$, inhibitory and excitatory reversal potentials $V_{inh} = -80\text{mV}$ and

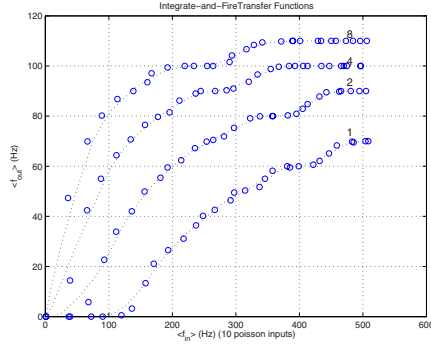


Fig. 6. Integrate-and-fire spike-rate transfer functions for a variety of input conductances. Each trace represents the output of one neuron given 10 Poisson inputs. Conductance units shown on each curve are in multiples of a canonical conductance.

$V_{inh} = 0\text{mV}$, leakage potential $V_L = -70\text{mV}$, reset potential $V_0 = -70.25\text{mV}$, spike threshold potential $V_\theta = -54\text{mV}$, membrane capacitance $C_m = 0.5\text{nF}$, and maximum firing rate $F_{max} = 100\text{Hz}$.

3.2 Circuit Architecture for Gaussian-Like Tuning in the Direction of the Input Vector

Neural Dynamics. In this section we present a circuit which performs *normalized* tuning: given a preferred stimulus encoded by the strengths of the input synapses of the circuit, we would like the output activity of the network to peak when the input activity “vector” is collinear with the preferred stimulus, and fall off to zero, much like a Gaussian does, when the input and preferred stimulus vectors move towards orthogonality. In addition, the output activity should be normalized, in some way, by the total input activity. This normalization can take on a variety of possible forms, and, depending on its strength, leads to behaviors where the output activity of the circuit decreases or remains constant while the input activity increases, for a given fixed angle between the inputs and the preferred stimulus. While the tuning circuit presented in section 2 approximates a multidimensional Gaussian in the input space, the model presented in this section exhibits the latter aforementioned behavior. The tuning function is Gaussian-like for the direction of the normalized input vector with respect to a preferred stimulus, but is not shaped like a Gaussian in the input space because the output activity does not decrease when the total activity increases for a fixed angle between the input and the preferred stimulus vectors. Normalization is still, however, applied so that the total output activity of the circuit does not increase when input activity increases, but instead remains constant.

We approximate normalized tuning by delivering both divisive, normalizing inhibition and weighted excitatory signals to units operating in nonlinear regions of the spike-rate transfer function. We define the spike-rate “transfer function”

to be the instantaneous output firing rate of a single integrate-and-fire neuron versus the sum of the instantaneous arrival rates of spikes at the unit’s excitatory synapses, assuming no additional background input or current applied to the neuron. This transfer function is of course sensitive to the conductance of the input synapses, and we illustrate how different synaptic conductances can lead to different transfer function shapes in Figure 6.

One possible tuning circuit architecture is shown in Figure 5 (right), where units in the lower layer sum input activity and then inhibit a global pooling unit which also receives the original input. We take as our desired tuning model the normalized dot-product formulation described in [2]

$$y = g\left(\frac{\sum_j w_j x_j^p}{k + (\sum_j x_j^q)^r}\right) \quad (11)$$

where $g(\cdot)$ is a sigmoid nonlinearity, $\mathbf{w} = (w_1, \dots, w_N)^T$ is the vector of synaptic strengths, and $\mathbf{x} = (x_1, \dots, x_N)^T$ is the vector of inputs to the circuit. If the (integer-valued) exponents in (11) are chosen so that $p < qr$, then the output y will peak when the input $\mathbf{x}$ is “close” to the preferred stimulus encoded by $\mathbf{w}$, but will fall off as the total input activity increases while maintaining the same angle with the preferred stimulus vector. If $r = 1$ and $p \approx q$, then the output will peak when the input is close to the preferred stimulus, and will remain at the peak activity level if the total input activity increases (but will not increase any further as one would observe in the absence of any normalization).

In the spiking model described here, we attempt to roughly approximate exponentiation in the numerator and denominator terms in Equation (11) by choosing suitable operating points on the spike-rate transfer curves for units computing x^q , and separately, x^p . With r set to 1, $p \leq q$ means that neurons implementing the denominator in Equation (11) should exhibit locally steeper, more nonlinear transfer functions compared to numerator neurons. In the simulations that follow in section 3.3, the steepness of the initial rise in the spike-rate transfer function is controlled by adjusting the conductance of the inputs synapses only, however we found that even with a denominator conductance much larger than that of the numerator, the circuit behaves as if $p \approx q$ with $r = 1$. In this case, the normalization compensates for increasing input activity, but not to the extent that the total output activity begins to decrease, as discussed above: the “shape” of the tuning function in the input space is not a symmetric, multidimensional Gaussian as shown in section 2.

Circuit Organization & Simulation. After selecting parameters yielding suitable operating points and transfer functions, the circuit performs the tuning function in two feed-forward stages. First, the sum of the input activities is computed and represented by an output spike train 10 separate times. These 10 outputs are connected via shunting inhibitory synapses to a single pooling unit which also receives multiple copies of the input. The inputs at the pooling unit are each assigned specific conductance strengths which together encode a preferred stimulus. In the second stage of the circuit, the weighted sum of the

input activity is combined with the effect of the inhibitory inputs in a divisive manner, giving normalized tuning at the output of the pooling unit. As before, multiple copies of the entire circuit are utilized in order to meet realistic timing requirements. One simple unit (S_n , for $n > 1$) in the model described by [2], would therefore correspond to the 10 integrate-and-fire output (pooling) neurons for each of the 10 circuit copies. In general, each tuning unit requires $K(P + 1)$ neurons, for K circuit copies and P inhibitory normalization neurons per circuit.

In the simulations presented in Section 3.3, we implement all units with integrate-and-fire neurons with refractory periods, synaptic conductance functions, and numerical simulation methods identical to those described in Section 3.1. We note, however, that unlike the maximum circuit, the tuning architecture is purely feed-forward and therefore allows for simplified computer simulation. In particular, the normalization and pooling components are decoupled and may be computed independently, while the effect of the alpha-function on the synaptic conductances may be computed using convolution (as shown in Equation (9)) before numerical integration of the integrate-and-fire dynamics.

The physical cell parameters in the tuning circuit were chosen as follows: leakage conductance $g_L = 25 \times 10^{-9} S$, excitatory input conductance $g_{ex} = 8g_L$ for the lower row of normalization units and excitatory conductance $g_{ex} = 4g_L$ with inhibitory conductance $g_{inh} = 20g_L$ for the upper pooling unit, refractory delay $\tau_{ref} = 8.5\text{ms}$, and inhibitory reversal potential equal to the leak potential $V_{inh} = V_L = -70\text{mV}$. All other physical parameters were identical to those described in section 3.1.

3.3 Poisson Spike Code Simulations

Spike Process Model. In the simulations that follow, we use a Poisson distribution to model the spike arrival process. It is worth mentioning that the variance of a homogeneous Poisson spike process with rate parameter λ is also λ , and thus it is possible to occasionally generate biophysically unrealistic instantaneous firing rates in simulation. One advantage of this fact, however, is that the Poisson model leads to a conservative estimate as to the number of circuit copies necessary to achieve an accurate max computation within a small time window (by averaging over multiple circuits); restricting the variance of our spike processes can only improve simulation performance.

Maximum Circuit Simulations. To simulate the maximum circuit, we applied as inputs several sets of features returned by the $S1$ layer of the model described in [2]. These features were computed by filtering a set of natural images with oriented edge detectors at different scales, positions and orientations, and passing the result through a sigmoid nonlinearity. The feature sets ranged in size from 64 (8x8 patches) to 484 (22x22 patches), while the feature values were quantized into 4 levels, including “zero”. Given the features, we generated 10 copies of each input by sampling 10 separate Poisson spike input trains with mean arrival rates proportional to the feature’s value. Thus, for N inputs, for example, we generated $10N$ Poisson spike processes, and applied them to each

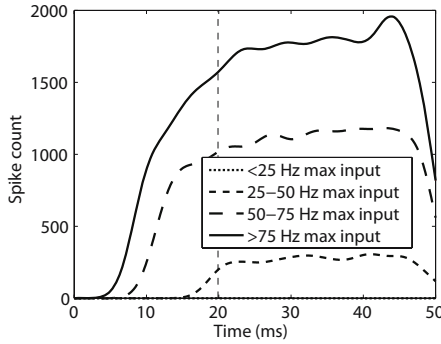


Fig. 7. Max circuit time response simulation. (This figure courtesy Tim Masquelier.)

input unit in each circuit copy. With 10 circuit copies, this corresponds to $100N$ separate input synapses. To evaluate the performance of the maximum circuit, we examine output spike rates as well as transient output characteristics (e.g. time to the first spike).

In Figure 7 we show the time course of the circuit's response to four 64-input stimuli in which the true maximum activity was one of the respective allowed input levels. Each trace represents the combined mean firing rate of 10 circuit copies estimated using a short Gaussian-shaped time window. The transient response of the circuit can be seen to carry a great deal of information, with larger input activities evoking earlier and larger responses. Figure 8 depicts the accuracy of the circuit given 64- and 144-input stimuli, over 2300 different stimulus instances. Each open circle marks the approximate instantaneous firing rate produced by each output neuron at a point 20ms following presentation of a distinct set of inputs. The instantaneous firing rate was estimated by computing the total output firing rate in the simulation time interval [15ms,25ms], and dividing by the total number of output neurons (64 or 144 of them, in this case) for all circuit copies (10 in these simulations). We then arrive at an instantaneous firing rate per output neuron, where there are $N \times K$ outputs if the dimensionality of the input is N and the number of circuit copies is K . Finally, because the circuit was presented with stochastic Poisson spike trains with mean spike rate equal to one of the levels of activity specified by the quantized image features, it should be noted that there is significant variance in the firing rates actually delivered to the circuit.

In order to score the accuracy of the maximum circuit, we divided up the range of the output firing rates into 4 bins that proportionally preserve the division of the input domain into its 4 levels. Using these bins, we then counted the number of instances where the output did not fall into the correct output bin, as determined by the bin of the maximum input. Thus if the output firing rate falls into the correct corresponding output bin, it is deemed a successful trial. Over all 2300 trials, only a small number did not meet this criteria: in 98.48% of the 64-input trials and 98.13% of the 144-input trials, the correct maximum bin was produced. We have also experimented with larger input sets, of size 256

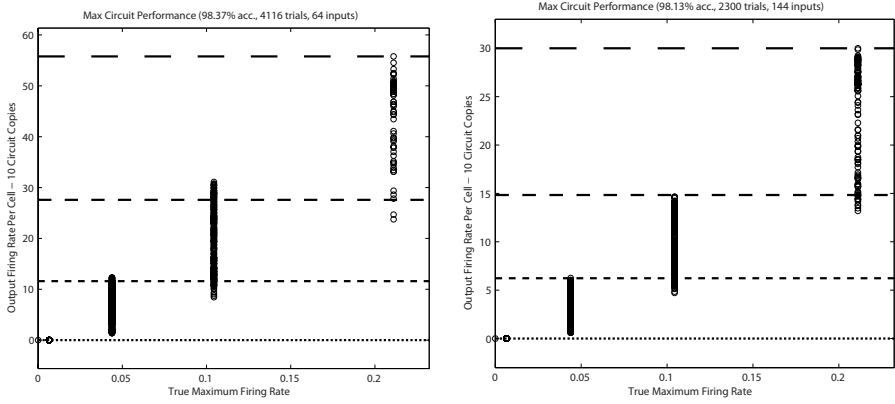


Fig. 8. Max simulations with 64 (left) and 144 (right) inputs (over 2300 trials). The abscissa gives the true, desired activity level, while the ordinate gives the estimated instantaneous firing rate per output neuron at a point 20ms after stimulus onset. In all trials the circuit is simulated with Poisson spike inputs so that input activities are never exactly equal to the ideal discrete levels shown along the x-axis.

and 484, and found that accuracy scales well with the number of inputs. The slight decrease in accuracy with input dimensionality seen in the above figures can be explained by the fact that more inputs offer more of an opportunity for there to be repeated inputs which take on the maximum value. Because the winner-take-all dynamics involve noisy spike processes rather than continuous quantities, if there are several inputs close to the maximum value, not all of them will be suppressed all of the time. This situation can lead to exaggerated spike rates and results in a response that occasionally falls into a higher level bin than appropriate.

Tuning Circuit Simulations. In this section we describe experiments which illustrate the behavior of the tuning circuit shown in Figure 5 (right). In simulating this circuit, care must be taken to ensure that inputs and outputs maintain realistic firing rates. This constraint, however, complicates testing the tuning properties of the circuit because constraints on the activities of the inputs translate into constraints on the way in which we are able to test the circuit given specific desired angles between the input and the preferred stimulus. In particular, we would like to choose a set of input activities $\{\mathbf{x}^j = (x_i^j, \dots, x_N^j)^T\}_{j=1}^J$ that sweep out a range of angles $\{\theta_1 = 0, \dots, \theta_J = \pi/2\}$ while satisfying

$$0 \leq x_i^j \leq F_{max}, \quad \text{and} \quad \sum_{i=1}^N x_i^j = M \quad (12)$$

where N is the number of inputs to the circuit, M is a fixed, total level of activity, F_{max} is the maximum allowed firing rate for a single neuron, and J

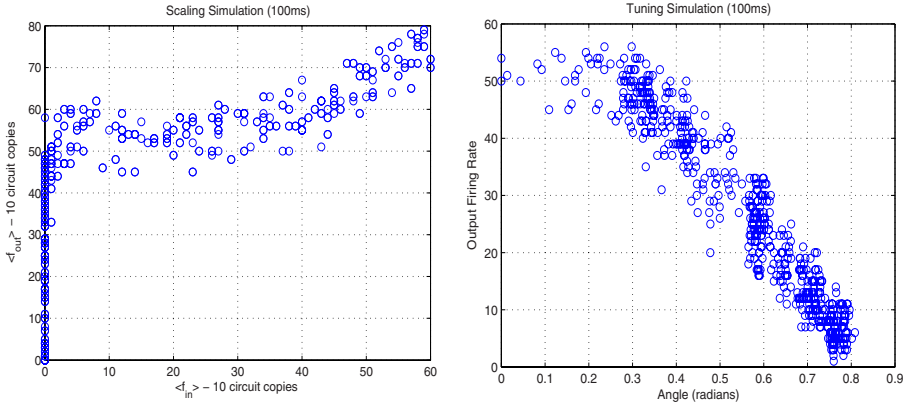


Fig. 9. Simulations illustrating scaling (left) and tuning (right) properties of the tuning circuit shown in Fig. 5 (right), with 10 circuit copies and 64 inputs. Although we have plotted the circuit response for only positive angles in the right-hand panel, the tuning profile can be considered Gaussian-like because it is symmetric in the angle.

is the chosen number of input instances that we will apply to the circuit to test its behavior. The second (normalization) constraint in (12) is imposed when testing the tuning properties in order to fix the operating point along the scaling

curve.⁴ If the total activities changed over the input vectors, then the tuning properties would change, and responses for different angles would not be directly comparable. Conversely, when testing the scaling properties of the tuning circuit, the angle between the inputs and the preferred stimulus vector must be fixed while varying the total activity over a chosen range.

To select vectors for testing the tuning behavior of the circuit, we set the first input vector in a set of inputs equal to the preferred stimulus, $\mathbf{x}^0 = \mathbf{w}$, and then generate a sequence of J vectors that successively drift away from collinearity with the weight vector by applying the following sampling process: At each iteration t , we begin by selecting two random components from $\mathbf{x}^{(t-1)}$, $x_u^{(t-1)}$ and $x_v^{(t-1)}$. Given a small but fixed adjustment amount Δ , we generate the new vector $\mathbf{x}^{(t)}$ by applying the following updates:

$$x_i^{(t)} = x_i^{(t-1)}, \quad i \neq u, v \quad (13a)$$

$$x_u^{(t)} = x_u^{(t-1)} + \Delta \quad (13b)$$

$$x_v^{(t)} = x_v^{(t-1)} - \Delta \quad (13c)$$

If a component falls outside of the range $0 \leq x_i \leq F_{max}$, then the candidate vector is rejected and a new one is resampled. The preferred stimulus was chosen

⁴ We stress however that normalized inputs are *not* required during ordinary operation of the circuit; we only normalize the total activity here to evaluate tuning vs. input angle while controlling for overall input activity.

randomly, but fixed for the duration of the analysis. The above procedure works well for finding vectors that slowly travel from collinear to angles near $\pi/4$ with respect to the weight vector when using the convention that we always take the smallest positive solution θ to $\theta = \cos^{-1}(z)$ for $z \geq 0$.⁵ To generate inputs near orthogonality, we searched for a vector $\mathbf{v} \perp \mathbf{w}$ by solving the constrained optimization problem

$$\mathbf{v} = \arg \min_{\mathbf{x}} \left\{ \cos^{-1} \left(\frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\| \|\mathbf{x}\|} \right) - \frac{\pi}{2} \right\}^2$$

subject to the constraints (12), and again using the convention that we select the smallest positive angle satisfying $\theta = \cos^{-1}(z)$. We then applied the sampling procedure (13) to $\mathbf{v}$ and generated a sequence of inputs whose angles with the preferred stimulus drifted from orthogonal to approximately $\pi/4$. Finally, the scalar components of the activity vectors found above were used as mean arrival rates to sample Poisson spike input trains that were then applied to the circuit during simulation. Vectors for the scaling simulation were generated by fixing the angle to be collinear with the preferred stimulus for all test inputs, and simply varying the length of the test vector over a range which ensured that the individual activities did not exceed the maximum firing rate. In order to evaluate the performance of the circuit over a wide range of conditions, we did not constrain the scalar input vector components or Poisson arrival rates to take on one of the 4 allowed values represented in the model [2].

In Figure 9 we show the behavior of the circuit with fixed input angle over a range of activities (left panel), and with fixed activity over a range of input angles (right panel). For fixed input activities, it is clear that the circuit exhibits an appropriate fall-off in output activity as the input moves away from the preferred stimulus in angle. We have shown only the positive angles corresponding to each response, however the tuning curve is symmetric for negative angles as well, and thus approximates a Gaussian.

If the circuit is performing a *normalized* dot-product, as opposed to the canonical dot-product, then regardless of the magnitude of the input, and for a fixed angle, the response should either decrease or stay roughly constant depending on the choice of the exponents p, q and r in Equation (11). The left-hand panel in Figure 9 verifies that this is approximately the case over a wide range of input activities: when the average input firing rate is in between 5 and 20Hz, the output firing rate decreases slightly. Between 20 and 40Hz the output remains constant on average. Beyond 40Hz, the operating points of the neurons in the circuit become shifted into a regime where the divisive normalization no longer over-compensates for the “length” of the input, and we see a proportional increase in

⁵ Normalized random vectors with positive uniformly distributed components form an absolute angle (modulo 2π) near $\pi/4$ with the optimal stimulus on average, and will rarely form angles near $\pi/2$ or 0 radians. For this reason, we sample input vectors starting at 0 radians moving to $\pi/4$, and then from $\pi/2$ back to $\pi/4$.

the output firing rate with respect to the input firing rate. Whether tuning in the higher levels of visual cortex takes the shape of a multi-dimensional Gaussian or not is still, however, under discussion. Several researchers have identified and modeled cells exhibiting a wide range of contrast normalization and tuning characteristics, including units whose output activity decreases or remains constant as a function of input activity [37, 38].

In all simulations, we assumed 10 circuit copies, and 10 input copies, 64 inputs, and took as “output” the sum of the average spike rates of the 10 circuit outputs over the first 50ms of the simulation. The dynamics were integrated using Euler step sizes $\delta = 0.1\text{ms}$.

3.4 Problems and Future Goals

- The tuning circuit presented above may be made more flexible and possibly configured to behave like a true Gaussian function in the input space through a more careful evaluation of the trade-off between conductances in neurons implementing the numerator and denominator of Equation (11).
- The resolution of the output and the time delay from input presentation to stable outputs in both circuits may be possibly improved by adding additional circuit copies and adjusting the integrate-and-fire membrane time constants (subject to biophysical constraints).
- The potential role of feedback in the tuning circuit should be explored. It is likely that feedback connections can be used to modulate the shape of the Gaussian-like tuning function by sharpening neural responses, and (separately) might facilitate normalization if used in a gain-control configuration.
- Learning the parameters that determine the Gaussian-like “centers” – that is the optimal stimulus for the tuning circuit – is also an open problem in this case.
- More careful measurements of the time required by the computation and of the bandwidth of the output activity are needed.

4 Discussion

We presented two different sets of circuits which implement the maximum and tuning operations under two different coding schemes. The circuits serve as a proof of concept that these computations can be performed by spiking neural circuits in the cortex.

We introduced a coding scheme in section 2 that is compatible with the spontaneous and evoked firing rates observed in rat barrel cortex and presented a canonical microcircuit that is able to perform both the max and the tuning operations, depending on different values for its synaptic weights. This not only argues for the plausibility of the model of object recognition presented in [2], but also provides support for the idea of a canonical microcircuit present in

multiple (at least sensory) cortical areas [20]. The main architecture of the circuit follows the observation of monosynaptic excitation combined with disynaptic inhibition found in layer IV of primary sensory areas. Although this connectivity has mostly been investigated in the thalamo-cortical circuit, it is conceivable that higher cortical areas might follow the same architecture, even though they receive their inputs from other cortical areas instead of thalamic nuclei. Both the max and tuning configurations of the presented microcircuit exhibit trial-to-trial variability. Whether this variability is feasible for high-level computational models such as [2] needs to be investigated, however, it appears to be compatible with the variability of neurons in cortex. More stringent characterization of the variability in both cases is needed in order to quantitatively compare models and experimental data.

Under the assumption that the peak bandwidth for communication between visual areas is limited to approximately 2 bits of information over time intervals of approximately 10-20 ms, the maximum architecture described in section 3.1 is capable of accommodating large numbers of inputs while maintaining high accuracy. Because the circuit was designed to operate given Poisson distributed inputs, it is relatively robust to noise in the form of spike addition, deletion, and timing jitter. The use of multiple redundant circuit copies is a critical feature that facilitates computation which satisfies biophysical timing and resolution constraints. Thus, several circuits may be connected together in series, as would be required by the model in [2], while maintaining realistic timing properties. The tuning circuit presented in section 3.2 was shown to have the desired tuning and normalization characteristics, and can also scale to accommodate large numbers of inputs. As in the case of the max circuit, multiple tuning circuit copies were utilized in order to satisfy the stringent timing and resolution constraints imposed by the neuroscience of object recognition.

There are several important assumptions underlying the circuits presented in this paper that must be considered when attempting to connect spiking artificial circuits to circuits in cortex. The winner-take-all configuration of the max circuit in section 3.1 requires all-to-all inhibition between units within a circuit copy, as well as some mechanism for self-excitation. Chemical positive feedback loops are probably more likely to be found than autapses, if this circuit is implemented in cortex. All-to-all inhibitory connectivity is an idealization that simplifies numerical and theoretical analyses, but is less likely to be found in the brain than, for example, dense but random connectivity. The winner-take-all computation is, however, thought to be ubiquitous in the brain, possibly underlying some aspects of attention and decision making. Indeed, there are several network designs that exhibit winner-take-all behavior, including networks with a single global inhibitory neuron instead of all-to-all inhibition [39]. It is therefore likely that the all-to-all requirement in this max circuit (section 3.1) can be relaxed, while maintaining the desired performance; the circuit we have presented relies on winner-take-all computations, but does not critically depend on the particular instantiation of winner-take-all shown in Figure 5.

In the tuning circuit discussed in section 3.2, divisive, “shunting” inhibition was chosen as the normalization mechanism, while the selection of an operating point in a nonlinear region of the units’ spike-rate transfer function was used to approximate the effect of the exponents in the normalized dot-product (11). Although several alternatives exist for both division and multiplication in spiking neural networks, we believe the choices we have made are among the most plausible in light of the design constraints. If one adopts the reasonable position that evolution has led to solutions which minimize both energy consumption and real-estate in the brain, the architecture discussed in section 3.2 can be seen to minimize the number of units necessary to accomplish normalization and exponentiation. It could also be that single cells are capable of performing normalized tuning using intrinsic mechanisms, in which case the number of cells in a cortical implementation is likely to be smaller than the number used in our circuit of simplistic integrate-and-fire units.

In order to improve the match with physiological data, we plan to extend our models to be more faithful to the morphology, biophysical properties and connectivity of different specific subtypes of cortical neurons in layers IV and II/III of somatosensory and visual cortices, utilizing the limited quantitative data available from other researchers [3, 4] and our own recordings.

In order to decide which of the presented models and variations, if any, best describes biophysical reality, new experiments are needed to help distinguish between the alternatives. It would be interesting to change the timing of the presented stimuli to test the temporal precision that is needed for the maximum and tuning effects to occur, and what kind of behavior can be observed outside of that precision window. In addition, experiments should try to extend from two to three or even more inputs (stimuli). This is difficult for visual stimuli because of the small receptive field sizes of the cells under investigation but it would help to tease apart different possible mechanisms as their behavior for more than two inputs can be significantly different. The rat vibrissae and barrel cortex system provides a suitable preparation for this task since it is much easier to stimulate more than two vibrissae at the same time. Finally, it is quite possible, despite the general skepticism of cortical physiologists, that the two operations described here may be performed by *circuits of mostly non-spiking neurons*, as proposed in [2]. Such models may need to be reconsidered and improved.

Acknowledgements

The authors would like to thank Timothe Masquelier for help with the experiments in section 3.3, and Thomas Serre, Christof Koch, Michael Okun and Ilan Lampl for helpful discussions and suggestions.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at

MIT, the Department of Brain & Cognitive Sciences, and the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from: the US Defense Advanced Research Projects Agency, US Office of Naval Research, US National Science Foundation, - National Institutes of Health (NIMH). Additional support was provided by: Daimler-Chrysler AG, Eastman Kodak Company, Honda Research Institute USA, Inc., Komatsu Ltd., Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, Toyota Motor Corporation, and the Eugene McDermott Foundation.

References

- [1] Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. *Nature Neurosci.* 2, 1019–1025 (1999)
- [2] Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T.: A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *AI Memo 2005-036 / CBCL Memo 259*, MIT CSAIL and CBCL, Cambridge, MA (2005)
- [3] Gawne, T.J., Martin, J.M.: Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophys.* 88, 1128–1135 (2002)
- [4] Lampl, I., Ferster, D., Poggio, T., Riesenhuber, M.: Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J. Neurophys* 92, 2704–2713 (2004)
- [5] Carandini, M., Heeger, D.J.: Summation and division by neurons in primate visual cortex. *Science* 264, 1333–1336 (1994)
- [6] Chelazzi, L., Duncan, J., Miller, E.K., Desimone, R.: Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophys* 80, 2918–2940 (1998)
- [7] Poggio, T., Reichardt, W., Hausen, W.: A neuronal circuitry for relative movement discrimination by the visual system of the fly. *Network* 68, 443–466 (1981)
- [8] Reichardt, W., Poggio, T., Hausen, K.: Figure-ground discrimination by relative movement in the visual system of the fly – II: Towards the neural circuitry. *Biol. Cyb.* 46, 1–30 (1983)
- [9] Grossberg, S.: Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics* 52, 213–257 (1973)
- [10] Abbott, L.F., Varela, J.A., Sen, K., Nelson, S.B.: Synaptic depression and cortical gain control. *Science* 275, 220–224 (1997)
- [11] Yu, A.J., Giese, M.A., Poggio, T.: Biophysically plausible implementations of the maximum operation. *Neural Comp.* 14(12), 2857–2881 (2002)
- [12] Kouh, M., Poggio, T.: A canonical cortical circuit for gaussian-like and max-like operations (submission, 2007)
- [13] Hung, C., Kreiman, G., Poggio, T., DiCarlo, J.: Fast read-out of object identity from macaque inferior temporal cortex. *Science* 310, 863–866 (2005)
- [14] Sclar, G., Maunsell, J.H., Lennie, P.: Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res.* 30(1), 1–10 (1990)
- [15] Lennie, P.: Single units and visual cortical organization. *Perception* 27, 889–935 (1998)

- [16] Woolsey, T A, der Loos, H.V.: The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex. The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Res.* 17(2), 205–242 (1970)
- [17] Vogels, T.P., Abbott, L.F.: Signal propagation and logic gating in networks of integrate-and-fire neurons. *J. Neurosci.* 25(46), 10786–10795 (2005)
- [18] Destexhe, A., Contreras, D.: Neuronal computations with stochastic network states. *Science* 314(5796), 85–90 (2006)
- [19] Poggio, T.: Stochastic linearization, central limit theorem and linearity in (nervous) “black-boxes”. In: *Atti di III Congresso Nazionale di Cibernetica E Biofisica*, pp. 349–358 (1975)
- [20] Douglas, R.J., Martin, K.A.C.: Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* 27, 419–451 (2004)
- [21] Moldakarimov, S., Rollenhagen, J.E., Olson, C.R., Chow, C.C.: Competitive dynamics in cortical responses to visual stimuli. *J. Neurophys.* 94(5), 3388–3396 (2005)
- [22] Okun, M., Lampl, I.: Synchronized excitation and inhibition during spontaneous and evoked response in the barrel cortex. In: *Computational and Systems Neuroscience*, Salt Lake City, UT (2007)
- [23] Pinto, D.J., Brumberg, J.C., Simons, D.J.: Circuit dynamics and coding strategies in rodent somatosensory cortex. *J. Neurophys.* 83(3), 1158–1166 (2000)
- [24] Wehr, M., Zador, A.M.: Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* 426(6965), 442–446 (2003)
- [25] Wilent, W.B., Contreras, D.: Dynamics of excitation and inhibition underlying stimulus selectivity in rat somatosensory cortex. *Nature Neurosci.* 8(10), 1364–1370 (2005)
- [26] Yoshimura, Y., Dantzker, J.L.M., Callaway, E.M.: Excitatory cortical neurons form fine-scale functional networks. *Nature* 433(7028), 868–873 (2005)
- [27] Yoshimura, Y., Callaway, E.M.: Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nature Neurosci.* 8(11), 1552–1559 (2005)
- [28] Swadlow, H.A.: Efferent neurons and suspected interneurons in S-1 vibrissa cortex of the awake rabbit: receptive fields and axonal properties. *J. Neurophys.* 62(1), 288–308 (1989)
- [29] Zhu, Y., Stornetta, R.L., Zhu, J.J.: Chandelier cells control excessive cortical excitation: characteristics of whisker-evoked synaptic responses of layer 2/3 nonpyramidal and pyramidal neurons. *J. Neurosci.* 24(22), 5101–5108 (2004)
- [30] Destexhe, A., Mainen, Z.F., Sejnowski, T.J.: Kinetic models of synaptic transmission. In: Segev, I., Koch, C. (eds.) *Methods in Neuronal Modeling: From Ions to Networks*, pp. 1–26. MIT Press, Cambridge (1998)
- [31] Hartline, H K, Ratliff, F.: Spatial summation of inhibitory influences in the eye of limulus, and the mutual interaction of receptor units. *Journal of General Physiology* 41(5), 1049–1066 (1957)
- [32] Amari, S.-I., Arbib, M.A.: Competition and cooperation in neural nets. In: Metzler, J. (ed.) *Systems Neuroscience*, pp. 119–165. Academic Press, London (1977)
- [33] Hahnloser, R.H.R., Seung, H., Slotine, J.J.: Permitted and forbidden sets in symmetric threshold-linear networks. *Neural Comp.* 15 (2003)
- [34] Jin, D., Seung, H.: Fast computation with spikes in a recurrent neural network. *Physical Review E* 65 (2002)
- [35] Oster, M., Liu, S.C.: Spiking inputs to a winner-take-all network. In: Weiss, Y., Platt, B.S. (eds.) *Advances in Neural Information Processing Systems*, vol. 18, pp. 1051–1058. MIT Press, Cambridge (2005)

- [36] Gerstner, W., Kistler, W.M.: Spiking Neuron Models. Cambridge University Press, Cambridge (2002)
- [37] Giese, M.A., Leopold, D.A.: Physiologically inspired model for the encoding of face spaces. *Neurocomputing*, 65–66 (2005)
- [38] Carandini, M., Heeger, D.J., Movshon, J.A.: Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* 17, 8621–8644 (1997)
- [39] Ermentrout, B.: Complex dynamics in winner-take-all neural nets with slow inhibition. *Neural Networks* 5(3), 415–431 (1992)

Cognitive Architectures and the Challenge of Cognitive Social Simulation

Ron Sun

Rensselaer Polytechnic Institute, Troy, NY 12180, USA

rsun@rpi.edu

<http://www.cogsci.rpi.edu/~rsun>

Abstract. A cognitive architecture is a domain-generic computational cognitive model that may be used for a broad, multiple-domain analysis of cognition and behavior. It embodies generic descriptions of cognition in computer algorithms and programs. Social simulation with multi-agent systems can benefit from incorporating cognitive architectures, as they provide a realistic basis for modeling individual agents (as argued in Sun 2001). In this article, an example of a cognitive architecture will be given, and its application to social simulation will be outlined.

1 What Is a Cognitive Architecture?

A cognitive *architecture* is a broadly-scoped, domain-generic computational cognitive model, capturing the essential structure and process of the mind, to be used for a broad, multiple-level, multiple-domain analysis of cognition and behavior (Newell 1990, Sun 2002).

An analogy is appropriate here: The architecture for a building consists of its overall framework and its overall design, as well as roofs, foundations, walls, windows, floors, and so on. Furniture and appliances can be easily rearranged and/or replaced and therefore they are not part of the architecture. By the same token, a cognitive architecture includes overall structures, essential divisions of modules, relations between modules, basic representations, essential algorithms, and a variety of other aspects (Sun 2004). In general, an architecture includes those aspects of a system that are relatively invariant across time, domains, and individuals. It deals with componential processes of cognition in a structurally and mechanistically well defined way.

In relation to understanding the human mind (i.e., cognitive science), a cognitive architecture provides a concrete framework for more detailed modeling of cognitive phenomena, through specifying essential structures, divisions of modules, relations between modules, and so on. Its function is to provide an essential framework to facilitate more detailed modeling and understanding of various components and processes of the mind. Research in computational cognitive modeling explores the essence of cognition and various cognitive functionalities through developing detailed, process-based understanding by specifying computational models of mechanisms and processes. It embodies descriptions of cognition in computer algorithms and programs. That is, it produces runnable

computational models. Detailed simulations are then conducted based on the computational models. In this enterprise, a cognitive architecture may be used for a broad, multiple-level, multiple-domain analysis of cognition.

In relation to building intelligent systems, a cognitive architecture specifies the underlying infrastructure for intelligent systems, which includes a variety of capabilities, modules, and subsystems. On that basis, application systems can be more easily developed. A cognitive architecture carries also with it theories of cognition and understanding of intelligence gained from studying the human mind. Therefore, the development of intelligent systems can be more cognitively grounded, which may be advantageous in many circumstances.

2 Why Are Cognitive Architectures Important?

While there are all kinds of cognitive architectures in existence, in this article I am concerned specifically with psychologically oriented cognitive architectures (as opposed to software engineering oriented “cognitive” architectures): their importance and their applications. Psychologically oriented cognitive architectures are particularly important because (1) they are “intelligent” systems that are cognitively realistic (relatively speaking) and therefore they are more human-like in many ways, (2) they shed new light on human cognition and therefore they are useful tools for advancing the science of cognition, (3) furthermore, they may (in part) serve as a foundation for understanding collective human behavior and social phenomena (to be detailed later). Let us examine the importance of this type of cognitive architecture.

For cognitive science, the importance of such cognitive architectures lie in the fact that they are enormously useful in terms of understanding the human mind. In understanding cognitive phenomena, the use of computational simulation on the basis of cognitive architectures forces one to think in terms of process, and in terms of detail. Instead of using vague, purely conceptual theories, cognitive architectures force theoreticians to think clearly. They are critical tools in the study of the mind. Researchers who use cognitive architectures must specify a cognitive mechanism in sufficient detail to allow the resulting models to be implemented on computers and run as simulations. This approach requires that important elements of the models be spelled out explicitly, thus aiding in developing better, conceptually clearer theories.

An architecture serves as an initial set of assumptions to be used for further modeling of cognition. These assumptions, in reality, may be based on either available scientific data (for example, psychological or biological data), philosophical thoughts and arguments, or ad hoc working hypotheses (including computationally inspired such hypotheses). An architecture is useful and important precisely because it provides a comprehensive initial framework for further modeling in a variety of task domains.

Cognitive architectures also provide a deeper level of explanation. Instead of a model specifically designed for a specific task (often in an ad hoc way), using a cognitive architecture forces modelers to think in terms of the mechanisms

and processes available within a generic cognitive architecture that are not specifically designed for a particular task, and thereby to generate explanations of the task that is not centered on superficial, high-level features of a task, that is, explanations of a deeper kind. To describe a task in terms of available mechanisms and processes of a cognitive architecture is to generate explanations centered on primitives of cognition as envisioned in the cognitive architecture, and therefore such explanations are deeper explanations. Because of the nature of such deeper explanations, this style of theorizing is also more likely to lead to unified explanations for a large variety of data and/or phenomena, because potentially a large variety of task data and phenomena can be explained on the basis of the same set of primitives provided by the same cognitive architecture. Therefore, using cognitive architectures leads to comprehensive theories of the mind (Newell 1990, Anderson and Lebiere 1998, Sun 2002).

On the other hand, for the fields of artificial intelligence and computational intelligence (AI/CI), the importance of cognitive architectures lies in the fact that they support the central goal of AI/CI—building artificial systems that are as capable as human beings. Cognitive architectures help us to reverse engineer the only truly intelligent system around—the human being, and in particular, the human mind. They constitute a solid basis for building truly intelligent systems, because they are well motivated by, and properly grounded in, existing cognitive research. The use of cognitive architectures in building intelligent systems may also facilitate the interaction between humans and artificially intelligent systems because of the similarity between humans and cognitively based intelligent systems.

3 Multiple Levels of Explanations

A broader perspective on the social and behavioral sciences may lead to a view of multiple “levels” of analysis encompassing multiple disciplines in the social and cognitive sciences. That is, a set of related disciplines, may be readily cast as a set of different levels of analysis, from the most macroscopic to the most microscopic. These different *levels* include: the sociological level, the psychological level, the componential level, and the physiological level. In other words, as has been argued in Sun et al (2005), one may view different disciplines as different levels of abstraction in the process of exploring essentially the same broad set of questions (cf. Newell 1990). See Figure 1.

level	object of analysis	type of analysis	model
1	inter-agent processes	social/cultural	collections of agent models
2	agents	psychological	individual agent models
3	intra-agent processes	componential	modular constr. of agent models
4	substrates	physiological	biological realization of modules

Fig. 1. A hierarchy of four levels

First of all, there is the sociological level, which includes collective behaviors of agents, inter-agent processes, sociocultural processes, social structures and organizations, as well as interactions between agents and their (physical and sociocultural) environments. Although studied extensively by sociology, anthropology, political science, and economics, this level has traditionally been very much ignored in cognitive science. Only recently, cognitive science, as a whole, has come to grip with the fact that cognition is, at least in part, a sociocultural process.¹

The next level is the psychological level, which covers individual experiences, individual behaviors, individual performance, as well as beliefs, concepts, and skills employed by individual agents. In relation to the sociological level, the relationship of individual beliefs, concepts, and skills with those of the society and the culture, and the processes of change of these beliefs, concepts, and skills, independent of or in relation to those of the society and the culture, may be investigated (in inter-related and mutually influential ways). At this level, one may examine human behavioral data, compared with models (which may be based on cognitive architectures) and with insights from the sociological level and details from the lower levels.

The third level is the componential level. At this level, one studies and models cognitive agents in terms of components (e.g., in the form of a cognitive architecture), with the theoretical language of a particular paradigm (for example, symbolic computation or connectionist networks, or their combinations thereof). At this level, one may specify computationally an overall architecture consisting of multiple components therein. One may also specify some essential computational processes of each component as well as essential connections among components. That is, one imputes a computational process onto a cognitive function. Ideas and data from the psychological level (that is, the psychological constraints from above), which bear significantly on the division of components and their possible implementations, are among the most important considerations. This level may also incorporate biological/physiological facts regarding plausible divisions and their implementations (that is, it can incorporate ideas from the next level down — the physiological level, which offers the biological constraints). This level results in *mechanisms* (though they are computational and thus somewhat abstract compared with physiological-level details).²

Although this level is essentially in terms of intra-agent processes, computational models (cognitive architectures) developed therein may be used to capture processes at higher levels, including interaction at a sociological level whereby multiple individuals are involved. This can be accomplished, for example, by examining interactions of multiple copies of individual agent models (based on the same cognitive architecture) or those of different individual agent models (based on different cognitive architectures). One may use computation as a means for constructing cognitive architectures at a sub-agent level (the componential level),

¹ See Sun (2001) for a more detailed argument for the relevance of sociocultural processes to cognition and vice versa.

² The importance of this level has been argued for, for example, in Anderson and Lebiere (1998), and Sun et al (2004).

but one may go up from there to the psychological level and to the sociological level (see the discussion regarding mixing levels in Sun et al 2005).

The lowest level of analysis is the physiological level, that is, the biological substrate, or the biological implementation, of computation. This level is the focus of a range of disciplines including biology, physiology, computational neuroscience, cognitive neuroscience, and so on. Although biological substrates are not our main concern here, they may nevertheless provide useful input as to what kind of computation is likely employed and what a plausible architecture (at a higher level) should be like. The main utility of this level is to facilitate analysis at higher levels, that is, analysis using low-level information to narrow down choices in selecting computational architectures as well as choices in implementing componential computation.³

In this enterprise of multiple levels in cognitive and social sciences, a cognitive architecture may serve as a centerpiece, tying together various strands of research. It may serve this purpose due to the comprehensiveness of its functionality and the depth with which it has been developed (at least for some psychologically oriented/grounded cognitive architectures). Thus, detailed mechanisms are developed within a cognitive architecture, which may be tied to low-level cognitive processes, while a cognitive architecture as a whole may function at a very high level of cognitive and social processes.

4 An Example of a Cognitive Architecture

4.1 An Overview

Below I will describe a cognitive architecture CLARION. It has been described extensively in a series of previous papers, including Sun and Peterson (1998), Sun et al (2001), and Sun (2002, 2003).

CLARION is an integrative architecture, consisting of a number of distinct subsystems, with a dual representational structure in each subsystem (implicit versus explicit representations). Its subsystems include the action-centered subsystem (the ACS), the non-action-centered subsystem (the NACS), the motivational subsystem (the MS), and the meta-cognitive subsystem (the MCS). The role of the action-centered subsystem is to control actions, regardless of whether the actions are for external physical movements or for internal mental operations. The role of the non-action-centered subsystem is to maintain general knowledge, either implicit or explicit. The role of the motivational subsystem is to provide underlying motivations for perception, action, and cognition, in terms of providing impetus and feedback (e.g., indicating whether outcomes are satisfactory or not). The role of the meta-cognitive subsystem is to monitor, direct, and

³ Work at this level is basically the reverse-engineering of biological systems. In such a case, what needs to be done is to pinpoint the most basic primitives that are of relevance to the higher-level functioning that is of interest. (While many low-level details are highly significant, clearly not all low-level details are significant or even relevant.) After identifying proper primitives, one may study processes that involve those primitives, in mechanistic/computational terms.

modify the operations of the action-centered subsystem dynamically as well as the operations of all the other subsystems.

Each of these interacting subsystems consists of two levels of representation (i.e., a dual representational structure): Generally, in each subsystem, the top level encodes explicit knowledge and the bottom level encodes implicit knowledge. The distinction of implicit and explicit knowledge has been amply argued for before (see Reber 1989, Cleeremans et al 1998, Sun 2002). The two levels interact, for example, by cooperating in actions, through a combination of the action recommendations from the two levels respectively, as well as by cooperating in learning through a bottom-up and a top-down process (to be discussed below). Essentially, it is a dual-process theory of mind. See Figure 2.

4.2 Some Details

The Action-Centered Subsystem. First, let us focus on the action-centered subsystem (the ACS) of CLARION. The operation of the action-centered subsystem may be described as follows:

1. Observe the current state x .
2. Compute in the bottom level the Q-values of x associated with each of all the possible actions a_i 's: $Q(x, a_1), Q(x, a_2), \dots, Q(x, a_n)$.
3. Find out all the possible actions ($b_1, b_2, \dots, b_m$) at the top level, based on the input x (sent up from the bottom level) and the rules in place.
4. Compare or combine the values of the selected a_i 's with those of b_j 's (sent down from the top level), and choose an appropriate action b .
5. Perform the action b , and observe the next state y and (possibly) the reinforcement r .
6. Update Q-values at the bottom level in accordance with the *Q-Learning-Backpropagation* algorithm
7. Update the rule network at the top level using the *Rule-Extraction-Refinement* algorithm.
8. Go back to Step 1.

In the bottom level of the action-centered subsystem, implicit reactive routines are learned: A Q-value is an evaluation of the "quality" of an action in a given state: $Q(x, a)$ indicates how desirable action a is in state x (which consists of some sensory input). The agent may choose an action in any state based on Q-values. To acquire the Q-values, the *Q-learning* algorithm (Watkins 1989) may be used, which is a reinforcement learning algorithm. It basically compares the values of successive actions and adjusts an evaluation function on that basis. It thereby develops sequential behaviors (Sun 2003).

The bottom level of the action-centered subsystem is modular; that is, a number of small neural networks co-exist each of which is adapted to specific modalities, tasks, or groups of input stimuli. This coincides with the modularity claim (Fodor 1983, Hirschfield and Gelman 1994) that much processing is done by limited, encapsulated (to some extent), specialized processors that are highly efficient. These modules can be developed in interacting with the world (computationally through various decomposition methods; e.g., Sun and Peterson

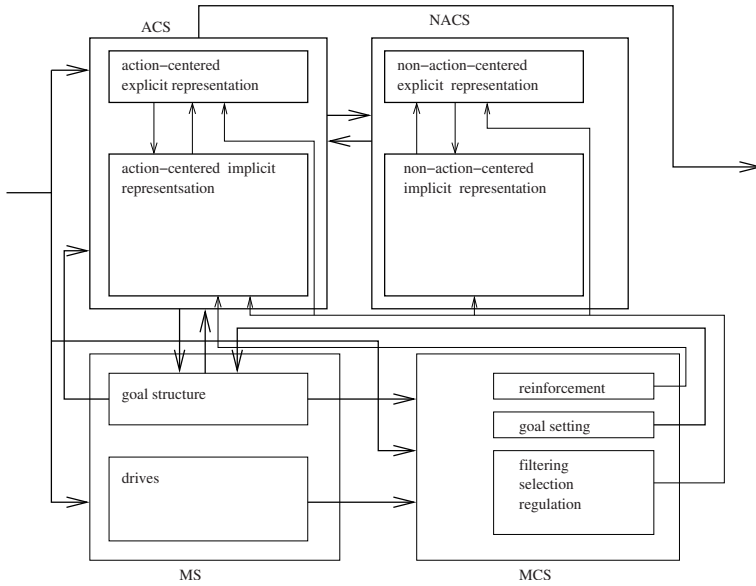


Fig. 2. The CLARION Architecture

1999). Some of them, however, are formed evolutionarily, that is, given a priori to agents, reflecting their hardwired instincts and propensities (Hirschfield and Gelman 1994).

In the top level of the action-centered subsystem, explicit conceptual knowledge is captured in the form of rules. See Sun (2003) for details. There are many ways in which explicit knowledge may be learned, including independent hypothesis-testing learning and “bottom-up learning” as discussed below.

Autonomous Generation of Explicit Conceptual Structures. Humans are generally able to learn implicit knowledge through trial and error, without necessarily utilizing a priori knowledge. On top of that, explicit knowledge can be acquired also from on-going experience in the world, through the mediation of implicit knowledge (i.e., bottom-up learning; see Sun 2002). The basic process of bottom-up learning is as follows: if an action implicitly decided by the bottom level is successful, then the agent extracts an explicit rule that corresponds to the action selected by the bottom level and adds the rule to the top level. Then, in subsequent interaction with the world, the agent verifies the extracted rule by considering the outcome of applying the rule: if the outcome is not successful, then the rule should be made more specific and exclusive of the current case; if the outcome is successful, the agent may try to generalize the rule to make it more universal (Michalski 1983).⁴ After explicit rules have been learned, a variety of explicit reasoning methods may be used. Learning explicit conceptual

⁴ The detail of the bottom-up learning algorithm can be found in Sun and Peterson (1998).

representation at the top level can also be useful in enhancing learning of implicit reactive routines at the bottom level (e.g., Sun et al 2001).

Assimilation of Externally Given Conceptual Structures. Although CLARION can learn even when no a priori or externally provided knowledge is available, it can make use of it when such knowledge is available. To deal with instructed learning, externally provided knowledge, in the forms of explicit conceptual structures such as rules, plans, categories, and so on, can (1) be combined with existent conceptual structures at the top level (i.e., internalization), and (2) be assimilated into implicit reactive routines at the bottom level (i.e., assimilation). This process is known as top-down learning. See Sun (2003) for more details.

The Non-Action-Centered Subsystem. The non-action-centered subsystem (NACS) may be used to represent general knowledge about the world, for performing various kinds of memory retrievals and inferences. Note that the non-action-centered subsystem is under the control of the action-centered subsystem (through its actions).

At the bottom level, “associative memory” networks encode non-action-centered implicit knowledge. Associations are formed by mapping an input to an output. The regular backpropagation learning algorithm can be used to establish such associations between pairs of inputs and outputs (Rumelhart et al 1986).

On the other hand, at the top level of the non-action-centered subsystem, a general knowledge store encodes explicit non-action-centered knowledge (cf. Sun 1994). In this network, chunks are specified through dimensional values. A node is set up in the top level to represent a chunk. The chunk node connects to its corresponding features represented as individual nodes in the bottom level of the non-action-centered subsystem. Additionally, links between chunks encode explicit associations between pairs of chunks, known as associative rules. Explicit associative rules may be formed (i.e., learned) in a variety of ways (Sun 2003).

In addition to applying associative rules, similarity-based reasoning may be employed in the non-action-centered subsystem. During reasoning, a known (given or inferred) chunk may be automatically compared with another chunk. If the similarity between them is sufficiently high, then the latter chunk is inferred (see Sun 2003 for details).

As in the action-centered subsystem, top-down or bottom-up learning may take place in the non-action-centered subsystem, either to extract explicit knowledge in the top level from the implicit knowledge in the bottom level or to assimilate explicit knowledge of the top level into implicit knowledge in the bottom level.

The Motivational and the Meta-Cognitive Subsystem. The motivational subsystem (the MS) is concerned with drives and their interactions (Toates 1986), which leads to actions. It is concerned with why an agent does what it does. Simply saying that an agent chooses actions to maximize gains, rewards, or payoffs leaves open the question of what determines these things. The

relevance of the motivational subsystem to the action-centered subsystem lies primarily in the fact that it provides the context in which the goal and the payoff of the action-centered subsystem are set. It thereby influences the working of the action-centered subsystem, and by extension, the working of the non-action-centered subsystem.

A bipartite system of motivational representation is in place in CLARION. The explicit goals (such as “finding food”) of an agent (which is tied to the working of the action-centered subsystem) may be generated based on internal drive states (for example, “being hungry”). (See Sun 2003 for details.)

Beyond low-level drives (concerning physiological needs), there are also higher-level drives. Some of them are primary, in the sense of being “hard-wired”. For example, Maslow (1987) developed a set of these drives in the form of a “need hierarchy”. While primary drives are built-in and relatively unalterable, there are also “derived” drives, which are secondary, changeable, and acquired mostly in the process of satisfying primary drives.

The meta-cognitive subsystem (the MCS) is closely tied to the motivational subsystem. The meta-cognitive subsystem monitors, controls, and regulates cognitive processes for the sake of improving cognitive performance (Nelson 1993). Control and regulation may be in the forms of setting goals for the action-centered subsystem, setting essential parameters of the action-centered subsystem and the non-action-centered subsystem, interrupting and changing on-going processes in the action-centered subsystem and the non-action-centered subsystem, and so on. Control and regulation can also be carried out through setting reinforcement functions for the action-centered subsystem. All of the above can be done on the basis of drive states and goals in the motivational subsystem. The meta-cognitive subsystem is also made up of two levels: the top level (explicit) and the bottom level (implicit).

5 Applications to Social Simulation

One application of CLARION to social simulation is in understanding organizational decision making and the interaction between organizational structures and cognitive factors in affecting organizational decision making (Sun and Naveh 2004).

In terms of organizational structures, there are two major types: (1) teams, in which agents act autonomously, individual decisions are treated as votes, and the organizational decision is the majority decision; and (2) hierarchies, which are characterized by agents organized in a chain of command, such that information is passed from subordinates to superiors, and the decision of a superior is based solely on the recommendations of his/her subordinates. In addition, organizations are distinguished by the structure of information accessible by each agent. Two varieties of information access are: (1) distributed access, in which each agent sees a different subset of attributes (no two agents see the same subset of attributes), and (2) blocked access, in which several agents see exactly the same subset of attributes.

Agent/Org.	Team(B)	Team(D)	Hierarchy(B)	Hierarchy(D)
Human	50.0	56.7	46.7	55.0
Radar-Soar	73.3	63.3	63.3	53.3
CORP-P-ELM	78.3	71.7	40.0	36.7
CORP-ELM	88.3	85.0	45.0	50.0
CORP-SOP	81.7	85.0	81.7	85.0

Fig. 3. Human and simulation data for the organizational decision task. D indicates distributed information access, while B indicates blocked information access. All numbers are percent correct.

Several simulation models were considered in Carley et al (1998). The experiments by Carley et al (1998) were done in a 2 x 2 fashion (organization x information access). In addition, human data for the experiment were compared to the results of the four models (Carley et al 1998).⁵ See Figure 3.

In their work, the agent models used were very simple, and the “intelligence” level in these models was low. Moreover, learning in these simulations was rudimentary: there was no complex learning process as one might observe in humans. With these shortcomings in mind, it is worthwhile to undertake a simulation that involves more complex agent models that more accurately capture human performance. Moreover, with the use of more cognitively realistic agent models, one may investigate individually the importance of different cognitive capacities and process details in affecting organizational performance (see Sun and Naveh 2004).

Hence, a simulation with CLARION used for modeling individual agents in an organization was conducted. The results (see Figure 4) closely accord with the patterns of the human data, with teams outperforming hierarchal structures, and distributed access proving superior to blocked access. Also, as in humans, performance is not grossly skewed towards one condition or the other, but is

Agent/Org.	Team(B)	Team(D)	Hierarchy(B)	Hierarchy(D)
Human	50.0	56.7	46.7	55.0
CLARION	53.2	59.3	45.0	49.4

Fig. 4. Simulation data for agents running for 3,000 cycles. The human data from Carley et al (1998) are reproduced here. Performance of CLARION is computed as percent correct over the last 1,000 cycles.

⁵ Among them, CORP-ELM produced the most probable classification based on an agent’s own experience, CORP-P-ELM stochastically produced a classification in accordance with the estimate of the probability of each classification based on the agent’s own experience, CORP-SOP followed organizationally prescribed standard operating procedure (which involved summing up the values of the attributes available to an agent) and thus was not adaptive, and Radar-Soar was a (somewhat) cognitive model built in Soar, which is based on explicit, elaborate search in problem spaces (Rosenbloom et al 1991).

roughly comparable across all conditions, unlike some of the simulation results from Carley et al (1998). The match with the human data is far better than in the simulations conducted in Carley et al (1998). The better match is due, at least in part, to a higher degree of cognitive realism in our simulation. See Sun and Naveh (2004) for further details, including the interesting effects of varying cognitive parameters.

Another application of CLARION to social simulation is in capturing and explaining the essential process of publication in academic science and its relation to cognitive processes (Naveh and Sun 2006). Science develops in certain ways. In particular, it has been observed that the number of authors contributing a certain number of articles to a scientific journal follows a highly skewed distribution, corresponding to an inverse power law. In the case of scientific publication, the tendency of authorship to follow such a distribution was known as Lotka's law. Simon (1957) developed a simple stochastic process for approximating Lotka's law. One of the assumptions underlying this process is that the probability that a paper will be published by an author who has published i articles is equal to a/i^k , where a is a constant of proportionality. Using Simon's work as a starting point, Gilbert (1997) attempted to model Lotka's law. He obtains his simulation data based on some very simplified assumptions and a set of mathematical equations. To a significant extent, Gilbert's model is not cognitively realistic. The model assumes that authors are non-cognitive and interchangeable; it therefore neglects a host of cognitive phenomena that characterize scientific inquiry (e.g., learning, creativity, evolution of field expertise, etc.).

Using a more cognitively realistic model, one can address some of these omissions, as well as exploring other emergent properties of a cognitively based model and their correspondence to real-world phenomena. The results of the simulation based on CLARION are shown in Figures 5 and 6, along with results (reported by Simon 1957) for *Chemical Abstracts* and *Econometrica*, and estimates obtained from previous simulations by Simon (1957) and Gilbert (1997). The figures in the tables indicate number of authors contributing to each journal, by number of papers each has published.

The CLARION simulation data for the two journals could be fit to the power curve $f(i) = a/i^k$, resulting in an excellent match. The results of the curve fit are shown in Figure 7, along with correlation and error measures.

Note that, in our simulation, the number of papers per author reflected the cognitive ability of an author, as opposed to being based on auxiliary assumptions such as those made by Gilbert (1997). This explains, in part, the greater divergence of our results from the human data: whereas Gilbert's simulation consists of equations selected to match the human data, our approach relies on much more detailed and lower-level mechanisms—namely, a cognitive agent model that is generic rather than task-specific. The result of the CLARION based simulation is therefore emergent, and not a result of specific and direct attempts to match the human data. That is, we put more distance between mechanisms and outcomes, which makes it harder to obtain a match with the human data. Thus,

# of Papers	Actual	Simon's estimate	Gilbert's simulation	CLARION simulation
1	3991	4050	4066	3803
2	1059	1160	1175	1228
3	493	522	526	637
4	287	288	302	436
5	184	179	176	245
6	131	120	122	200
7	113	86	93	154
8	85	64	63	163
9	64	49	50	55
10	65	38	45	18
11 or more	419	335	273	145

Fig. 5. Number of authors contributing to *Chemical Abstracts*

# of Papers	Actual	Simon's estimate	Gilbert's simulation	CLARION simulation
1	436	453	458	418
2	107	119	120	135
3	61	51	51	70
4	40	27	27	48
5	14	16	17	27
6	23	11	9	22
7	6	7	7	17
8	11	5	6	18
9	1	4	4	6
10	0	3	2	2
11 or more	22	25	18	16

Fig. 6. Number of authors contributing to *Econometrica*

Journal	a	k	Pearson R	R-square	RMSE
CA	3806	1.63	0.999	0.998	37.62
E	418	1.64	0.999	0.999	4.15

Fig. 7. Results of fitting CLARION data to power curves. CA stands for Chemical Abstracts and E stands for Econometrica.

the fact that we were able to match the human data reasonably well shows the power of our cognitive architecture based approach.

6 The Challenges of Cognitive Social Simulation

An important development in the social sciences has been that of agent-based social simulation (ABSS). This approach consists of instantiating a population of agents, allowing the agents to run, and observing the interactions between them. The use of agent-based social simulation as a means for computational study of societies mirrors the development of cognitive architectures in cognitive science. Thus, it is time to tackle sociality and social processes through cognitive architectures. So far, however, the two fields of social simulation and cognitive architectures have developed separately from each other (with some exceptions; e.g., Sun 2006). Thus, most of the work in social simulation assumes very rudimentary cognition on the part of the agents.

The two fields of social simulation and cognitive architectures can be profitably integrated. This is an important challenge. As has been argued before (Sun and Naveh 2004), social processes ultimately rest on the choices and decisions of individuals, and thus understanding the mechanisms of individual cognition can lead to better theories describing the behavior of aggregates of individuals. Although most agent models in social simulation have been extremely simple, a more realistic cognitive agent model, incorporating realistic tendencies, inclinations and capabilities of individual cognitive agents can serve as a more realistic basis for understanding the interaction of individuals.⁶

At the same time, by integrating social simulation and cognitive modeling, one can arrive at a better understanding of individual cognition. Traditional approaches to cognitive modeling have largely ignored the potentially decisive effects of social aspects of cognition (including norms and so on). By modeling cognitive agents in a social context, one can learn more about the sociocultural processes that influence individual cognition.

The most fundamental challenge in this regard is to develop better ways of conducting detailed social simulation based on cognitive architectures as basic building blocks. This is not an easy task. Although some initial work has been done (e.g., Sun and Naveh 2004, Sun 2006), much more work is needed.

One specific challenge is how to enhance cognitive architectures for the purpose of accounting for sociality in individual cognitive agents. There are many questions in this regard. For example, what are the characteristics of a proper cognitive architecture for modeling the interaction of cognitive agents? What additional sociocultural representations (for example, “motive”, “obligation”, or “norm”) are needed in cognitive modeling of multi-agent interaction? See, for example, Sun (2006) for further discussions.

⁶ Although some cognitive details may ultimately prove to be irrelevant, this cannot be determined *a priori*, and thus simulations are useful in determining which aspects of cognition can be safely abstracted away.

There is also the challenge of computational complexity and thus scalability that needs to be addressed. Social simulation could involve a large number of agents, up to thousands. Computational complexity is thus already high, even without involving cognitive architectures as agent models. To incorporate cognitive architectures into social simulation, one has to deal with a great deal of added complexity. Thus, scalability is a significant issue.

7 Concluding Remarks

We can expect that the field of cognitive architectures will have a profound impact on cognitive science as well as on social simulations, both in terms of better understanding cognition and in terms of better understanding sociality. As such, it should be considered an important future research direction, and correspondingly, a significant amount of collective research effort should be put into it.

Acknowledgments

This work was carried out while the author was supported in part by ARI grants DASW01-00-K-0012 and W74V8H-04-K-0002 (to Ron Sun and Bob Mathews). Thanks are due to Ning Zhong for his invitation to contribute this article.

References

- Anderson, J., Lebiere, C.: *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ (1998)
- Carley, K., Prietula, M., Lin, Z.: Design versus cognition: The interaction of agent cognition and organizational design on organizational performance. *Journal of Artificial Societies and Social Simulation* 1(3) (1998)
- Cleeremans, A., Destrebecqz, A., Boyer, M.: Implicit learning: News from the front. *Trends in Cognitive Sciences* 2(10), 406–416 (1998)
- Fodor, J.: *The Modularity of Mind*. MIT Press, Cambridge (1983)
- Gilbert, N.: A simulation of the structure of academic science. *Sociological Research Online* 2(2) (1997) Available online at <http://www.socresonline.org.uk/socresonline/2/2/3.html>
- Hirschfeld, L., Gelman, S. (eds.): *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge University Press, Cambridge (1994)
- Marr, D.: *Vision*. W.H. Freeman, New York (1982)
- Maslow, A.: *Motivation and Personality*, 3rd edn. Harper and Row, New York (1987)
- Naveh, I., Sun, R.: A cognitively based simulation of academic science. *Computational and Mathematical Organization Theory* 12, 313–337 (2006)
- Nelson, T. (ed.): *Metacognition: Core Readings*, Allyn and Bacon (1993)
- Newell, A.: *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA (1990)
- Simon, H.: *Models of Man, Social and Rational*. Wiley, New York (1957)

- Sun, R.: Integrating Rules and Connectionism for Robust Commonsense Reasoning. John Wiley and Sons, New York (1994)
- Sun, R.: Cognitive science meets multi-agent systems: A prolegomenon. *Philosophical Psychology* 14(1), 5–28 (2001)
- Sun, R.: Duality of the Mind. Lawrence Erlbaum, Mahwah (2002)
- Sun, R.: A Tutorial on CLARION. Technical report, Cognitive Science Department, Rensselaer Polytechnic Institute (2003),
<http://www.cogsci.rpi.edu/~rsun/sun.tutorial.pdf>
- Sun, R.: Desiderata for cognitive architectures. *Philosophical Psychology* 17(3), 341–373 (2004)
- Sun, R.: Prolegomena to integrating cognitive modeling and social simulation. In: Sun, R. (ed.) *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, Cambridge University Press, Cambridge (2006)
- Sun, R.: Introduction to computational cognitive modeling. In: Sun, R. (ed.) *The Cambridge Handbook of Computational Cognitive Modeling*, Cambridge University Press, Cambridge (2007)
- Sun, R., Coward, L.A., Zenzen, M.J.: On levels of cognitive modeling. *Philosophical Psychology* 18(5), 613–637 (2005)
- Sun, R., Naveh, I.: Simulating organizational decision-making using a cognitively realistic agent model. *Journal of Artificial Societies and Social Simulation* 7(3) (2004),
<http://jasss.soc.surrey.ac.uk/7/3/5.html>
- Sun, R., Peterson, T.: Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks* 9(6), 1217–1234 (1998)
- Sun, R., Peterson, T.: Multi-agent reinforcement learning: Weighting and partitioning. *Neural Networks* 12(4-5), 127–153 (1999)
- Watkins, C.: Learning with Delayed Rewards. Ph.D Thesis, Cambridge University, Cambridge, UK (1989)

ACT-R Meets fMRI

Yulin Qin^{1,2}, Daniel Bothell², and John R. Anderson²

¹ Department of Psychology and Behavior Science, Zhejiang University
Hangzhou, Zhejiang, China

² Department of Psychology, Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Abstract. ACT-R (Adaptive Control of Thought - Rational) is a theory and computational model of human cognitive architecture. It consists of a set of modules with their own buffers, each devoted to processing a different kind of information. A production rule in the core production system can be fired based on the chunks in these buffers and then it changes the chunks in the buffer of the related modules or the state of the related modules, which may leads to fire a new production rule and so on to generate the cognitive behavior. It has successfully predicted and explained a broad range of cognitive psychological phenomena and found applications in the human-computer interface and other areas (see <http://act-r.psy.cmu.edu>) and may have potential applications in Web intelligence.

In recent years, a series of fMRI experiments have been performed to explore the neural basis of cognitive architecture and to build a two-way bridge between the information processing model and fMRI. The patterns of the activations of brain areas corresponding to the buffers of the major modules in ACT-R were highly consistent across these experiments; and ACT-R has successfully predicted the Blood Oxygenation Level-Depend (BOLD) effect in these regions. The approach of ACT-R meets fMRI may shed light on the research of Web Intelligence (WI) meets Brain Informatics (BI).

1 ACT-R

ACT-R (Adaptive Control of Thought - Rational) is a theory and computational model of human cognitive architecture. As a theory, it proposes the systematical hypothesis on the basic structure of human cognitive system and functions of these structures in information processing to generate the human cognitive behavior; as a computational model, it offers a computer software system for the development of computational models to quantitatively simulate and predict the human behavior for a wide range of cognitive tasks.

1.1 Two Kinds of Knowledge

There are two kinds of knowledge represented in ACT-R – declarative knowledge and procedural knowledge. Declarative knowledge corresponds to things we are aware we know and can usually describe to others. Examples of declarative

knowledge include "George Washington was the first president of the United States" and "An atom is like the solar system". Procedural knowledge is the knowledge which we display in our behavior but which we are not conscious of. For instance, no one can describe the rules by which we speak a language and yet we do. In ACT-R declarative knowledge is represented in structures called chunks whereas procedural knowledge is represented as rules called productions. Thus chunks and productions are the basic building blocks of an ACT-R model.

Chunks in ACT-R. In ACT-R a chunk is defined by its type and its slots. One can think of chunk-types as categories (e.g., birds) and slots as category attributes (e.g., color or size). A chunk also has a name which can be used to reference it, but the name is not considered to be a part of the chunk itself. Below is a representation of a chunk that encodes the fact that $4 + 3 = 7$. The name of the chunk is Fact3+4. The isa slot is special and specifies the type of the chunk which is addition-fact in this example, and the other slots are addend1, addend2, and sum.

```
Fact3 + 4
  isa addition-fact
  addend1 three
  addend2 four
  sum seven
```

Productions in ACT-R. A production rule is a statement of a particular contingency that controls behavior. An example might be

```
IF the goal is to add two digits d1 and d2 in a column
  and  $d1 + d2 = d3$ 
THEN set as a subgoal to write d3 in the column
```

The condition of a production rule (the IF part) consists of a specification of the chunks in various buffers. The action of a production rule (the THEN part) consists of modifications of the chunks in the buffers, requests for other chunks to be placed into the buffers, and/or requests for other actions to be taken.

Representations of Chunks and Productions in ACT-R. In previous sessions there are informal English specifications of chunks and production rules. This session shows examples of them represented in ACT-R. The examples are from an ACT-R model for counting numbers.

All ACT-R commands are Lisp functions and therefore are specified in parentheses. The first term after the left parenthesis is the command name. That is followed by the details of the item and then a right parenthesis. In this section we will show how to use the commands to create the knowledge representations in ACT-R.

An ACT-R model begins from creating the chunk type. To create a new type of chunk, one needs to specify a frame for the chunk using the chunk-type

command. This requires specifying the name of the chunk type and the names of the slots that it will have. The general chunk type specification looks like this:

```
(chunk-type name slot-name-1 slot-name-2 ... slot-name-n)
```

To model count, the chunk types can be created as:

```
(chunk-type count-order first second)
(chunk-type count-from start end count state)
```

The command to create a set of chunks and automatically add them to declarative memory is `add-dm`. It takes any number of chunk specifications as its arguments. If a slot is without initial value, the value of this slot is `nil` (as in Lisp), and this slot can be missing in the chunk, as the slots of count and state in the goal chunk here.

```
(add-dm
 (b ISA count-order first 1 second 2)
 (c ISA count-order first 2 second 3)
 (d ISA count-order first 3 second 4)
 (e ISA count-order first 4 second 5)
 (f ISA count-order first 5 second 6)
 (first-goal ISA count-from start 2 end 5))
```

The general form of a production rule is

```
(P Production-Name
  condition part
  ==>
  action part
)
```

The production that counts from one number to the next is as follows:

(P counting-example	English Description
<pre>=goal> ISA count state incrementing count =num1</pre>	If the goal is to count the state slot is incrementing the number in count slot is called =num1 and the chunk in the retrieval buffer
<pre>=retrieval> ISA count-order first =num1 second =num2</pre>	is of type count-order the first slot has the value =num1 and the second slot has a value we will call =num2
<pre>==></pre>	Then

=goal>		change the goal
count	=num2	to continue counting from =num2
+retrieval>		and request a retrieval
ISA	count-order	of a count-order chunk to
first	=num2	find the number that follows =num2
)		

The condition of the preceding production specifies a pattern to match to the goal buffer (indicated by =goal>) and a pattern to match to the retrieval buffer (indicated by =retrieval>). The slot values start with "=", such as =num1 and =num2, are variables. +retrieval> means requesting a retrieval in declarative memory.

1.2 Modules in ACT-R

ACT-R consists of a set of modules with their own buffers, each devoted to processing a different kind of information. Besides the declarative memory module mentioned above, there is a visual module for identifying objects in the visual field, a manual module for controlling the hands, a goal module for keeping track of current goals and intentions and other modules. The information in these modules is largely encapsulated and the modules communicate only by firing the productions based on the chunks in their buffers. Within a module and among different modules, the information processing can go on in parallel and asynchronously. However, there are two sequential processing restrictions in ACT-R: Only one chunk can be in the buffer of any module at any time and there is only one production rule can be fire at each processing cycle.

The Vision Module. Many tasks involve interacting with visible stimuli and the vision module provides a model with a means for acquiring visual information. It is designed as a system for modeling visual attention. It assumes that there are lower level perceptual processes that generate the representations with which it operates, but does not model those perceptual processes in detail. It includes some default mechanisms for parsing text and other simple visual features from a window and provides an interface that one can use to extend it when necessary.

The vision module has two buffers. There is a visual buffer that can hold a chunk that represents an object in the visual scene and a visual-location buffer that holds a chunk which represents the location of an object in the visual scene. As with all modules, it also responds to queries of the buffers about the state of the module. One should always check to make sure that the visual buffer is free before making any requests to avoid jamming it.

?visual>

state free

is true when the buffer is available.

The Motor Module. When we speak of motor actions in ACT-R we are only concerned with hand movements. The hand movement can be the finger presses

at a keyboard or other devices. The hands can also be used to move a mouse or other device.

The buffer for interacting with the motor module is called the manual buffer. Unlike other buffers, however, the manual buffer never holds a chunk. It is used only to issue commands and to query the state of the motor module. The manual buffer is used to request actions be performed by the hands. As with the vision module, one should always check to make sure that the motor module is free before making any requests to avoid jamming it.

1.3 A Simple Example of ACT-R Model

The example shown in this session includes the processes of reading the screen to get the task of finding the answer of $3 + 4$, retrieving the arithmetic fact, $3 + 4 = 7$, to solve the problem, and then moving the hand to press the key of 7 in the keyboard to give the output of the system. This example is, of course, very simple, but involves the major steps of the interaction between human and the Web. We will give the ACT-R model with explanations in this session, but put other detailed issues in the appendix.

```
;; Define Chunk-types (Comments in Lisp begin with ;)

(chunk-type addition-fact addend1 addend2 sum)
(chunk-type arithmetic-task argu1 operator argu2 state)

;; Create Chunks in declarative memory

(add-dm
  (fact34 ISA addition-fact addend1 "3" addend2 "4" sum "7")
  (first-goal ISA arithmetic-task state start)
)
```

Note again that if a slot is without initial value, this slot can be missing in the chunk definition (as of slots `argu1`, `argu2`, and `operator` in the `first-goal` chunk here).

```
;; Productions for receiving visual stimulus
```

The first three productions form a typical process of receiving visual stimulus in ACT-R models. The first production is to find the location of an object in the visual scene, the second one is to switch attention to the object, and the third is to harvest the object.

```
(P find-unattended-txt
  =goal>
  ISA          arithmetic-task
  state        start
```

```

==>
+visual-location>
  ISA          visual-location
  screen-x     lowest
  :attended    nil
=goal>
  state        find-location
)

```

This production makes a request of the visual-location buffer and changes the goal state to find-location. The visual-location request asks the vision module to find the location of an object in its visual scene (which is the experiment window for this model) that meets the specified requirements (in this production it is the leftmost (with lowest x coordinate) unattended (:attended nil) object), to build a chunk to represent the location of that object if one exists, and place that chunk in the visual-location buffer.

```

(P attend-txt
=goal>
  ISA          arithmetic-task
  state        find-location
=visual-location>
  ISA          visual-location
?visual>
  state        free
==>
+visual>
  ISA          move-attention
  screen-pos   =visual-location
=goal>
  state        attend
)

```

If the goal state is find-location, there is a visual-location in the visual-location buffer, and the vision module is not currently active (state free), then a query is made of the visual buffer which isa move-attention and it specifies the screen-pos[ition] as the visual location in the visual-location buffer. A request of the visual buffer for a move-attention is a request for the vision module to move its attention to the specified location, encode the object that is there, and place that object into the visual buffer as a chunk.

```

(P encode-first-txt
=goal>
  ISA          arithmetic-task
  state        attend
)

```



```

    argu1      nil
  =visual>
    ISA        text
    value      =letter
  ==>
  =goal>
    argu1      =letter
    state      start
)

```

After a visual object has been placed in the visual buffer, it can be harvested by a production like this one, which stores the object that was read into the argu1 slot of the goal and sets the state slot to start to fire the first production to find the location of the next object.

```

(P encode-second-txt
  =goal>
    ISA        arithmetic-task
    state      attend
    argu1      =t1
    operator    nil
  =visual>
    ISA        text
    value      =letter
  ==>
  =goal>
    operator    =letter
    state      start
)

```

```

(P encode-third-txt
  =goal>
    ISA        arithmetic-task
    state      attend
    argu1      =t1
    operator    "+"
    argu2      nil
  =visual>
    ISA        text
    value      =letter
  ==>
  =goal>
    argu2      =letter
    state      respond
  +retrieval>

```

```

      ISA          addition-fact
      addend1      =t1
      addend2      =letter
    )

```

After reading all characters in the screen, the task is to do the addition of 3 + 4, which can be done by retrieving the arithmetic fact chunk from the declarative memory.

```

(P respond-answer
=goal>
  ISA          arithmetic-task
  state        respond
  arg1         =num1
  operator     "+"
  arg2         =num2
=retrieval>
  ISA          addition-fact
  addend1      =num1
  addend2      =num2
  sum          =sum
?manual>
  state        free
==>
+manual>
  ISA          press-key
  key          =sum
-goal>
)

```

After the arithmetic fact chunk $3 + 4 = 7$ has been placed into the buffer of declarative memory, and the motor module is available (the state of motor buffer is free), a request is made to press the key corresponding to the value of the sum slot of the chunk in the retrieval buffer, and the goal is deleted to finish the task. The type of action requested of the hands is specified in the isa slot of the manual buffer request. The press-key request assumes that the hands are located over the home row and the fingers will be returned there after the key has been pressed.

```
(setf *actr-enabled-p* t)
```

The global variable `*actr-enabled-p*` is set to `t` to run the ACT-R model. If it is set to `nil`, the human subject, instead of the model, will perform this task.

```
(goal-focus first-goal)
```

The chunk first-goal is declared to be the current goal (placed into the goal buffer).

Typing (do-experiment) to run this model. The output, called the trace of the model, is as follows. Each line of the trace represents one event in the model and shows the time in seconds, the module that generated the event and the details of that event.

> (do-experiment)

```

0.000 GOAL          SET-BUFFER-CHUNK GOAL FIRST-GOAL REQUESTED NIL
0.000 VISION        SET-BUFFER-CHUNK VISUAL-LOCATION LOC0 REQUESTED NIL
0.000 PROCEDURAL    CONFLICT-RESOLUTION
0.050 PROCEDURAL    PRODUCTION-FIRED FIND-UNATTENDED-TXT
0.050 PROCEDURAL    CLEAR-BUFFER VISUAL-LOCATION
0.050 VISION        Find-location
0.050 VISION        SET-BUFFER-CHUNK VISUAL-LOCATION LOC1
0.050 PROCEDURAL    CONFLICT-RESOLUTION
0.100 PROCEDURAL    PRODUCTION-FIRED ATTEND-TXT
0.100 PROCEDURAL    CLEAR-BUFFER VISUAL-LOCATION
0.100 PROCEDURAL    CLEAR-BUFFER VISUAL
0.100 PROCEDURAL    CONFLICT-RESOLUTION
0.185 VISION        Encoding-complete LOC1-0 NIL
0.185 VISION        SET-BUFFER-CHUNK VISUAL TEXT3
0.185 PROCEDURAL    CONFLICT-RESOLUTION
0.235 PROCEDURAL    PRODUCTION-FIRED ENCODE-FIRST-TXT
0.235 PROCEDURAL    CLEAR-BUFFER VISUAL
0.235 PROCEDURAL    CONFLICT-RESOLUTION
0.285 PROCEDURAL    PRODUCTION-FIRED FIND-UNATTENDED-TXT
0.285 PROCEDURAL    CLEAR-BUFFER VISUAL-LOCATION
0.285 VISION        Find-location
0.285 VISION        SET-BUFFER-CHUNK VISUAL-LOCATION LOC3
0.285 PROCEDURAL    CONFLICT-RESOLUTION
0.335 PROCEDURAL    PRODUCTION-FIRED ATTEND-TXT
0.335 PROCEDURAL    CLEAR-BUFFER VISUAL-LOCATION
0.335 PROCEDURAL    CLEAR-BUFFER VISUAL
0.335 PROCEDURAL    CONFLICT-RESOLUTION
0.420 VISION        Encoding-complete LOC3-0 NIL
0.420 VISION        SET-BUFFER-CHUNK VISUAL TEXT4
0.420 PROCEDURAL    CONFLICT-RESOLUTION
0.470 PROCEDURAL    PRODUCTION-FIRED ENCODE-SECOND-TXT
0.470 PROCEDURAL    CLEAR-BUFFER VISUAL
0.470 PROCEDURAL    CONFLICT-RESOLUTION
0.520 PROCEDURAL    PRODUCTION-FIRED FIND-UNATTENDED-TXT
0.520 PROCEDURAL    CLEAR-BUFFER VISUAL-LOCATION
0.520 VISION        Find-location
0.520 VISION        SET-BUFFER-CHUNK VISUAL-LOCATION LOC5
0.520 PROCEDURAL    CONFLICT-RESOLUTION
0.570 PROCEDURAL    PRODUCTION-FIRED ATTEND-TXT
0.570 PROCEDURAL    CLEAR-BUFFER VISUAL-LOCATION
0.570 PROCEDURAL    CLEAR-BUFFER VISUAL
0.570 PROCEDURAL    CONFLICT-RESOLUTION
0.655 VISION        Encoding-complete LOC5-0 NIL
0.655 VISION        SET-BUFFER-CHUNK VISUAL TEXT5
0.655 PROCEDURAL    CONFLICT-RESOLUTION
0.705 PROCEDURAL    PRODUCTION-FIRED ENCODE-THIRD-TXT
0.705 PROCEDURAL    CLEAR-BUFFER VISUAL
0.705 PROCEDURAL    CLEAR-BUFFER RETRIEVAL
0.705 DECLARATIVE   START-RETRIEVAL
0.705 PROCEDURAL    CONFLICT-RESOLUTION
0.755 DECLARATIVE   RETRIEVED-CHUNK FACT34
0.755 DECLARATIVE   SET-BUFFER-CHUNK RETRIEVAL FACT34
0.755 PROCEDURAL    CONFLICT-RESOLUTION
0.805 PROCEDURAL    PRODUCTION-FIRED RESPOND-ANSWER
0.805 PROCEDURAL    CLEAR-BUFFER GOAL
0.805 PROCEDURAL    CLEAR-BUFFER RETRIEVAL
0.805 PROCEDURAL    CLEAR-BUFFER MANUAL

```

0.805 MOTOR	PRESS-KEY 7
0.805 PROCEDURAL	CONFLICT-RESOLUTION
1.055 PROCEDURAL	CONFLICT-RESOLUTION
1.105 PROCEDURAL	CONFLICT-RESOLUTION
1.205 MOTOR	OUTPUT-KEY # (7 2)
1.205 PROCEDURAL	CONFLICT-RESOLUTION
1.355 PROCEDURAL	CONFLICT-RESOLUTION
1.355 —	Stopped because no events left to process

The trace of the model could be used to predict the human subjects' behavior (such as the reaction time, RT) and the BOLD effect of the related brain areas when a subject is performing this task. We will discuss this issue in the coming sections.

2 ACT-R Meets fMRI

2.1 Regions of Interest

Our original efforts in ACT-R meeting fMRI (Anderson et al, 2003) used an exploratory study to help with the identification of the parietal, prefrontal, and motor regions. Based on this study we identified three regions, each 5 voxels wide, 5 voxels long, and 4 voxels high, about $16 \times 16 \times 13 \text{ mm}^3$ (The voxel in our experiments is 3.125 mm long and wide and 3.2 mm high). These regions were subsequently used in a number of studies, and several new regions have also been identified in these studies. Figure 1 shows the location of the 8 regions of interest (ROI) we have identified.

The detailed information of these regions, including the corresponding buffers in ACT-R (inside the parentheses), the Talairach coordinates of the center, the Brodmann Area and the size (if not $5 \times 5 \times 4$), is as follows:

1. **Motor1 (Manual):** Centered at $x = +/ - 37$, $y = -25$, $z = 47$. This includes parts of Brodmann Areas 2 and 4 at the central sulcus. The activation in the left hemisphere of this region will be much stronger than that in the right if moving the right hand.

2. **Anterior Cingulate (Goal):** Centered at $x = +/ - 5$, $y = 10$, $z = 38$. This includes parts of Brodmann Areas 24 and 32. This is a 5-voxel-long, 3-voxel-wide, and 4-voxel-high region.

3. **Motor2 (Vocal):** Centered at $x = +/ - 44$, $y = -12$, $z = 29$. This includes parts of Brodmann Areas 2 and 4 at the central sulcus.

4. **Parietal (Problem State or Imaginal):** Centered at $x = +/ - 23$, $y = -64$, $z = 34$. This includes parts of Brodmann Areas 7, 39, and 40 at the border of the intraparietal sulcus. The activation in the left hemisphere of this region was stronger than that in the right in most of our experiments.

5. **Prefrontal(Retrieval):** Centered at Talairach coordinates $x = +/ - 40$, $y = 21$, $z = 21$. This includes parts of Brodmann Areas 45 and 46 around the inferior frontal sulcus. The activation in the left hemisphere of this region was stronger than that in the right in most of our experiments.

6. **Caudate (Procedural):** Centered at $x = +/ - 15$, $y = 9$, $z = 2$. This is a subcortical structure, with 4-voxel-long, 4-voxel-wide, and 4-voxel-high.

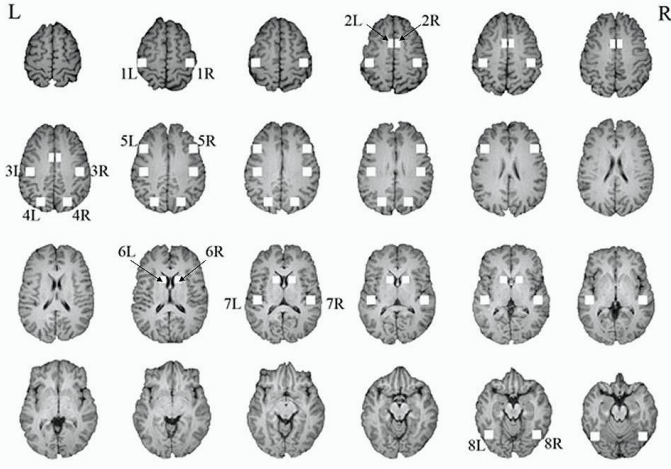


Fig. 1. Brain regions corresponding to the buffers of 8 ACT-R modules. Each slice is with 3.2 mm thick, The AC-PC line is on slice 19. The ROI number is shown in the first slice each ROI, L means the left hemisphere, R means the right hemisphere.

7. Auditory Cortex (Aural): Centered at $x = +/ - 47$, $y = -22$, $z = 4$. This includes parts of Brodmann Areas 21,22, and 42 in the region known as the auditory cortex. Note, however, this region excludes Brodmann Area 41, which is the primary auditory cortex.

8. Fusiform Gyrus (Visual): Centered at $x = +/ - 42$, $y = -60$, $z = -8$. This includes parts of Brodmann Area 37.

This completes the mapping of existing ACT-R modules onto brain regions. Of course, there are many brain regions not included above and cognitive functions not yet represented in ACT-R.

Declarative memory retrieval and problem state representation are involved in most of the cognitive tasks, including working with the Web. As an example of the consistency of the activation patterns across the cognitive tasks of above identified regions, figure 2-4 show the activation patterns in the left prefrontal (ROI 5) and left posterior parietal areas (ROI 4) across the event related fMRI experiments of algebra equations solving (Anderson et al., 2003), artificial equations solving (Qin et al, 2003), and a simple task related to memory retrieval and mental manipulation with different input and output modality (Anderson et al., 2007). We can see that the left parietal region was consistently sensitive to the task load of mental representation and mental manipulation in these tasks, and the left prefrontal region was consistently sensitive to the task load of memory retrieval in these tasks. These figures also show that the predictions of ACT-R theory are very highly correlated with the scanned data. The methods to make the predictions will be discussed in next session.

Before going ahead, we need to state a few qualifying remarks about these brain-module associations. It might seem that we are claiming that the cortical regions

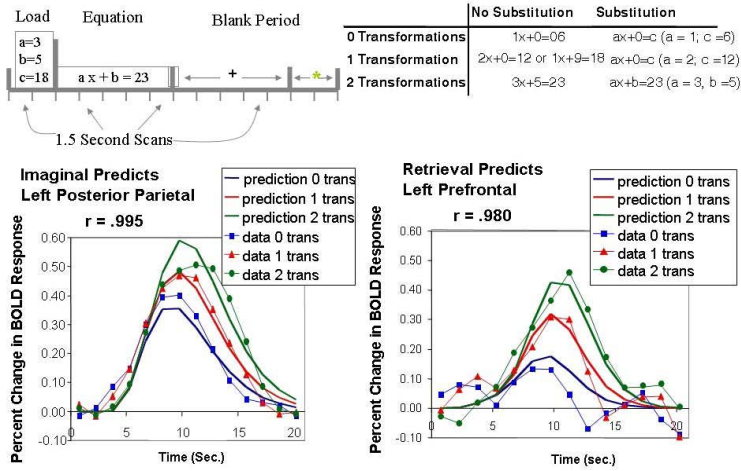


Fig. 2. The design and result of the Event-Related fMRI experiment on algebra equation solving. To solve an equation, one needs to retrieve related algebra rules and arithmetic facts, and to change the representation of the equation. One of the two factors in this design is the complexity of the equation. There are three levels of this factor. 0-transformation means no transformation needed to get the solution. 1-transformation means only one transformation (e.g., in $2x+0=12$, only division is needed) needed to get the solution, and 2-transformation means 2 transformation operations needed to get the solution. The result shows the effect of this factor in both left posterior parietal cortex (ROI 4) and left prefrontal cortex (ROI 5) and the prediction of the ACT-R model based on the activation time course of imaginal buffer and retrieval buffer respectively.

listed perform the function associated with each module. While this is a plausible inference, it is not necessary to the logic of this approach. It is only necessary that the activity of the brain region reliably reflect a particular information-processing function. Even if the function is performed in that region, there is no reason to suppose that its activity will only reflect that function. Finally, there is no claim that the ascribed function is restricted to a specific region.

2.2 Using ACT-R to Predict fMRI Data

A number of researchers (e.g., Boyton et al., 1996; Cohen, 1997; Dale and Buckner, 1997) have proposed that the Blood Oxygenation Level-Depend (BOLD) response to an event in an fMRI experiment varies according to the following function of time, t , since the event

$$B(t) = t^a e^{-t}$$

where estimates of the exponent, a , have varied between 2 and 10 (and we will constrain our estimates within these bounds). This is essentially a gamma function which will reach maximum at $t = a$ time units after the event.

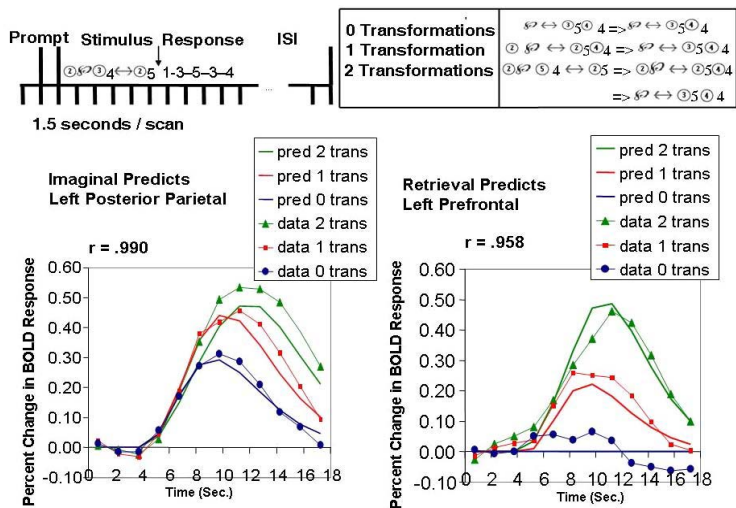


Fig. 3. The design and result of the Event-Related fMRI experiment on artificial equation solving. The result also shows the effect of the complexity of the equation in both left posterior parietal cortex (ROI 4) and left prefrontal cortex (ROI 5) and the prediction of the ACT-R model based on the activation time course of imaginal buffer and retrieval buffer respectively.

As we have seen, the output trace of an ACT-R model is the time course of the activations of the modules involved in the cognitive task and can be used to predict the behavior data such as RT. Further more, with the information of when and how long of each activation of the modules, one can predict the BOLD effect of the brain regions corresponding to the ACT-R module listed above using the gamma function.

Figure 5 illustrates the general idea about how we map from events in an information-processing model onto the predictions of the BOLD function in our early efforts. Basically, one can think of the observed BOLD function in a region (e.g., Total line in Figure 5) as reflecting the sum of separate BOLD functions for each period of time the buffer is active (e.g., the First, Second, and Third events in Figure 5). Each period of activity is going to generate a BOLD function according to a gamma function as illustrated. The peak of the BOLD functions reflects roughly when the buffer was active but is offset because of the lag in the hemodynamic response. The height of the BOLD function reflects the duration of the event since the integration makes the height of the function proportional to duration over short intervals.

In Anderson et al. (2003) it was proposed that while a buffer is active it is constantly producing a change that will result in a BOLD response according above gamma function. The observed fMRI response is integrated over the time that the buffer is active. Therefore, the observed BOLD response will vary with time as

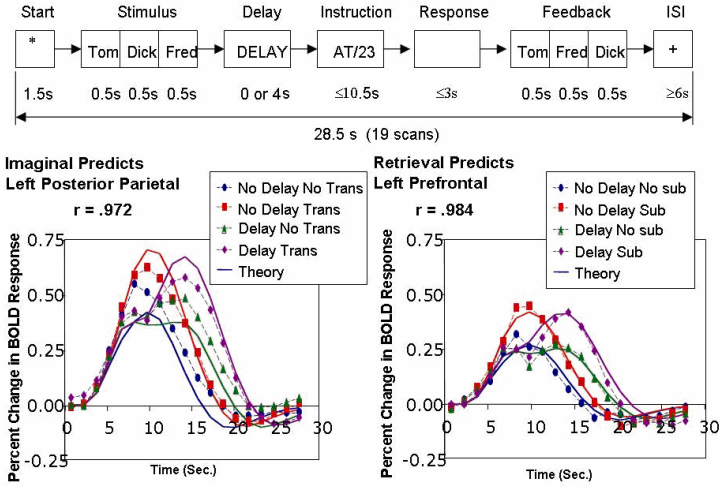


Fig. 4. The design and result of the third Event-Related fMRI experiment. The task is to change the order of the three names according to the instruction. The instruction can be a two digit number. For example, if it is 23, it means shifting the order of the second name with the third name. This is the case of with transformation. There are 1/3 trials no transformation required (e.g., the number is 24, there is no fourth name in stimulus). If the instruction is a word, such as AT, this is a trial with substitution (needs to retrieve the associated number, 23 in this example. The participants were trained well for this kind of associations before scan). There are 1/2 trials with no substitution needed. The result shows the effect of the transformation in left posterior parietal cortex (ROI 4) and the effect of the substitution in the left prefrontal cortex (ROI 5) and the prediction of the ACT-R model based on the activation time course of imaginal buffer and retrieval buffer respectively.

$$CB(t) = M \int_0^t i(x) B\left(\frac{t-x}{s}\right) dx$$

where M is the magnitude scale for response, s is the latency scale, and $i(x)$ is 1 if the buffer is occupied at time x and 0 otherwise. Note because of the scaling factor, the prediction is that the BOLD function will reach maximum at roughly $t = a * s$ seconds.

Figure 6 shows an example of the activation time course of the retrieval and imaginal modules formed from the trace of the ACT-R model for the algebra equation solving (Anderson et al., 2003). Figure 2 shows the prediction of the BOLD effect based on this kind of time course and $CB(t)$. For the left prefrontal region (corresponding to the retrieval buffer), the parameters of scale (s) is estimated 0.691, exponent (a), 8.180, and magnitude $MF(a+1)$, 0.933 (This is a more meaningful measure since the height of the function is determined by the exponent as well as M .), the result correlation coefficient between the scan data and the prediction among the condition of the complexity of the equations is 0.980; For the left posterior parietal region (corresponding to the imaginal

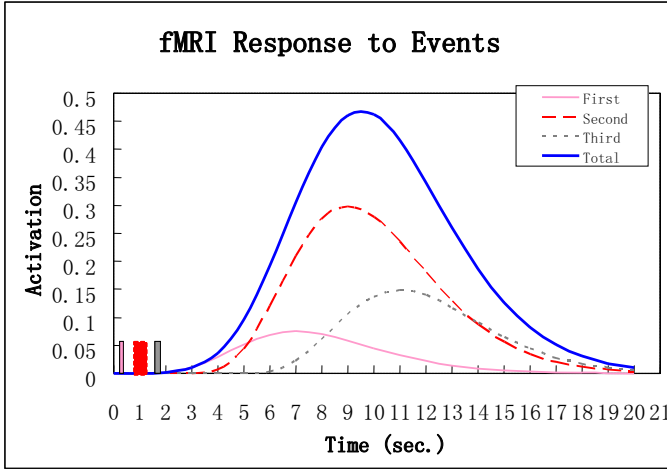


Fig. 5. The observed BOLD effect (total) in a ROI is the sum of the BOLD effects of three events in this ROI with different starting time and length of duration

buffer), s is 1.647, a is 3.054, and $M\Gamma(a + 1)$ is 3.486, the result correlation coefficient between the scan data and the prediction among the condition of the complexity of the equations is 0.995. Three parameters for the left motor region has also estimated and the result correlation coefficient between the scan data and the prediction among the condition of the complexity of the equations is 0.955, but the detail will not be given here.

These parameters were estimated by trying to minimize the following quantity:

$$\sum_{i \in ROIs} \sum_{j \in Conditions} \sum_{k \in Scans} (\bar{B}_{ijk} - \hat{B}_{ijk})^2 / S_i^2$$

where $\bar{B}_{ijk}$ is the mean BOLD response, $\hat{B}_{ijk}$ is the predicted response, and S_i^2 is the mean error in the BOLD response for ROI i calculated by the interaction between the 84 values (6 conditions x 14 scans) by 8 participants interaction term. Under the hypothesis that all deviations are normally distributed noise, this quantity is distributed as a chi-square with degrees of freedom equal to the number of observations (252) minus parameters (9)-that is 243 degrees of freedom. The value of this quantity is 340.83, which is quite significant indicating, not surprisingly, that there are things in the data not predicted by the model. On the other hand, it is not radically different from its expected value (the degrees of freedom), indicating that we are capturing most of the systematic variance (Anderson et al, 2003).

The methods of making the prediction of BOLD effect have been improved in the later studies (e.g., Anderson et al, 2007), but we will not go ahead to cover the detailed issues here.

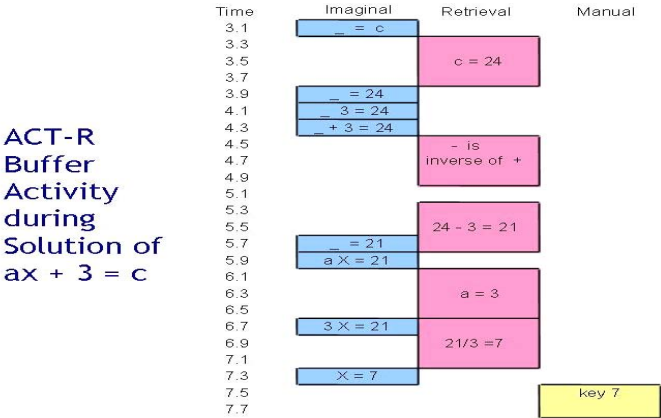


Fig. 6. The time course of the activations in the imaginal and retrieval buffers in solving the equation of $ax+3=c$, where $a=3$, $c=24$

3 Discussion

The central points discussed in this article are:

- 1. ACT-R is a theory and computational model on the components of information processing in human brain and how the interaction among these components generates human cognitive behavior. This theory and computational model may have potential application in Web intelligence;
- 2. The empirical findings on the brain regions corresponding to these components;
- 3. The approach of how the detailed processing of an information-processing theory like ACT-R can make precise predictions about the BOLD response.

This approach can be seen as a two way bridge. From one direction, it allows us to interpret the significance of the fMRI data in terms of precise information-processing operations rather than in terms of relatively diffuse concepts. On the other hand, we can use fMRI experiments to test the predictions of the theory and to help us to find ways to improve the theory. Therefore, this approach may shed light on the research of Web Intelligence meets Brain Informatics.

An important issue to make a precise prediction of fMRI data is to make a precise prediction of behavior data, which may need to estimate some parameters in an ACT-R model. For the detail, one can check the tutorial in <http://act-r.psy.cmu.edu/actr6>.

We have only introduced the symbolic level of ACT-R in this article. There is a subsymbolic level of ACT-R. For example, the activation of a chunk i is a sum of a base-level activation B_i , reflecting the recency and frequency of practice of the chunk, and an associative activation, reflecting its relevance to the current context:

$$A_i = B_i + \sum_k \sum_j W_{kj} S_{ji} + \varepsilon \quad (1)$$

where, W_{kj} is the amount of activation from source j in buffer k , S_{ji} is the strength of association from source j to chunk i . ε is the noise value, and the base-level activation

$$B_i = \ln\left(\sum_{j=1}^n t_j^{-d}\right) \quad (2)$$

where n is the number of presentations for chunk i . t_j is the time since the j th presentation. d is the decay parameter which is almost always set to 0.5.

The probability of recall the chunk i depends on the expected activation, A_i , the retrieval activation threshold τ , and the amount of noise in the system which is controlled by the parameter s :

$$\text{recall probability}_i = \frac{1}{1 + e^{\frac{\tau - A_i}{s}}} \quad (3)$$

Also the time to retrieve chunk i is depended on A_i :

$$\text{Time} = Fe^{-A_i} \quad (4)$$

The detailed information on the subsymbolic level of ACT-R can be found in Anderson et al (2004) and the ACT-R 6 tutorial in the ACT-R website shown above.

References

1. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of Mind. *Psychological Review* 111, 1036–1060 (2004)
2. Anderson, J.R., Qin, Y., Sohn, M.-H., Stenger, V.A., Carter, C.S.: An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin and Review* 10, 241–261 (2003)
3. Anderson, J.R., Qin, Y., Yung, K.-J., Carter, C.S.: Information-Processing Modules and their Relative Modality Specificity. *Cognitive Psychology* 54, 185–217 (2007)
4. Boyton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J.: Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience* 16, 4207–4221 (1996)
5. Cohen, M.S.: Parametric analysis of fMRI data using linear systems methods. *NeuroImage* 6, 93–103 (1997)
6. Dale, A.M., Buckner, R.L.: Selective Averaging of Rapidly Presented Individual Trials Using fMRI. *Human Brain Mapping* 5, 329–340 (1997)
7. Qin, Y., Sohn, M.-H., Anderson, J.R., Stenger, V.A., Fissell, K., Goode, A., Carter, C.S.: Predicting the practice effects on the blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task. *Proceedings of the National Academy of Sciences of the United States of America* 100, 4951–4956 (2003)

Appendix

Here is the whole model of solving the problem of 3+4 with the Lisp functions simulating the visual stimulus in the experiment window. This experiment can be performed by a human subject or by an ACT-R 6 model.

```
(clear-all)
  (defvar *response* nil)
  (defmethod rpm-window-key-event-handler ((win rpm-window) key)
    (setf *response* (string key)))
    ;; define the experiment window

(defun do-experiment ()

(reset)

(let* ((window (open-exp-window "addition")))

  (add-text-to-exp-window :text "3" :x 75 :y 175)
  (add-text-to-exp-window :text "+" :x 125 :y 175)
  (add-text-to-exp-window :text "4" :x 175 :y 175)
    ;;To simulate the stimulus of the experiment
  (setf *response* nil)

  (if *actr-enabled-p*
      (progn
        (install-device window)
        (proc-display)
        (run 10 :real-time t))
        ;; if *actr-enabled-p* = t, ACT-R model performs the task
    (while (null *response*)
      (allow-event-manager window))))
    ;; otherwise, a human subject will perform the task

(define-model addition34

  (sgp :v t :show-focus t :needs-mouse nil)
    ;; set :v to t is to show the trace, set :show-focus to t is to show the red
    ;; focus circle on the experiment window
  (sgp :esc t :lf .05 :trace-detail medium)
    ;; to set retrieval time as 0.05 seconds

  ;; Define Chunk-types
  < The ACT-R model shown in the text >
)
```

The Neural Mechanism of Human Numerical Inductive Reasoning Process: A Combined ERP and fMRI Study

Peipeng Liang¹, Ning Zhong^{1,2}, Shengfu Lu¹, Jiming Liu^{1,3}, Yiyu Yao^{1,4},
Kuncheng Li⁵, and Yanhui Yang⁵

¹ The International WIC Institute, Beijing University of Technology, China
ppliange@emails.bjut.edu.cn, lusf@mail.bjut.edu.cn

² Dept. of Life Science and Informatics, Maebashi Institute of Technology, Japan
zhong@maebashi-it.ac.jp

³ Dept. of Computer Science, Hong Kong Baptist University, Hong Kong
jiming@comp.hkbu.edu.hk

⁴ Dept. of Computer Science, University of Regina, Canada
yyao@cs.uregina.ca

⁵ Dept. of Radiology, Xuanwu Hospital, Capital Medical University, China
likuncheng1955@yahoo.com.cn, yanhui826@gmail.com

Abstract. Inductive reasoning is one of the most important higher level cognitive functions of the human brain, and we still know very little about its neural mechanism. In the present study, event-related potential (ERP) and event-related fMRI are used to explore the dynamic spatiotemporal characteristics of inductive reasoning process. We hypothesize that the process of numerical inductive reasoning is partially dissociable over time and space. A typical task of inductive reasoning, function-finding, was adopted. Induction tasks and calculation tasks were performed in the experiments, respectively. ERP results suggest that the time course of inductive reasoning process is partially dissociable as the following three sub-processes: number recognition (the posterior P100 and N200), strategy formation (P300) and hypothesis generation and verification (the positive slow waves). fMRI results show many activations, including pre-frontal gyrus (BA 6), inferior parietal lobule (BA 7, 40), and occipital cortex (BA 18). After the respective discussions, the two kinds of data are combined qualitatively, then the dynamic spatiotemporal characteristic of inductive reasoning process are depicted using a conceptual figure. This study is a preliminary effort towards deeply understanding the dynamic information processing mechanism of human inductive reasoning process.

1 Introduction

1.1 Background

Web intelligence (WI) is a new direction for scientific research and development that explores the fundamental roles as well as practical impacts of Artificial

Intelligence (AI) and advanced Information Technology (IT) on the next generation of Web-empowered systems, services, and environments [26,39]. Furthermore, a new perspective of WI research is proposed from the viewpoint of Brain Informatics (BI) [40,41]. BI is a new interdisciplinary field to study human information processing mechanism systematically from both macro and micro points of view by cooperatively using experimental brain/cognitive technology and WI centric advanced information technology. In particular, it attempts to understand human intelligence in depth, towards a holistic view at a long-term, global field of vision, to understand the principles, models and mechanisms of human multi-perception, reasoning and inference, problem solving, learning, discovery and creativity [41].

Web inference engine (WIE) is one of the core research contents of WI [26,41]. As a must, WIE will be designed to combine multiple reasoning methods in PSML (Problem Solver Markup Language) representation and distributed inference engines efficiently and effectively. The design and development of inference/reasoning engine have been the bottleneck of intelligent systems in AI and WI. On the one hand, in the domain of traditional AI research, in which the single inference mode (i.e., deductive reasoning) is applied on the single knowledge base, inference engine is facing many difficulties, although many useful results have been produced. Furthermore, when used to develop the Wisdom Web [26], the existing inference techniques have shown their insufficiency and cannot meet the real world needs in the age of the World Wide Web. In particular, the traditional inference techniques, including deductive reasoning, probabilistic reasoning, fuzzy reasoning, and so on, mainly lack the ability of discovering new knowledge, i.e., inductive reasoning ability. On the other hand, WIE requests real-time application. The introduction of multiple inference modes to WIE has brought the difficulty of the inference speed when the ability of inference has been strengthened. Inductive reasoning is a much slower process than deductive reasoning.

However, human brain can perform complex reasoning tasks smoothly. From the urgent needs of WIE, we are interested in the question: how do human brain implement inductive reasoning process. There must be a great benefit for the development of intelligent systems if the information processing mechanism of human inductive reasoning can be clarified and understood.

1.2 Related Work

Inductive reasoning is defined as the process of inferring a general rule (conclusion) by observation and analysis of specific instances (premises) [19]. Inductive reasoning is used when generating hypotheses, formulating theories and discovering relationships, and is essential for scientific discovery.

There have been some preliminary studies of human inductive reasoning, including brain lesion and brain imaging experiments, and some models have been proposed [32,33,36]. On the one hand, results from patients mainly show the evidence of left hemisphere lateralization and the critical role of left prefrontal cortex in inductive reasoning [9,10,16,23,35]. On the other hand, by using PET

and fMRI, researchers have performed some experiments to study the neural mechanism of inductive reasoning, and found some brain regions related to inductive reasoning [4,5,12,13,14,15,29]. However, the previous studies have not adequately addressed the following three issues. Firstly, the previous experiments need button-pressing response, which may confuse the understanding of some brain areas, such as BA 6 and cerebellum. Secondly, the effects of background knowledge cannot be eliminated for both sentential tasks and RPM tasks, i.e., there exist significant individual differences among subjects. Finally, these studies employ PET or fMRI technique only. Due to their low temporal resolution, we still know little about the time course of inductive reasoning process in the human brain.

In order to overcome the above three insufficiencies, we will use, in the present study, one kind of (numerical) function-finding tasks to explore the time course and spatial localization of human inductive reasoning process. The function-finding task is a typical inductive reasoning task [19]. The main advantage of the (numerical) function-finding task is that all subjects have almost the same background knowledge. Event-related potential and event-related fMRI experimental techniques will be used, respectively, to study inductive reasoning process.

1.3 Experimental Hypothesis

As a higher level cognitive function, inductive reasoning is much more a serial processing course. According to Sternberg's cognitive components theory, inductive reasoning task can be broken down into its underlying cognitive sub-processes (i.e., components). Haverty and colleagues [19] divided a function-finding task into three stages: data gathering, pattern finding and hypothesis generation. Lin and colleagues [25] presented an in-depth discussion of inductive reasoning in function-finding tasks. From this point of view, this study can be regarded as exploring the neural basis of the cognitive sub-processes in inductive reasoning process.

Our hypotheses are that: 1) the process of numerical inductive reasoning is composed of number recognition(encoding), strategy formation, hypothesis generation and validation; 2) the above three components are (partially) dissociable over time; 3) the above three components can be (partially) separated in spatial localization. The objective of the present study is to give evidence of our hypothesis using ERP and fMRI, and then construct the dynamic spatiotemporal characteristics of human inductive reasoning process by preliminarily combining the two kinds of brain data.

2 ERP Experiment

2.1 Subjects

Eleven paid undergraduate or graduate students (8 males and 3 females) from the Beijing University of Technology participated in the experiment. The subjects

ranged in age from 20 to 26 years (mean age 23.0 years). All subjects were right-handed and had the normal or corrected-to-normal vision. None of the subjects reported any history of neurological or psychiatric diseases. The Institutional Review Board of the Beijing University of Technology approved all experimental procedures, and written informed consent was obtained from each participant.

2.2 Stimuli and Tasks

The experimental tasks are adapted from one kind of intelligence test problems. The basal element of the task is a reverse triangle as shown in Figure 1. The three numbers located in three different positions may constitute a calculation rule, i.e., an equation like $Z = X + Y$.

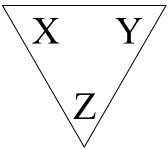


Fig. 1. A triangle with numbers

Figure 2 gives the inductive reasoning task, which is assembled with three reverse triangles as mentioned above.

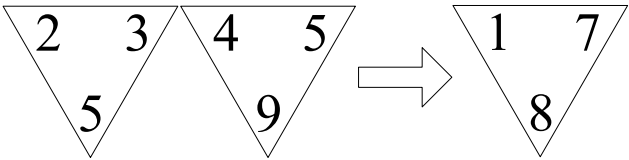


Fig. 2. An example of the inductive reasoning tasks

The range of the number involved is set to be 0-9. Addition and subtraction are used to form the calculation rule for induction tasks. Two kinds of tasks are designed in the present study as shown in Figure 3. Calculation tasks are designed as the baseline of induction tasks. Each of induction tasks has the corresponding calculation task in which $+/-$ operators located in the bottom corner, and the numbers in the top left corner and the top right corner are identical to the corresponding induction tasks. The $+/-$ symbols used in the three reverse triangles are balanced for the calculation tasks.

According to the location of the maximal number, there are three styles for induction tasks including $Z = X + Y$, $Y = X + Z$, and $X = Y + Z$. Ten tasks are designed for each style of induction tasks, and then 30 corresponding calculation tasks are made. In order to reduce the influences of button-pressing to ERP waveforms for the induction tasks, the oddball paradigm is introduced: the third

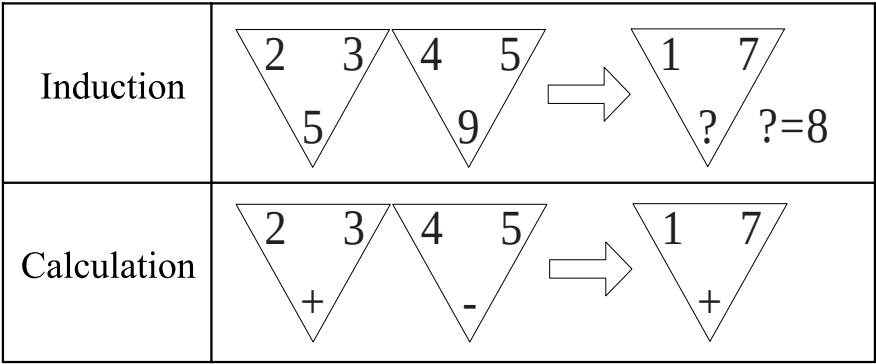


Fig. 3. Two kinds of experimental tasks

reverse triangle (with the value of question mark) fitted the rule generalized from the first two for 80% of tasks (called congruent), while not fitted for the remaining 20% (called incongruent). However, the oddball paradigm is used in the opposite direction, i.e., we focus on the congruent induction tasks while the incongruent tasks which need button-pressing response are used to reduce the learning effect in the experimental course.

The reverse triangles are equilateral triangles with 44 mm edges. All the numbers in reverse triangles are written with 36 pt Times New Roman font. All tasks (the first two reverse triangles) constitute a picture of the horizontal visual angle of 5.87° and the vertical visual angle of 2.61°. The tasks were presented to subjects in the style of black-background and white-foreground.

2.3 Procedure

The subjects seat in a comfortable chair in an electrically shielded room. A computer display for visual stimulus delivery is placed 1000 mm away from the subjects’ eyes. Stimuli presentation is controlled by the Presentation software as shown in Figure 4. All time parameters are determined by a pilot study.

Firstly, a fixation mark “+” is shown in the center of screen followed by a blank. After a while, the first reverse triangle is presented for 1 s, and then followed by the presentation of the second reverse triangle for 1 s,¹ then the third reverse triangle is displayed. Subjects are asked to generalize a calculation rule from the first two reverse triangles and judge the value of the question mark after the presentation of the third reverse triangle. If the value is wrong

¹ In our previous ERP study [24], two reverse triangles are simultaneous presented to subjects. However, the strategies used by subjects are complex from the open questionnaires after experiments. Hence, the first two reverse triangles are shown one by one with 1 s interval in the present study in order to regulate and simplify the strategy. Here, 1 s is from pilot study, which allow subjects see straight the corresponding reverse triangle while have no time to think more about the task.

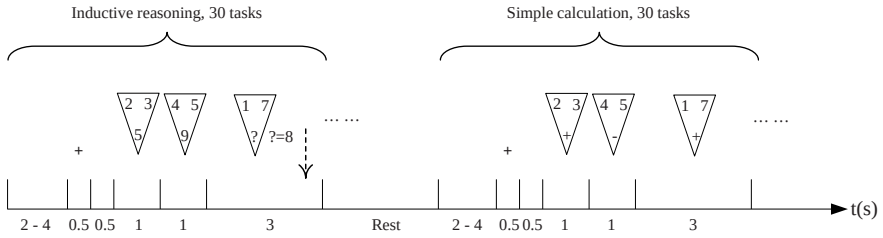


Fig. 4. Paradigm of stimuli presentation for ERP experiment

(incongruent) subjects require to press the return button with the right index finger as quickly as possible, else if the value is right (congruent), they do not need to press the button. For simple calculation tasks, subjects are instructed just to do the given arithmetics without button-pressing.

In order to keep the consistent mental state, induction tasks and calculation tasks are presented in two blocks respectively, while the block order is balanced among subjects. Tasks are randomized in every block. The two tasks both need about 5 minutes, and there is an interval for rest after the first block. After the ERP experiment, there is a questionnaire in which subjects are required to write down the process how they solve the problems.

2.4 ERP Recording

Electroencephalographic activity is recorded using a 128 channel BrainAmp amplifier (Brain Products, Munich, Germany) with a 128 electrode cap as Figure 5. The electrodes are placed according to the extended 10-20 System and referenced during recording to an additional reference electrode placed between Fz and Cz, with impedance of less than 8 $k\Omega$. Both the vertical and horizontal electrooculogram (EOG) are recorded. The sampling rate is 500 Hz/channel.

2.5 Data Analysis

Analyzer software (from Brainproducts Inc. in Germany) is used to deal with the EEG data. EEG data are recalculated using linked mastoids as reference $[(TP9 + TP10)/2]$, and then band pass filtered at 0.01 – 100 Hz (slope 12 dB/octave), notch filtered at 50 Hz. The analyzed time epoch is 3600 ms (400 ms before and 3200 ms after the presentation of the first reverse triangle). All calculation tasks are included in analysis, while among the aforementioned 80% congruent induction tasks, only tasks, in which subjects do not press the return button after the third reverse triangle is shown, are included in analysis. To avoid eye movement and other artifacts, all epochs exceeding $\pm 100 \mu V$ in any channel are excluded from further analysis. A baseline correction is performed, and then epoches are averaged.

This study is different from a previous study [24]. With the change of stimuli presentation, the inductive reasoning process has been changed. The time course

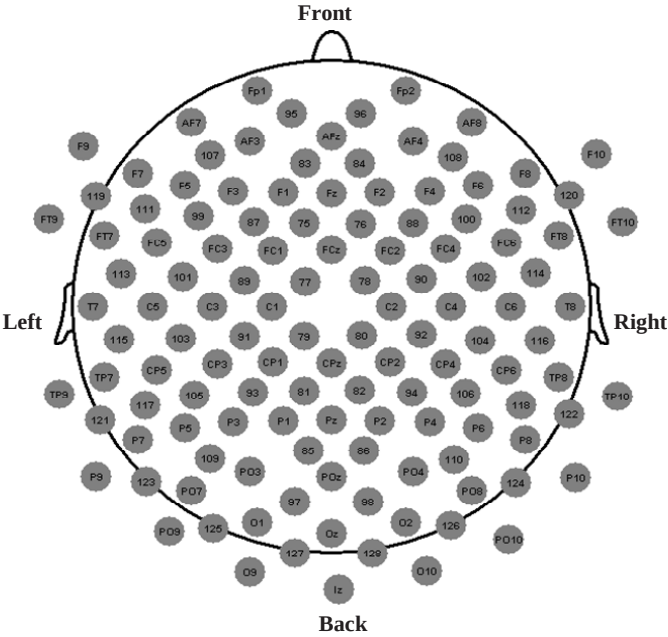


Fig. 5. Distribution of 128 electrodes on the cap

is firstly partitioned into three segments: 0-1000 ms (presentation for the first reverse triangle), 1000-2000 ms (presentation for the second reverse triangle) and 2000-3200 ms (presentation for the third reverse triangle and left for problem solving). According to our design, we concentrate on the third segment, in which subjects need perform inductive reasoning for the induction tasks while just simple calculation is needed for the calculation tasks. The first and second segments both have 1000 ms, which is just enough for encoding the stimuli. According to the grand ERP waveforms and the current source density (CSD) maps, the time course of inductive reasoning process is divided into the following three time windows: 2000-2250 ms, 2250-2600 ms, 2600-3200 ms.

The following 15 sites are chosen for statistical analysis: F3/4, Fz, FCz, in the frontal scalp, C3/C4 in the central scalp, and P3/4, Pz, CPz, O1/2, Oz, 117/118 in the posterior scalp. The P100 and N200 in the posterior sites, the frontal P300 are measured in 120-180 ms, 180-220 ms, and 250-350 ms respectively for all the three segments. The amplitudes of the posterior P100 and N200, and the frontal P300 are chosen to repeat the measure analysis of variance (ANOVAs) for three segments respectively. The mean amplitude of the posterior P300 (or LPC, late positive component) is computed in 250-500 ms for all three segments. The mean amplitude of the positive slow wave (PSW) is computed for the third segment in the following seven time windows: 500 - 600 ms, 600 - 700 ms, 700 - 800 ms, 800 - 900 ms, 900 - 1000 ms, 1000 - 1100 ms and 1100 - 1200 ms, then the mean amplitudes are used to ANOVAs.

For the posterior P100 and N200, the frontal P300 and the posterior P300, the within-subject effects are considered in the following two factors: task (2 levels, induction and calculation) and location (i.e., the electrode site; 6 levels for the frontal and 9 levels for posterior scalps respectively). While three factors are considered: task, location and time (7 levels as aforementioned seven time windows) for the positive slow wave. The p values are corrected using the Greenhouse-Geisser method.

2.6 Results

ERP Results. In order to reduce the effect of button-pressing, oddball experimental paradigm is adopted for ERP experiment. Behavioral data is not collected for ERP session. Figure 6 shows the whole view of grand ERP waveforms. Figure 7 displays the ERP components elicited. Figure 8 denotes the current source density diagrams of inductive reasoning process after the presentation of the third segment.

For the posterior P100 amplitude, the main effect of task is not significant for all the three segments. Only the interaction effect of $task * location$ is significant [$F(8, 80) = 3.378, p < 0.05$]; The other effects are all not significant. For the posterior N200 amplitude, all effects of task are not significant except the main effect of location for the first segment [$F(8, 80) = 5.419, p < 0.05$] and the second segment [$F(8, 80) = 3.429, p < 0.05$].

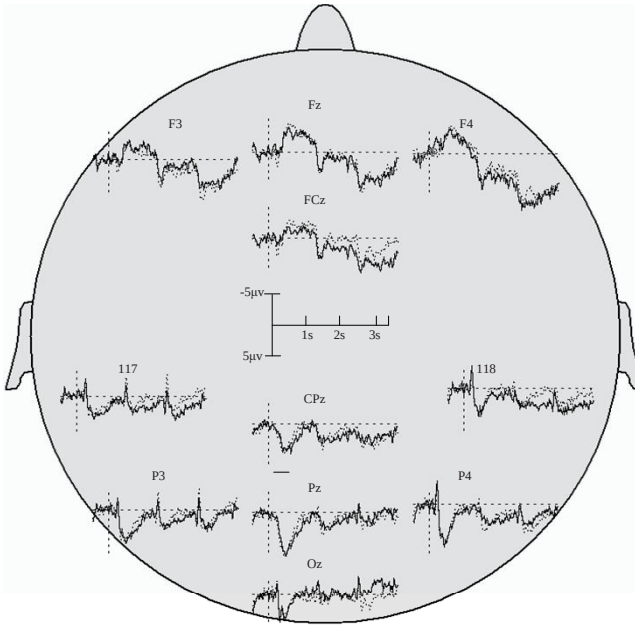


Fig. 6. Overlapped are the grand average ERPs (11 participants) to induction (solid line) and calculation (dotted line) recorded from 11 scalp sites

For the frontal P300, except the main effect of location for the third segment [$F(5, 50) = 4.097, p < 0.05$], all the other effects are not significant. All the effects for the posterior P300 are not significant.

The following statistical analysis is only for the third segment. For the PSW (from 500 ms to 1200 ms), *task * location * time* statistical analysis revealed that: the main effect of time is significant [$F(6, 60) = 5.517, p < 0.01$]; the interaction effect of *location * time* is significant [$F(84, 840) = 3.9, p < 0.05$]; the interaction effect of *task * location * time* is significant [$F(84, 840) = 2.853, p < 0.05$]; all the other effects are not significant. Furthermore, repeated measured ANOVA is applied to every time window for the frontal positive slow wave and the posterior positive slow wave respectively. Except the main effect of location for the frontal 500-600ms [$F(5, 50) = 9.299, p < 0.01$], the frontal 600-700ms [$F(5, 50) = 7.772, p < 0.01$], the frontal 700-800ms [$F(5, 50) = 4.554, p < 0.05$], the posterior 500-600 ms [$F(8, 80) = 3.297, p < 0.05$] and the posterior 600-700 ms [$F(8, 80) = 5.923, p < 0.01$], all the other effects are not significant.

Questionnaires. The open questionnaires are collected, classified, and deciphered qualitatively. From the questionnaires, the process of numerical inductive reasoning for most of subjects can be summarized as follows. In general, most of subjects reported that, after the display of the third reverse triangle, they will generate the hypothesis with respect to a calculation rule from the first reverse triangle, and verify the hypothesis with the second one. In the end, they can judge the third reverse triangle using the obtained rule.

2.7 Discussions

According to our experimental design, combined with subjective reports from the questionnaires, we can describe the problem solving course of induction tasks as follows: subjects encode the first reverse triangle, encode second one, and encode the third one; generate some strategies (based on background knowledge and habits); form the hypothesis (form the calculation rule) and verify it. For the calculation tasks, people just need to compute the corresponding arithmetic formulas.

For all the three segments, the posterior P100 may reflect number recognition. This is consistent with [38], in which N1-P2 complex elicited by number additions is thought of indicating visual encoding. The posterior N200 may relate with encoding as well, for encoding will last around 200ms [6]. Furthermore, the posterior N200 may also indicate attention [1,2,20,27]. The corresponding statistical results support the above statements, for the main effect of task is not significant for the three segments.

P300 has been elicited from both the frontal and posterior scalp for three segments. According to the most popular theory for the P300 components [7], the amplitude varies as a function of memory load and the latency reflects the time needed for updating of working memory. In the light of experimental design and the above problem solving course, the P300 may relate to the understanding of tasks in working memory, in which some strategies/tendency may be generated

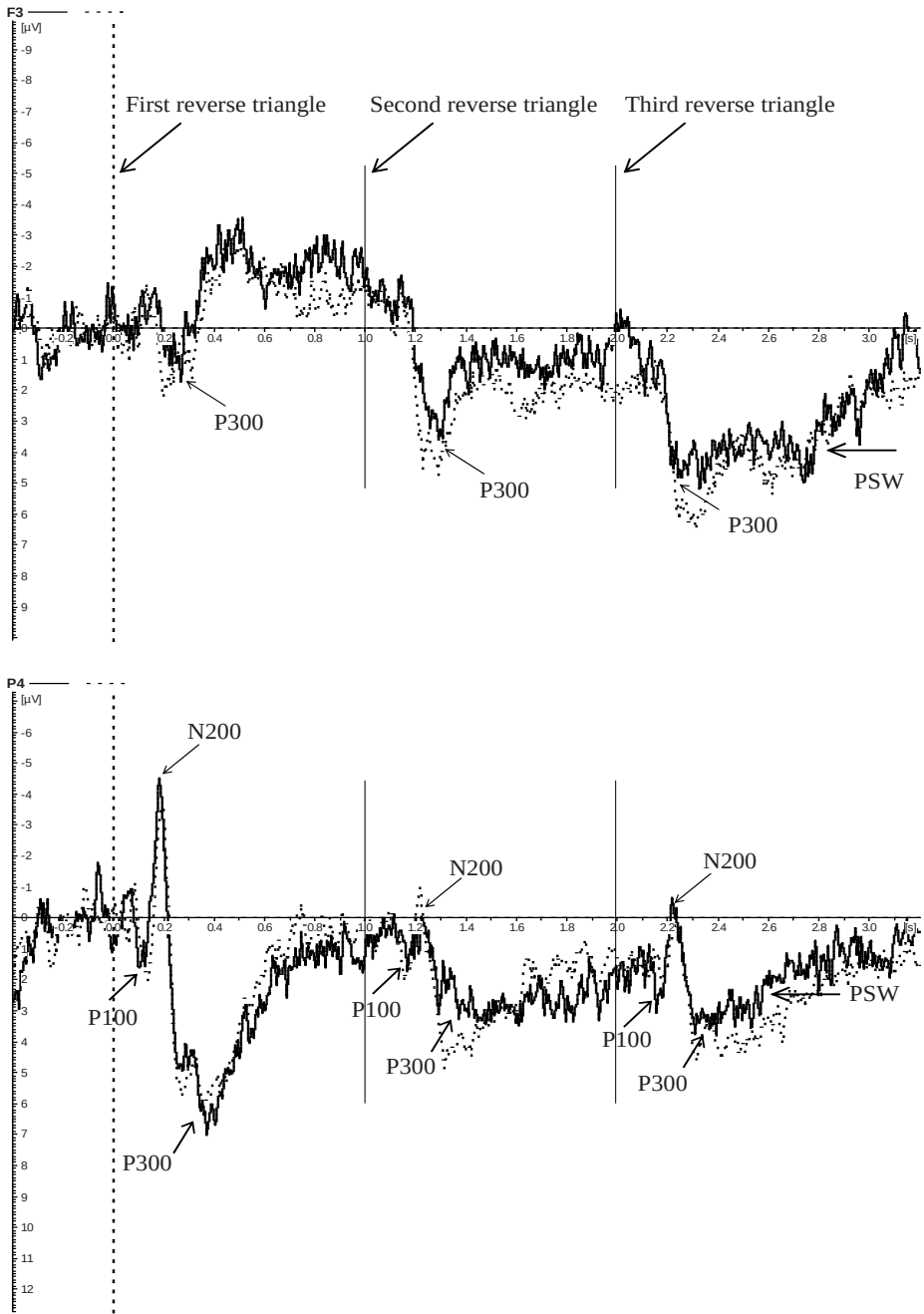


Fig. 7. Grand average ERPs at F3 and P4 electrodes. ERPs are time-locked to the onset of the sequential stimuli. The vertical bars indicate the onset of the first, second and third reverse triangles. EPRs are compared in two experimental conditions: induction (solid line) and calculation (dotted line).

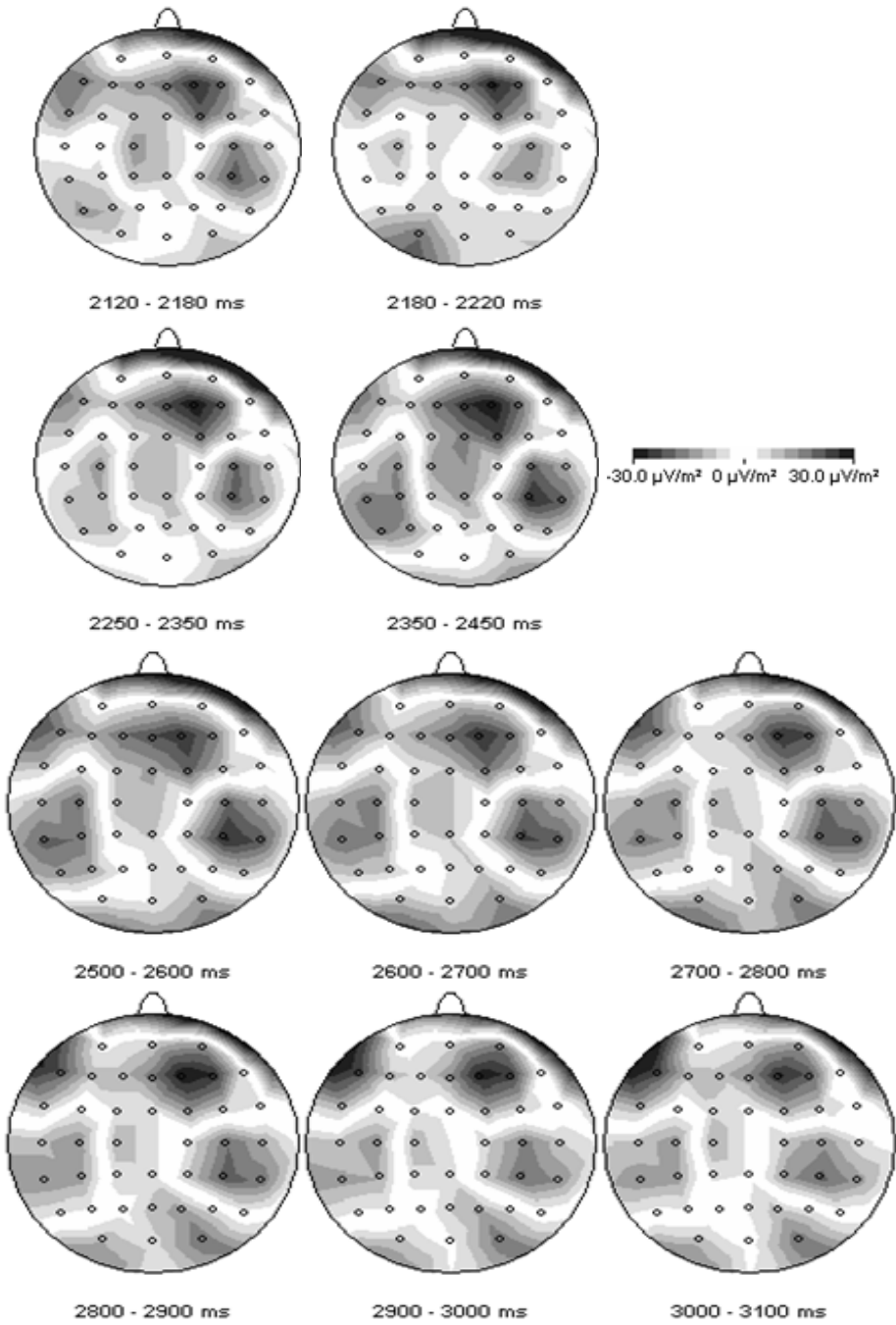


Fig. 8. Current source density diagrams of inductive reasoning process after the presentation of the third reverse triangle

automatically. However, statistical results do not reveal the evident split between induction and calculation for the frontal P300 and the posterior P300, which may be the reason that the work load of induction and calculation in this experiment is similar. Different from the previous study [24], no significant LNC for the third segments is elicited from the frontal scalp. Although statistical results did not give us more support, the ERP waveforms suggest that the serial presentation of the first two reverse triangles have mediated the course of information processing in numerical inductive reasoning tasks.

Slow waves [21,24,31] in the time window of 500-1200 ms may reflect hypothesis formation and verification. There should have no cycle in the course of solving the two kinds of problems, for only simple arithmetic operations are involved. Then it can be inferred that hypothesis formation and hypothesis verification may be sequential organized. However, where is the interface of hypothesis formation and hypothesis verification, that is, the time point when the hypothesis formation has finished and the hypothesis verification will begin, is unclear.

In summary, the inductive reasoning process is located in the time window of 2000-3200 ms (after the presentation of the third reverse triangle) for the sequential stimuli presentation. The time course of inductive reasoning process may be (partially) dissociable according to the grand ERPs, CSD and questionnaires. The sub-processes of inductive reasoning process are number recognition/encoding, strategy formation, hypothesis generation and verification.

3 fMRI Experiment

3.1 Subjects

20 paid undergraduate or graduate male students from the Beijing University of Technology participated in the experiment. The subjects ranged in age from 20 to 26 years (mean age 24.4 years). All subjects are right-handed and have the normal or corrected-to-normal vision. None of the subjects reported any history of neurological or psychiatric diseases. The Institutional Review Board of the Beijing University of Technology approved all experimental procedures, and written informed consent was obtained from each participant.

3.2 Stimuli and Tasks

The stimuli and tasks used are the same to ERP experiment. However, the stimuli presentation paradigm in the fMRI session is different from the ERP session as shown in Figure 9, i.e., the first two reverse triangles are presented for 2 s respectively.

The fMRI experiment is the further amendment of experimental design based on the previous ERP experiment. The aim of the fMRI experiment is to describe a three stage inductive reasoning process by adjusting the present time of the first two reverse triangles from 1 s to 2 s. For studying inductive reasoning process by combining ERP and fMRI, only activated areas for the whole inductive reasoning process (vs simple arithmetic) are reported in this paper. The three stage model of inductive reasoning will be described in another paper.

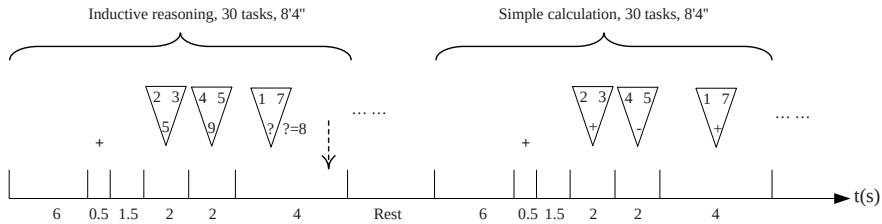


Fig. 9. Paradigm of stimuli presentation for fMRI experiment

3.3 Image Recording

Scanning is performed on a 3 T Siemens system using a standard whole-head coil. Functional data are acquired using a gradient echo planar pulse sequence (TR=2 s, TE=30 ms, 25 axial slices, $3.75 \times 3.75 \times 4 \text{ mm}^3$ voxels, 0.8 mm inter-slice gap, 242 volumes per run, 90° flip angle, 64×64 matrix size in $240 \times 240 \text{ mm}^2$ field of view). The imaging sequence is optimized for detection of the BOLD effect including local shimming and 10 s of scanning prior to data collection to allow the MR signal to reach equilibrium.

Visual stimuli are projected onto a screen that is viewed through a mirror. Subjects need to perform 60 experimental trials over the course of two event-related fMRI scans (eight minutes and four seconds for each session). Across two runs, a total of 484 whole-brain volumes are acquired.

3.4 fMRI Data Analysis

Data are analyzed using Statistical Parametric Mapping (SPM2). Images are corrected for differences in timing of slice acquisition, followed by rigid body motion correction. The data are realigned and normalized to the standard SPM2 EPI template. The registration of the EPI data to the template is checked for each individual subject. Head movement is $< 2 \text{ mm}$ in all cases. The fMRI data are then smoothed with a 8 mm FWHM isotropic Gaussian kernel. The hemodynamic response to the three reverse triangles is modeled with event-related delta functions, which are convolved with the canonical hemodynamic response function and its time derivative employed in SPM2. No scaling is implemented for global effects. The resulting time series across each voxel are high-pass filtered with a cut-off of 128 s to remove section-specific low frequency drifts in the BOLD signal. An auto-regression AR(1) is used to exclude the variance explained by the previous scan. The contrast images for each subject are then used in a random effects analysis to determine which regions are the most consistently activated across subjects using a one-sample t test. The activations reported survived an uncorrected voxel-level intensity threshold of $p < 0.001$ with a minimum cluster size of 20 contiguous voxels.

3.5 Results

The activated areas are reported in Table 1. The results are organized according to frontal lobe, parietal lobe, and occipital lobe. Our results (as shown in Figure 10) indicate that inductive reasoning (induction vs calculation) engages a neural network comprising bilateral prefrontal cortex (BA 6), left precuneus (BA 7) and inferior parietal lobule (BA 40), right superior and inferior parietal lobule (BA 7, 40), and bilateral occipital lobes (BA 18).

Table 1. Activations of Induction vs Calculation

Location	Nox	Hemi.	BA	X	Y	Z	T
Frontal lobe							
Superior frontal gyrus	263	L	6	-4	6	64	5.48
Medial frontal gyrus		R	6	4	10	48	5.01
Middle frontal gyrus	101	L	6	-26	2	62	5.23
				-36	4	54	3.84
Medial frontal gyrus	93	R	6	30	8	64	4.75
				44	10	52	4.53
Parietal lobe							
Precuneus	1534	L	7	-26	-72	32	6.50
				-16	-70	48	5.66
Inferior parietal lobule	76	L	40	-42	-42	40	4.26
Inferior parietal lobule	928	R	40	34	-52	34	6.16
Superior parietal lobule			7	26	-70	42	4.97
Occipital lobe							
Fusiform gyrus	1204	L	18	-20	-90	-24	7.97
Inferior occipital gyrus				-34	-90	-14	7.37
Middle occipital gyrus	1325	R	18	22	-86	-14	7.60
Inferior occipital gyrus				34	-84	-18	7.04

3.6 Discussions

The results are consistent with previous neuroimaging and patient studies of inductive reasoning [5,10,12,13,14,15,23] on the whole, including the activation of left prefrontal and occipital cortex, and left hemisphere lateralization. The significant difference from previous studies is that we do not find the activation of temporal cortex. Combined with experimental tasks and previous studies, the activated brain areas are partitioned into the following four parts for discussion: left PFC, right PFC, inferior parietal lobule, and occipital cortex.

In order to depict the dynamic spatiotemporal characteristics of inductive reasoning process, Every activated brain area survived for induction vs calculation should be discussed of its own role in inductive reasoning. Different areas may play different roles (direct or indirect, sufficient or necessary) in different time windows (with a certain time resolution of ms level) for inductive reasoning process.

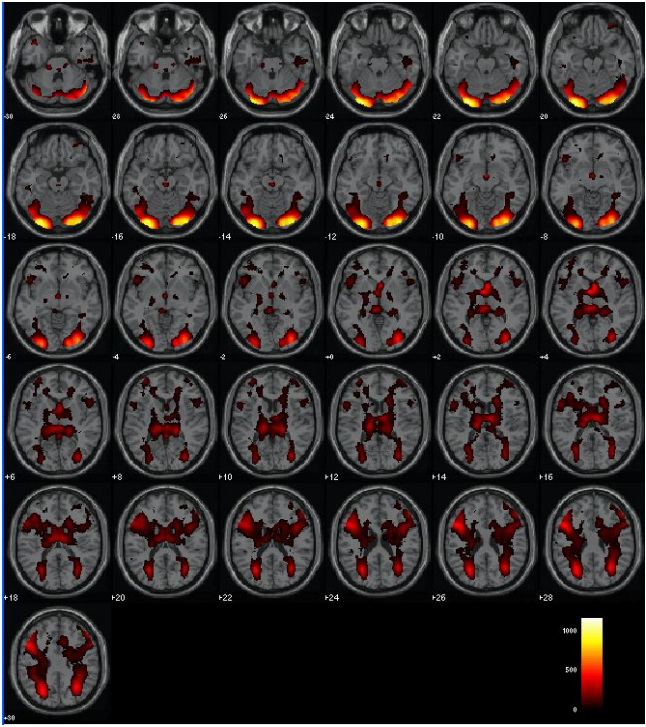


Fig. 10. Slice view of activations for Induction vs Calculation

Left Prefrontal Cortex. Left prefrontal cortex is considered to have relation with many higher level cognition, including plan, executive function, monitoring etc. many researchers have suggested the key role of left PFC in inductive reasoning [12,14]. However, our results mainly show the activation of superior frontal (BA 6), the middle/medial frontal gyrus (BA 6), and have some difference from some related studies. Some researchers reported the activation of the left ventrolateral prefrontal cortex, for instance, Geake [11] conjectured the activation of left ventrolateral (BA 45, 46) may reflect the need of relation integration. Goel [15] also found the activation of the lateral PFC (BA 46) in hypothesis formation. Some studies found the activation of the left dorsolateral prefrontal cortex, for example, Strange [34] reckoned that DLPFC maybe mediate rule application after rule acquisition. Goel [14] postulated that DLPFC relate with access of world knowledge in the process of hypothesis formation and validation.

Compared with these previous studies, our results show activation of superior and medial/middle parts of frontal gyrus, which are also coincident with some related works. Prabhakaran [30] thought of bilateral prefrontal cortex may relate with monitoring of each processing steps and answer generation when performing RPM tasks. Geake [11] attributed the left superior frontal gyrus to recollection of rule knowledge in analogical reasoning (inductive reasoning is involved).

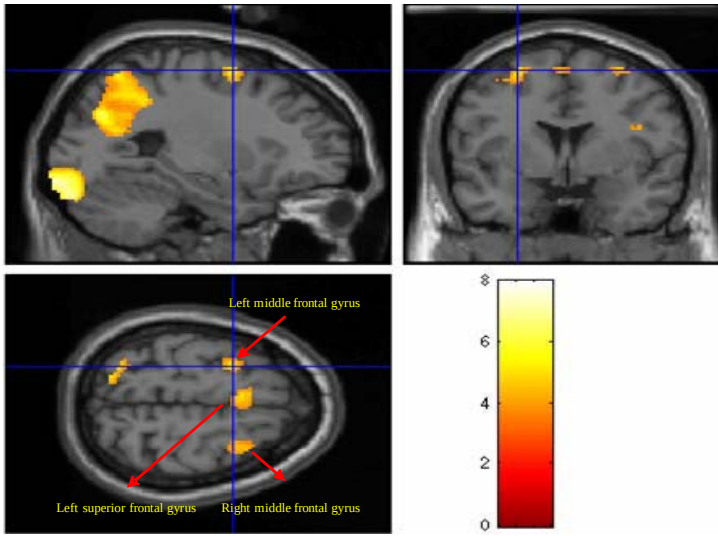


Fig. 11. Activations of bilateral prefrontal cortex (BA 6)

Combined the existing similar studies and our experimental design, we hypothesize that the left prefrontal gyrus in the present experiment may associate with access of corresponding background knowledge from the long-term memory.

Right Prefrontal Cortex. Many studies have found the activation of the right prefrontal cortex in inductive reasoning. For instance, Goel [15] found activation of the right ventral lateral PFC (BA 47) in hypothesis generation, while right dorsolateral PFC (BA 46) in hypothesis maintenance. Christoff [4] also observed the activation of right PFC in relation integration. Another study from Goel [13] even showed significant right hemisphere lateralization. Goel have reviewed some literatures and then concluded that RH play a special role to inductive reasoning, although this role is still ambiguous.

Inferior Parietal Lobule. We have found activation of the left and right inferior parietal lobule (IPL) in this study. The left and right IPL, although they both contribute to mental arithmetic, may not be functionally equivalent [3]. A task calling only for the internal manipulation of numerical quantity, is found to rely more on the right parietal lobule, whereas a task presumably requiring access to verbal memory is more strongly associated with the left parietal lobule.

Occipital Cortex. The activations in the occipital mainly include the left fusiform and inferior occipital gyrus (BA 18) and right occipital gyrus (BA 18). In general, the occipital cortex relates with visual information processing. In the present study, induction tasks and calculation tasks are designed to keep the same visual load, so the occipital cortex activations may reflect visual working memory and visual attention [8]. Some other studies showed the similar results.

4 Combined Discussion

From the ERP waveforms and fMRI activated areas, it may be inferred that inductive reasoning process is (partially) dissociable over time and space. We may combine the two kinds of brain data qualitatively to preliminarily depict the dynamic spatiotemporal characteristics of human inductive reasoning process (as shown in Figure 12) based on the above discussions.

Brain is complex, and need recruit many brain areas even for a simple sub-process. In addition to the present experiments, the following conceptual model also draws results from other related studies. This model will tend to grasp the primary brain areas in inductive reasoning process, although it may not be able to use all the involved areas.

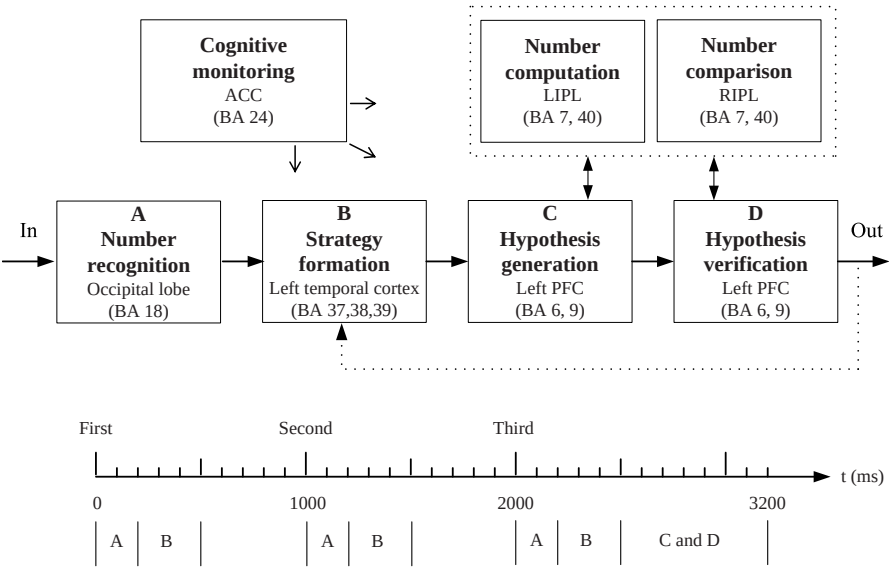


Fig. 12. Schematic representation of dynamic spatiotemporal characteristics for numerical inductive reasoning process

Number Recognition. The posterior P100 and N200 may reflect stimuli encoding, i.e., number recognition, corresponding with activation of bilateral occipital cortex (BA 18). The posterior N200 may also relate with attention.

Strategy Formation. P300 elicited in the present study and the frontal LNC (N400) elicited in our previous study are conjectured to associate with understanding tasks and forming strategies to solve the problem. Many studies have proposed that P300 in different tasks have different sources, in which temporal lobe is one of sources [18]. Some researches have suggested that the left temporal lobe are mainly the N400 neural generators [17,37].

Hypothesis Generation and Verification. Hypothesis generation and verification cannot be split over time in the current study, although some fMRI experiment have proposed their different neural mechanisms [15]. The positive slow wave may reflect these two sub-processes. There has been an agreement on the role of the left prefrontal gyrus to the hypothesis generation and verification.

Cognitive Monitoring. The most popular theory about the ACC (Anterior Cingulate Cortex) is its role in cognitive monitoring [22,28]. ACC is also included in the model, although activation is not found in ACC for the comparison of induction vs calculation.

Simple Arithmetic and Number Comparison. From the model, the main line of human inductive reasoning process is composed of the following sub-processes: input, number recognition, strategy formation, hypothesis generation and verification, output. However, for numerical inductive reasoning tasks, number calculation (addition and subtraction) and number comparison are concomitant after the encoding stage (number recognition).

5 Conclusions

This paper preliminarily depicts the dynamic information processing course of human inductive reasoning process using the three reverse triangles tasks by combining the ERP and fMRI experimental results. The results have validated our hypothesis to a certain extent, i.e., the inductive reasoning process is (partially) dissociable over time and space.

This work is just a beginning. On the one hand, the paradigms of stimuli presentation of the ERP and fMRI experiment are not identical. On the other hand, many other methods will be used to analyze the ERP and fMRI data, such as ICA, connectivity mapping, fMRI activation induced dipole localization analysis etc. These methods will offer a multi-aspect view of brain data, and be helpful to further construct the information processing course of inductive reasoning process. Additionally, learning effect during experiment should be considered, for it may confuse experimental data especially for human higher level cognition. The future work will include the following aspects: (1) Simultaneous ERP and fMRI recording experiment will be performed; (2) Multiple methods will be applied to ERP and fMRI data respectively; (3) ERP and fMRI data will be combined quantitatively.

Acknowledgements

The authors would like to thank Xiuqin Jia and other fellows in the International WIC Institute for their support in experiments. This work is partially supported by the grant-in-aid for scientific research (No. 18300053) from the Japanese Ministry of Education, Culture, Sports, Science and Technology, the Open Foundation of Key Laboratory of Multimedia and Intelligent Software

Technology (Beijing University of Technology) Beijing, and the Doctoral Research Fund of Beijing University of Technology (00243).

References

1. Bledowski, C., Prvulovic, D., Goebel, R., Zanella, F.E., Linden, D.E.: Attentional System in Target and Distractor Processing: a Combined ERP and fMRI Study. *NeuroImage* 22(2), 530–540 (2004)
2. Chen, A.T., Li, H., Qiu, J., Luo, Y.J.: Time Course of Visual Categorization: Electrophysiology Evidence of ERPs. *Chinese Science Bulletin* 51(11), 1281–1286 (2006)
3. Chochon, F., Cohen, L., van de Moortele, P.F., Dehaene, S.: Differential Contributions of the Left and Right Inferior Parietal Lobules to Number Processing. *Journal of Cognitive Neuroscience* 11(6), 617–630 (1999)
4. Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J.K., Holvoak, K.J., Gabrieli, J.D.: Rostrolateral Prefrontal Cortex Involvement in Relational Integration during Reasoning. *NeuroImage* 14(5), 1136–1149 (2001)
5. Christoff, K., Ream, J.M., Geddes, L.P.T., Gabrieli, J.D.E.: Evaluating Self-Generated Information: Anterior Prefrontal Contributions to Human Cognition. *Behavioral Neuroscience* 117(6), 1161–1168 (2003)
6. Coles, M.G.H., Rugg, M.D.: Event-related Brain Potentials: an Introduction. In: Rugg, M.D., Coles, M.G.H. (eds.) *Electrophysiology of Minds: ERPs and Cognition*, pp. 1–26. Oxford University Press (1995)
7. Donchin, E.: Surprise!... Surprise? *Psychophysiology* 18(5), 493–513 (1981)
8. Fink, G.R., Halligan, P.W., Marshall, J.C., Frith, C.D., Frackowiak, R.S.J., Dolan, R.J.: Where in the Brain Does Visual Attention Select the Forest and the Trees? *Nature* 382, 626–628 (1996)
9. Gazzaniga, M.S., Smylie, C.S.: Dissociation of Language and Cognition. *Brain* 107(1), 145–153 (1984)
10. Gazzaniga, M.S.: *The Mind's Past*. University of California Press, Berkeley, CA (1998)
11. Geake, J.G., Hansen, P.H.: Neural Correlates of Intelligence as Revealed by fMRI of Fluid Analogies. *Neuroimage* 26, 555–564 (2005)
12. Goel, V., Gold, B., Kapur, S., Houle, S.: The Seats of Reason? An Imaging Study of Deductive and Inductive Reasoning. *NeuroReport* 8(5), 1305–1310 (1997)
13. Goel, V., Dolan, R.J.: Anatomical Segregation of Component Processes in an Inductive Inference Task. *Journal of Cognitive Neuroscience* 12(1), 1–10 (2000)
14. Goel, V., Vartanian, O.: Dissociating the Roles of Right Ventral Lateral and Dorsal Lateral Prefrontal Cortex in Generation and Maintenance of Hypotheses in Set-shift Problems. *Cerebral Cortex* (2004)
15. Goel, V., Dolan, R.J.: Different Involvement of Left Prefrontal Cortex in Inductive and Deductive Reasoning. *Cognition* 93(3), B109–B121 (2004)
16. Goel, V., Shuren, J., Sheesley, L., Grafman, J.: Asymmetrical Involvement of Frontal Lobes in Social Reasoning. *Brain* 127(3), 1–8 (2004)
17. Guillem, F., Kaoub, N., Rougier, A., Claverie, B.: Intracranial Topography of Event-related Potentials (N400 vs P600) Eliciting during a Continuous Recognition Memory Task. *Psychophysiology* 32, 382–392 (1995)
18. Halgren, E., Marinkovic, K., Chauvel, P.: Generators of the Late Cognitive Potentials in Auditory and Visual Oddball Tasks. *Electroencephalography Clinical Neurophysiology* 106, 156–164 (1998)

19. Haverty, L.A., Koedinger, K.R., Klahr, D., Alibali, M.W.: Solving Induction Problems in Mathematics: Not-so-trivial Pursuit. *Cognitive Science* 24(2), 249–298 (2000)
20. Hillyard, S.A., Munte, T.F.: Selective Attention to Color and Location: an Analysis with Event-related Brain Potentials. *Percept. Psychophys.* 36, 185–198 (1984)
21. Iguchi, Y., Hashimoto, I.: Sequential Information Processing during a Mental Arithmetic Is Reflected in the Time Course of Event-related Brain Potentials. *Clinical Neurophysiology* 111(2), 204–213 (2000)
22. Kondo, H., Osaka, N., Osaka, M.: Cooperation of the Anterior Cingulate Cortex and Dorsolateral Prefrontal Cortex for Attention Shifting. *NeuroImage* 23, 670–679 (2004)
23. Langdon, D., Warrington, E.K.: The Role of the Left Hemisphere in Verbal and Spatial Reasoning Tasks. *Cortex* 36(5), 691–702 (2000)
24. Liang, P.P., Zhong, N., Wu, J.L., Lu, S.F., Liu, J.M., Yao, Y.Y.: Time Dissociative Characteristics of Numerical Inductive Reasoning: Behavioral and ERP Evidence. In: *Proceeding of the International Joint Conference of Neural Network (IJCNN 2007)* (Accepted)
25. Lin, T.K.: Modelling Inductive Reasoning Ability for Adaptive Virtual Learning Environment. In: *Proceeding of the IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2004)*, pp. 343–349. IADIS Press, Lisbon, Portugal (2004)
26. Liu, J.: Web Intelligence (WI): What Makes Wisdom Web? In: *Proceeding of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pp. 1596–1601 (2003)
27. Luo, Y.J., Nan, Y., Li, H.: Difference of Neural Correlates between Subitizing and Counting Reflected by ERPs. *Acta Psychologica Sinica* 36(4), 434–441 (2004)
28. MacDonald, A.W., Cohen, J.D., Stenger, V.A., Carter, C.S.: Dissociating the Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control. *Science* 288(9), 1835–1838 (2000)
29. Parsons, L.M., Osherson, D.: New Evidence for Distinct Right and Left Brain Systems for Deductive vs Probabilistic Reasoning. *Cerebral Cortex* 11(10), 954–965 (2001)
30. Prabhakaran, V., Smith, J.A.L., Desmond, J.E., Glover, G.H., Gabrieli, J.D.E.: Neural Substrates of Fluid Reasoning: an fMRI Study of Neocortical Activation during Performance of the Raven's Progressive Matrices Test. *Cognitive Psychology* 33, 43–63 (1997)
31. Ruchkin, D.S., Johnson, J.R., Mahaffey, D., Sutton, S.: Toward a Functional Categorization of Slow Waves. *Psychophysiology* 25, 339–353 (1988)
32. Sternberg, R.J.: *Beyond IQ: A Triarchic Theory of Human Intelligence*. East China Normal University Press, Shanghai, China (1999)
33. Sternberg, R.J.: Toward a Unified Theory of Human Reasoning. *Intelligence* 10, 281–314 (1986)
34. Strange, B.A., Henson, R.N.A., Friston, K.J., Dolan, R.J.: Anterior Prefrontal Cortex Mediates Rule Learning in Humans. *Cerebral Cortex*, 11, 1040–1046 (2001)
35. Varley, R., Siegal, M.: Evidence for Cognition without Grammar from Causal Reasoning and 'Theory of Mind' in an Agrammatic Aphasic Patient. *Current Biology* 10(12), 723–726 (2000)
36. Verguts, T., Maris, E., Boeck, P.D.: A Dynamic Model for Rule Induction Tasks. *Journal of Mathematical Psychology* 46, 455–485 (2002)

37. Wang, J.J., Zhang, M.D., Hokama, H.: The Neural Generators of the N400 Component of Event-related Potentials. *Shanghai Archives of Psychiatry* 18(1), 5–9 (2006)
38. El Yagoubi, R., Lemaire, P., Besson, M.: Different Brain Mechanisms Mediate Two Strategies in Arithmetic: Evidence from Event-related Brain Potentials. *Neuropsychologia* 41(7), 855–862 (2003)
39. Zhong, N., Liu, J., Yao, Y.Y. (eds.): *Web Intelligence*. Springer, Berlin (2003)
40. Zhong, N., Liu, J., Yao, Y.Y., Wu, J.: *Web Intelligence (WI) Meets Brain Informatics (BI)*. In: *Proceeding of the First International Conference on Complex Medical Engineering (CME 2005)* (2005)
41. Zhong, N.: *Toward Human-level Web Intelligence*. In: Ślęzak, D., Yao, J., Peters, J.F., Ziarko, W., Hu, X. (eds.) *RSFDGrC 2005. LNCS (LNAI)*, vol. 3642, pp. 23–28. Springer, Heidelberg (2005)

Central Nervous Processing for Acupuncture at Liv3 with fMRI: A Preliminary Experience

Wei Wang¹, Kuncheng Li¹, Baoci Shan², Jianyang Xu³, Bin Yan²,
Jing Hao¹, Yanhui Yang¹, Ke Li², and Na Lu²

¹ Department of Radiology, Xuanwu Hospital
Capital University of Medical Sciences, Beijing, China

² Institute of High Energy Physics
Chinese Academy of Sciences, Beijing, China

³ General Hospital of Chinese People's Armed Police Forces
Beijing, China

Abstract. Acupuncture is a traditional Chinese healing technique, which is gaining popularity as an alternative and complementary therapeutic intervention in many worldwide countries. The acupoints are arranged on so-called meridians, which represent a network of channels each connected to a functional organic system. Our experiment is to investigate the mechanism of acupuncture at Liv3 (Taichong) and possible post-effect of acupuncture. Functional magnetic resonance imaging (fMRI) of the whole brain was performed in 18 healthy right-handed young volunteers during two stimulation paradigms: ten subjects received real acupuncture (RA) at acupoints rights Liv3 (on the hand) and other 8 subjects received sham acupuncture (SA) near . Liv3.fMRI data were analyzed using SPM99. Acupuncture at Liv3 resulted in activation of bilateral cerebella, prefrontal lobe (PF), superior parietal lobule (SPL) and inferior parietal lobule (IPL), occipital lobe, parahippocampal gyrus, insula, thalamus, lentiform nucleus; contralateral temporal pole and anterior cingulate gyrus (ACG), posterior cingulate gyrus (PCG). The PE of RA activated bilateral cerebella, PF, SPL, and IPL, occipital lobe, lentiform nucleus; isolateral temporal pole, hippocampus, insula and thalamus; contralateral head of nucleus caudate, corpus callosum, ACG and PCG. Acupuncture at Liv3 resulted in activation of visual area, limbic system and subcortical gray structures, which was considered as the specific central nervous response within the brain to acupuncture at Liv3. Moreover, the activation still existed during PE of RA. fMRI provides an objective evidence for post-effect existence, which will establish the foundations of later scientific design in acupuncture experiments.

1 Introduction

Acupuncture, a traditional Chinese healing technique that can be traced back at least about 3,000 years, is gaining popularity as an alternative and complementary therapeutic intervention in many worldwide countries, especially in the treatment of pain, under a milestone of the National Institutes for Health (NIH) [1]. The acupoints are arranged on so-called “meridians”, which represent

a network of channels each connected to a functional organic system. However, the scientific basis of acupuncture remains unclear.

Functional magnetic resonance imaging (fMRI) or positron emission tomography (PET) provides a non-invasive, no radiation, repetitive, higher space and time resolution imaging modality to probe the nature of acupuncture [2-7]. These recent developments in fMRI can be used for the study of the correlation between various acupoints and specific functional areas of the brain with the hope of establishing a clear correlation between acupoints and organ disorders or diseases [2-4]. These studies provide primary proofs for the mechanism of acupuncture. Some scholars [2,4,5] reported that the acupoints specificity existed by using fMRI. Meanwhile, other researchers thought some structures, such as subcortical structures or hypothalamus, maybe serve as one key element in mediating efficacy of acupuncture stimulation [7-8]. These findings suggest that cerebral activation responding to acupuncture stimulation may shed light on the central mechanism of the acupuncture effect.

Traditionally, multiple blocks design was used in most of fMRI studies in order to increase the reliability of the statistics analysis results. Based on the theory of Traditional Chinese Medicine (TCM), the therapeutic effects will still last several minutes even several hours after the needle is drawn out from acupoint, which is called post-effect of acupuncture [9]. The effect requires that the experiment design of acupuncture should be different from other experiments such as visual or auditory experiments [10]. If multiple blocks design was used in the acupuncture study, the influence among multiple blocks was produced. Some researchers might have noticed this issue and adopted increasing the interval time between 2 blocks (in general 15-20 min) in their fMRI experiments [3,5,11]. However most of researchers ignored this issue.

In this preliminary study, we have used fMRI technique in normal young subjects to examine changes in brain activity associated with acupuncture at Liv3 to probe the mechanism of acupuncture.

2 Materials and Methods

2.1 Subjects

Eighteen right-handed normal, healthy young volunteer (eleven men and seven women, age 22-30 years, mean 26.5 ± 3.4 years) participated in this study. None had a history of psychiatric or neurological disorders or head trauma with loss of consciousness, or intake of tranquilizing drugs in the last 3 days. No normal subject was in pain or distress at the time of the study. The local ethics committee of Xuan Wu Hospital, Capital University of Medical Sciences approved the study. All subjects provided written informed consent.

2.2 Paradigm and Single Block Design

An acupuncturist with 20 years experience performed all the acupuncture procedures with manual acupuncture. Volunteers were told they would experience

acupuncture stimulation at one acupoint, but they did not know it was at real or sham acupoint.

We applied two models of stimulation, i.e., Liv3 (Taichong) on the dorsum of the foot, that is real acupuncture (RA) and one sham acupoint (SA) near Liv3, in the depression anterior approximately 10 mm to liv3. Everyone received only one model in the random order. Ten and eight subjects received stimulation at real acupoint and sham acupoint, respectively.

The experiment session was composed of special single block design to avoid the influence of post-effect: including rest state (R, no any stimulation, 3 min), stimulation state (S, RA or SA, 3 min) and post-effect state (PE, 14 min).

The acupuncturist entered the scanning room and sterilized local skin of acupoint routinely before fMRI scanning. After the 3 min resting period, one inch (TCM) stainless sliver needles (diameter 0.30 mm *times* 25 mm) were inserted 1 cm deep and slowly rotated (about 180) producing the characteristic “*De-Qi*” [3] (heaviness, tension and numbness) and manipulated about 3 min by 1 Hz with even method during the acupuncture period, and then pulled out needle.

2.3 Magnetic Resonance Imaging

MR imaging was performed with a standard head coil and a 1.5 T MR imager (Sonota, Siemens, Germany). The images covered the whole brain and were parallel to the AC-PC line. Functional images were obtained using a blood oxygenation level dependant (BOLD) T2*-weighted gradient echo EPI sequence with in plane resolution of 3.44 mm (TR 3000 ms, TE 50 ms, flip angle 90°, field of view 220 *times* 220 mm, matrix 64 *times* 64, 6 mm slice thickness and 1.2 mm gap). The subject was settled into the scanner and instructed to close his or her eyes and relax throughout the imaging session. Blinder and earplug were used in order to reduce the stimulation of light and sound. The head was immobilized with foam support cushions and adhesive strips to prevent motion.

2.4 Imaging Processing and Data Analysis

fMRI data were analyzed using Matlab (Math Works Inc., Natick, MA) and statistical parametric mapping (SPM99; Wellcome Department of Cognitive Neurology, London, UK) [12]. After realignment, the images were normalized to the Montreal Neurological Institute (MNI) space and then were smoothed spatially using a 9 *times* 9 *times* 18 mm³ Gaussian kernel. The smoothed data were analyzed voxel by voxel in two levels. The first level was fixed effects analysis based on a general linear model with a boxcar response function. The activated cerebral areas of acupuncture at RA/SA or PE contrast to the baseline were obtained respectively. The statistical parametric maps were formed, which showed activation above the height threshold $P < 0.001$ and were considered as specific areas by acupuncture at the acupoint. To acquire the specific active areas of acupuncture, the nonspecific factors, such as pain and psychological factors, needed to be removed from all of active areas of acupuncture, which was the aim of setting sham acupuncture test. Anatomically, Liv3 and the nearby non-acupoint

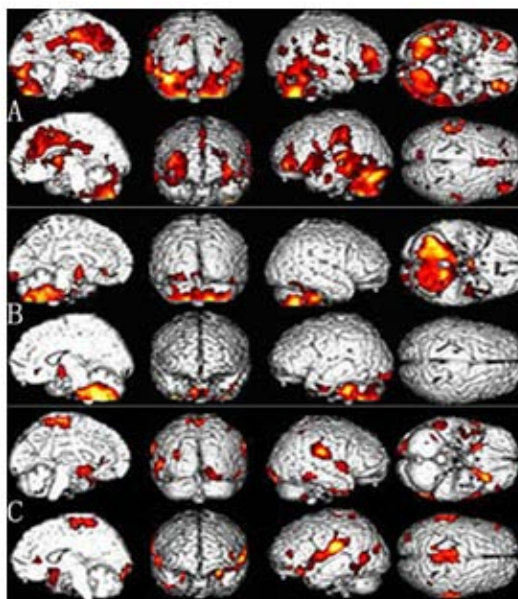


Fig. 1. (A) Activation areas of the real acupuncture at Liv 3 contrast to nearby sham acupoint, and (B) post-effect of RA contrast “PE” of SA, and (C) sham acupuncture. The red denoted active areas.

are both innervated by L5 spinal nerve, thus, we hypothesize the active areas of acupuncture Liv3 are consistent with those of its nearby sham acupoint. Hence, we can obtain the specific active brain areas of acupuncture by eliminating areas from nonspecific factors by using the mask function of SPM [13-14]. Finally, the coordinates of active areas were transformed into the Talairach space [15].

3 Results

Group results for fMRI (Fig. 1; Tables 1 and 2) during acupuncture at Liv3 (RA) and SA showed different activation (signal intensity increased during stimulation as compared with that during rest) patterns as follows:

The specific activation brain areas during RA (10 subjects), comparing with the corresponding SA (8 subjects), showed on bilateral cerebella, frontal lobe (BA 6, 8, 46), superior parietal lobule (SPL; BA 7) and inferior parietal lobule (IPL; BA 39, 40), occipital lobe (BA 18, 19), parahippocampal gyrus (BA 27/28), insula, thalamus, lentiform nucleus; isolateral postcentral gyrus (BA 1); contralateral temporal pole (BA 38); anterior cingulate gyrus (ACG; BA 24) and posterior cingulate gyrus (PCG; BA 23, 31) (Fig. 1, Table 1). The specific active areas during PE of RA (10 subjects), comparing that of “PE” of SA, indicated on bilateral cerebella, frontal lobe (BA 6, 8, 9, 10, 46, 47), SPL (BA 7) and IPL (BA 39, 40), occipital lobe (BA 17, 18, 19), lentiform nucleus; isolateral

Table 1. Brain areas activated by acupuncture at RA vs. SA and PE of RA vs. “PE” of SA ($P<0.001$). Note: indicate areas in which change in signal intensity did not reach the significance threshold (ie, $P<0.001$) or no activation at all; Abbreviations: R, right; L, left; BA, Brodmann Area.

	RA>SA				PE of RA>“PE” of SA			
	Talairach's				Talairach's			
Anatomy	x	y	z	t value	x	y	z	t value
R-Superior Frontal Gyrus BA 8	-	-	-	-	6	40	54	3.68
R-Superior Frontal Gyrus BA 9	-	-	-	-	2	24	46	4.24
L-Superior Frontal Gyrus BA 6	-10	18	66	7.87	-10	18	66	5.04
L-Superior Frontal Gyrus BA 8	-38	16	50	3.31	-6	28	50	3.68
L-Superior Frontal Gyrus BA 9	-	-	-	-	-20	40	26	6.15
L-Superior Frontal Gyrus BA 10	-	-	-	-	-14	62	24	3.19
R-Middle Frontal Gyrus BA 46	28	48	8	8.10	28	50	6	7.06
L-Middle Frontal Gyrus BA 9	-	-	-	-	-30	18	42	3.48
L-Middle Frontal Gyrus BA 46	-28	44	10	9.05	-	-	-	-
R-Inferior Frontal Gyrus BA 47	-	-	-	-	26	18	-14	3.44
R-Precentral Gyrus BA 4	-	-	-	-	28	-28	60	3.76
R-Postcentral Gyrus BA 1	48	-24	56	4.06	-	-	-	-
R-Superior Parietal Lobule BA 7	12	-70	40	4.53	12	-68	38	3.86
L-Superior Parietal Lobule BA 7	-32	-68	48	4.36	-34	-68	52	3.40
R-Inferior Parietal Lobule BA 39	-	-	-	-	52	-56	28	5.09
R-Inferior Parietal Lobule BA 40	50	-56	34	4.95	52	-54	36	4.99
L-Inferior Parietal Lobule BA 39	-	-	-	-	-40	-52	24	3.86
L-Inferior Parietal Lobule BA 40	-48	-38	54	3.94	-56	-56	30	3.29
R-Middle Temporal Gyrus BA 21	68	-20	2	3.70	64	-4	-18	4.03
L-Middle Temporal Gyrus BA 21	-	-	-	-	-22	64	10	3.46
L-Inferior Temporal Gyrus BA 20	-	-	-	-	-54	-34	-16	4.02
R-Temporal Pole BA 38	-	-	-	-	44	24	-28	3.95
L-Temporal Pole BA 38	-42	26	-22	4.43	-	-	-	-
R-Transverse Temporal Gyrus (Superficial Part) BA 42	58	-24	6	4.28	-	-	-	-
R-Occipital Lobe BA 17	-	-	-	-	8	-100	6	4.05
R-Occipital Lobe BA 18	34	-70	-2	4.07	-	-	-	-
L-Occipital Lobe BA 18	-	-	-	-	-22	-80	4	3.58
R-Occipital Lobe BA 19	34	-82	6	4.71	34	-82	6	4.71
L-Occipital Lobe BA 19	-34	-70	34	5.47	-30	-66	-	4.43
R-Cerebellum	6	-84	-20	9.82	44	-46	-30	3.97
L-Cerebellum	-44	-66	-36	10.06	-48	-66	-22	5.25
Anterior Cingulate Gyrus BA 24	0	-10	34	6.41	-2	8	30	3.50
Anterior Cingulate Gyrus BA 32	-	-	-	-	-26	44	14	6.31
Posterior Cingulate Gyrus BA 23	0	-42	24	6.17	0	-40	22	4.51
Posterior Cingulate Gyrus BA 31	0	-42	30	4.99	-6	-42	28	4.21
R-Parahippocampal Gyrus BA 28	30	6	-22	6.32	-	-	-	-
R-Parahippocampal Gyrus BA 27	-	-	-	-	22	-28	-4	5.00
L-Parahippocampal Gyrus BA 28	-28	8	-20	5.65	-30	6	-20	5.00
R-Hippocampus	-	-	-	-	24	-12	-10	3.25
Rostrum of Corpus Callosum	-	-	-	-	-8	24	14	4.35
Splenium of Corpus Callosum	-	-	-	-	2	-36	18	6.11
R-Insula	36	-2	-4	5.51	34	-4	-2	3.22
L-Insula	-20	10	16	3.82	-	-	-	-
R-Lentiform Nucleus	20	0	10	4.40	18	16	2	4.04
L-Lentiform Nucleus	-16	6	4	3.24	-20	-8	12	3.58
R-Thalamus	16	-18	14	4.08	14	-26	2	4.03
L-Thalamus	-16	-14	14	3.57	-	-	-	-
L-Caudate Nucleus (Head)	-	-	-	-	-16	16	16	3.49

precentral gyrus (BA 4), temporoal pole (BA 38), hippocampus, insula and thalamus; contralateral head of nucleus caudate, corpus callosum and ACG (BA 24, 32) and PCG (BA 23, 31) (Fig. 1, Table 1).

Table 2. Brain areas activated by acupuncture at SA and “PE” of SA ($P < 0.001$). Note: indicate areas in which change in signal intensity did not reach the significance threshold (ie, $P < 0.001$) or no activation at all; Abbreviations: R, right; L, left; BA, Brodmann Area.

	RA>SA				PE of RA> “PE” of SA			
	Talairach's				Talairach's			
Anatomy	x	y	z	t value	x	y	z	t value
R-Superior Frontal Gyrus BA 8	-	-	-	-	4	32	36	5.01
R-Superior Frontal Gyrus BA 10	-	-	-	-	24	60	8	4.02
L-Superior Frontal Gyrus BA 6	-4	18	70	7.21	-30	16	60	3.46
L-Superior Frontal Gyrus BA 9	-	-	-	-	-18	54	30	4.97
L-Superior Frontal Gyrus BA 10	-	-	-	-	-24	60	16	55.73
L-Superior Frontal Gyrus BA 12	-	-	-	-	-58	8	-8	6.25
R-Middle Frontal Gyrus BA 9	48	18	42	3.72	52	22	40	3.80
R-Middle Frontal Gyrus BA 9	-	-	-	-	-44	14	40	3.65
R-Inferior Frontal Gyrus BA 47	52	42	-4	8.79	-	-	-	-
L-Inferior Frontal Gyrus BA 47	-38	40	0	4.75	46	44	-8	6.22
R-Precentral Gyrus BA 4	10	-24	72	6.60	-	-	-	-
R-Superior Parietal Lobule BA 7	42	-66	44	4.21	-	-	-	-
L-Superior Parietal Lobule BA 7	-36	-60	56	4.99	-40	-62	56	4.57
L-Inferior Parietal Lobule BA 40	-64	-24	20	7.45	-64	-24	20	4.07
R-Superior Temporal Gyrus BA 22	64	0	0	9.30	64	-2	-2	4.66
L-Superior Temporal Gyrus BA 22	-58	8	-8	8.55	-	-	-	-
R-Inferior Temporal Gyrus BA 37	56	-64	-4	5.79	-	-	-	-
R-Superficial Part of Transverse Temporal Gyrus BA 42	66	-20	14	8.59	66	-12	10	4.03
L-Transverse Temporal Gyrus (Superficial Part) BA 42	-	-	-	-	-66	-28	12	3.29
R-Cerebellum	46	-56	-28	6.17	46	-60	-30	5.57
L-Cerebellum	-38	-76	-20	4.66	-28	-72	-50	5.44
Anterior Cingulate Gyrus BA 24	0	12	30	4.16	-	-	-	-
Anterior Cingulate Gyrus BA 32	-8	30	24	3.94	6	28	24	3.47

It can be seen that the results during acupuncture at Liv3 and PE of RA showed an activation pattern in the visual cortex, limbic system and subcortical gray structures. Except for ACG activation, there was no significant activation at those areas above mentioned during sham acupuncture.

Acupuncture at sham acupoint (8 subjects) resulted in the activation of bilateral cerebella, inferior frontal gyrus (BA 47), SPL (BA 7), superior temporal gyrus (BA 22); contralateral superior frontal gyrus (BA 6), IPL; isolateral precentral gyrus (BA 4), medial frontal gyrus (BA 9), superficial part of transverse temporal gyrus (BA 42), inferior temporal lobe (BA 37) and ACG (BA 24, 32) (Fig. 1, Table 2). The brain activation during “PE” of SA were bilateral cerebella, frontal lobe (BA 6, 8, 9, 10, 12, 47), superficial part of transverse temporal gyrus (BA 42); contralateral SPL (BA 7) and IPL (BA 40); isolateral superior temporal lobe (BA 22) and ACG (BA 32) (Table 2).

4 Discussion

4.1 Subjective Effects

The “*De-Qi*” is a unique sensation of numbness, tingling, fullness, and dull ache that develops at the site of acupuncture and may spread some distance from

the acupuncture point during needle manipulation. Presence of the *De-Qi* state, which is a prerequisite for and often an indicator of a clinical acupuncture effect [2]. After experiment, the questionnaire was performed on the subjects. All subjects experienced the *De-Qi* sensation to some degree during needle manipulation.

One of the aim of this study was to test the post-effect of acupuncture, and then we hypothesize the activated brain areas during PE of RA were similar with those during RA, or a little difference.

We mentioned above the post-effect would influence the results of acupuncture. However the exact duration of post-effect was unknown. Hence, a single block experiment design was adopted in this study to avoid this issue. In order to avoid shortcoming of single block, we added the period of acupuncture at Liv3 up to 3 min, corresponding to 3 or more block times at a general multiple block design, so that we can detect efficiently activated signals of acupuncture.

4.2 Brain Activation

Some scholars suggested that needle sensory signal of acupuncture traveled upwards into brain via spinal cord, and then went upwards to brain cortex via thalamus exchanging neuron; finally, the needle sensory was formed [9]. But other scholars thought that acupuncture at single point might activate bilateral thalami, which suggested that the activation associated with not only sensory but also attention and vigilance. In this study, we found bilateral thalami were activated during RA and PE of RA, which may be consistent with the viewpoint of the former; thus, thalamus probably was the specific brain areas of acupuncture.

During PE of RA, there was no activation on primary sensor cortex and second sensor cortex, because of no extra stimulation on acupuncture spot at all. Otherwise we presumed that the other reasons may be the long scanning time and lower intensity of signal.

It is well accepted that insula has an unspecific sensation integration function to multiple harmless stimulation, such as olfactory, taste or tactile sensation [16]. Based on fMRI studies and neuroanatomy knowledge, the activation on insula during RA or PE of RA implied that insula is a complex sensation area and key integration cortex.

ACG, which is thought to be part of the attention system accompanying pain sensation, participates in many complex somatic and visceral motor functions and pain response [17]; while PCG takes charge of supervisory sense, stereoscopic orientation and memory [18-19]. In general, ACG relates to the evaluation and modulation to the sensation and cognition information, the rostrum of ACG relates to supplementary feeling or psychological reaction [20-21]. Our findings indicated that ACG was activated during RA and PE of RA, which was in agreement with many results on acupuncture studies using fMRI [4,7]. We consider that the activation of ACG may be a complex integration of attention and feeling response to acupuncture stimulation. ACG may be not only the sensation to pain but also analgesia of acupuncture.

According to the theory of TCM about following meridian convey, Liv3 connects ocular system and then upwards to frontal lobe and vertex of the head. In our study, the extensive activation on visual cortex (Brodmann's area, BA 17) and visual communication cortex (BA 18, 19) were seen and obvious activation on bilateral frontal lobe during RA than those during sham acupuncture, which may be associated with this theory. In addition, many scholars reported that the acupoints specificity existed as mentioned above. Liv3 is one of important acupoints for the treatment of eye disorders suggested by TCM. In our study, we used blinder and dimmed light during MR scanning in order to avoid directly visual stimulation. It should be noticed that visual cortex was activated during RA and PE of RA. Meanwhile there was no activation on visual cortex during SA and "PE" of SA. Siedentopf et al [22] found that laser acupuncture on eye-related acupoints could activate the visual cortex, too. Whereas Gareus et al [23] proposed opposite opinion that no BOLD-response correlating with acupuncture was detected in the visual cortex. We concluded the activation may be the specific brain areas of acupuncture at Liv3 for visual cortex activation being seen in the study. Whether the theory of following meridian convey was associated with the mechanism of acupuncture or not, it needed further studies.

It is well known that cerebella not only are related to the modulation of voluntary motion, but also execute attention tasks based on time; the cerebella also participate to manage complex social or feeling actions and form complex network together with other related brain areas, which play an important role on modulation to the social cognition actions. Our findings about cerebella activation are in agreement with many previous studies reported [5-7]. In this study, bilateral cerebella were activated more obviously during RA or PE of RA than those during SA or "PE" of SA. Hence, the roles of cerebella on the mechanism of acupuncture should not be ignored.

The activation on prefrontal cortex (PF) and posterior part of parietal lobe was associated with the modulation to stimulation position and encodement in cognitive actions [16]. There is inter-modulation on the specific areas between PF and mediodorsal nucleus of thalamus, moreover, they form one system to decide the feeling reaction at present based on past experiences. In this study, there was activation on brain areas above, especially on bilateral PF and mediodorsal nucleus of thalami, which was directly reaction to the acupuncture at Liv3. Bilateral PF was also activated during SA, but the extent was smaller. There was no activation on mediodorsal nucleus of thalami during SA. So, the results implied that there was lack of intercommunication among brain areas during SA. Hence thalamus may be the specific brain areas to acupuncture.

In this study, the markedly flaky activation on bilateral PF, cerebella, thalami, lentiform nucleus, subcortical gray structures (such as bilateral head of caudate nucleus) and limbic system (bilateral hippocampus, temporal pole and insula) was visualized. The activation pattern during PE of RA was similar to those during RA, but the intensity and extent were smaller. The activation on bilateral PF during PE of RA was more obvious than those during "PE" of SA. Thus, we inferred that the post-effect of acupuncture existed.

In summary, acupuncture at Liv3 during RA or PE of RA activated obviously subcortical gray structures, limbic system (bilateral hippocampus, temporal pole and insula) and visual area cortex, while there was no significant activation during SA. Only activation on ACG was seen during SA. Hence we conclude acupuncture maybe modulated brain actions by activating subcortical gray structures and limbic system to treat the diseases. We consider subcortical gray structures, limbic system and visual area cortex may be the specific brain activation to Liv3 modulating within the human brain, moreover, the roles of cerebella on acupuncture mechanism should not be ignored.

We emphasize that single block experiment design can avoid the influence of post-effect of acupuncture. During manual acupuncture, acupuncturist should manipulate according to even method to keep consistent frequency and intensification of acupuncture. Our findings indicate that fMRI can visualize the specific brain response to the acupuncture and the post effect, which will establish the foundations of later scientific design in acupuncture experiments.

Acknowledgements

National Natural Science Foundation of China supports this study (grant number 90209030).

References

1. NIH Consensus Conference. Acupuncture. *JAMA* 280(17), 1518–1524 (1998)
2. Wu, M.T., Hsieh, J.C., Xiong, J., et al.: Central Nervous Pathway for Acupuncture Stimulation: Localization of Processing with functional MR Imaging of the Brain—preliminary Experience. *Radiology* 212, 133–141 (1999)
3. Wu, M.T., Sheen, J.M., Chuang, K.H., et al.: Neuronal Specificity of Acupuncture Response: A fMRI Study with Electroacupuncture. *Neuroimage* 16, 1028–1037 (2002)
4. Zhang, W.T., Jin, Z., Cui, G.H., et al.: Relations Between Brain Network Activation and Analgesic Effect Induced by Low vs. High Frequency Electrical Acupoint Stimulation in Different Subjects: A functional Magnetic Resonance Imaging Study. *Brain Research* 982, 168–178 (2003)
5. Zhang, W.T., Jin, Z., Luo, F., et al.: Evidence from Brain Imaging with fMRI Supporting Functional Specificity of Acupoints in Humans. *Neuroscience Letters* 354, 50–53 (2004)
6. Biella, G., Sotgiu, M.L., Pellegata, G., et al.: Acupuncture Produces Central Activations in Pain Regions. *Neuroimage* 14, 60–66 (2001)
7. Hsieh, J.C., Tu, C.H., Chen, F.P., et al.: Activation of the Hypothalamus Characterizes the Acupuncture Stimulation at the Analgesic Point in Humans: A Positron Emission Tomography Study. *Neuroscience Letters* 307, 105–108 (2001)
8. Hui, K.K.S., Liu, J., Makris, N., et al.: Acupuncture Modulates the Limbic System and Subcortical Gray Structures of the Human Brain: Evidence From fMRI Studies in Normal Subjects. *Hum. Brain Mapp.* 9, 13–25 (2000)
9. Li, Y.Q., Chen, J.H.: *JCAM*. 19(4), 56 (2003)

10. Joannie, S.H., Daniel, W.H.: Functional Neuroimaging: Review of Methods and Applications to Acupuncture Research. *Clinical Acupuncture and Oriental Medicine* 2, 163–167 (2001)
11. Fang, J.L., Krings, T., Weidemann, J., et al.: Functional MRI in Healthy Subjects During Acupuncture: Different Effects of Needle Rotation in Real and False Acupoints. *Neuroradiology* 46, 359–362 (2004)
12. Friston, K.J., Holmes, A.P., Poline, J.B., et al.: Analysis of fMRI Time-series Revisited. *Neuroimage* 2, 45–53 (1995)
13. Noppency, U., Price, C.J.: Functional Imaging of the Semantic System: Retrieval of Sensory-experienced and Verbally Learned Knowledge. *Brain and Language* 84, 120–133 (2003)
14. Friston, K.J., Holmes, A.P., Price, C.J., et al.: Multisubject fMRI Studies and Conjunction Analyses. *Neuroimage* 10, 385–396 (1999)
15. Talairach, J., Tournoux, P.: *Co-planar Stereotaxic Atlas of the Human Brain*. Thieme, New York (1988)
16. Baron, R., Baron, Y., Disbrow, E., et al.: Brain Processing of Capsaicin-induced Secondary Hyperalgesia. A functional MRI Study. *Neurology* 53, 548–557 (1999)
17. Davis, K.D., Taylor, S.J., Crawley, A.P., et al.: Functional MRI of Pain and Attention Related Activations in the Human Cingulate Cortex. *J. Neurophysiol.* 77, 3370–3380 (1997)
18. Vogt, B.A., Ro sene, D.L., Pandya, D.N.: Thalamic and Cortical Afferents Differentiate Anterior from Posterior Cingulate Cortex in the Monkey. *Science* 204, 205 (1979)
19. Vogt, B.A., Finch, D.M., Olson, C.K.: Functional Heterogeneity in Cingulate Cortex: the Anterior Executive and Posterior Evaluative Regions. *Cerebral Cortex* 2, 435 (1992)
20. Rainville, P., Duncan, G.H., Price, D.D., et al.: Pain Affect Encoded in Human Anterior Cingulated But Not Somatosensory Cortex. *Science* 277, 968–971 (1997)
21. Casey, K.L.: Forebrain Mechanisms of Nociception and Pain: Analysis Through Imaging. *Proc. Natl Acad Sci, USA* 96, 7668–7674 (1999)
22. Siedentopf, C.M., Golaszewski, S.M., Mottaghy, F.M., et al.: Functional Magnetic Resonance Imaging Detects Activation of the Visual Association Cortex During Laser Acupuncture of the Foot in Humans. *Neuroscience Letters* 327, 53–56 (2002)
23. Gareus, I.K., Lacour, M., Schulte, A.C., et al.: Is There a Bold Response of the Visual Cortex on Stimulation of the Vision-related Acupoint GB 37? *Journal of Magnetic Resonance Imaging* 15, 227–232 (2002)

A Role for Signal Propagation Through the Hippocampal CA2 Field in Memory Formation

Yuko Sekino^{1,3} and Tomoaki Shirao²

¹ Division of Neuronal Network, Institute of Medical Science
University of Tokyo, Tokyo, Japan

² Department of Neurobiology and Behavior
Gunma University Graduate School of Medicine, Maebashi, Japan

³ CREST, JST, Kawaguchi, Japan
yukos@ims.u-tokyo.ac.jp

Abstract. Memory is very vivid and long-lasting, particularly when we have strong emotions, such as pleasure, love, fun, surprise, fear, anxiety and anger. This fact leads us to imagine that memory storage in the hippocampus can be switched according to the emotion-related brain state. In this paper, we first review the two classical concepts for the hippocampal network. One is the trisynaptic circuit connecting the entorhinal cortex, the dentate gyrus, the CA3 field and the CA1 field. The other is the lamellar hypothesis which proposes that the lamellar organization of the trisynaptic circuit. We then summarize our studies highlighting a role of the hippocampal CA2 field in memory formation. Finally we propose a novel circuit in the hippocampal network, which contains the CA3-CA2-CA1 pathway in which the CA2 field functions as a gate controlled by the activity of the supramammillary nucleus to control signal propagation between hippocampal lamellar organizations.

1 Introduction

When we learn something new, we pay attention to it and repeat it until we remember it. In contrast, when we see something aimlessly, we often fail to remember it. In this sense, human memory is different from the memory function of computers, which stores information independent of context. Because we could not remember all the enormous amount of sensory information that we perceive, this selectivity of human memory is adaptive. However, once we judge information to be important, it is possible to memorize it instantly and clearly. We use the phrase “I can remember it like a picture” to express how clear such memories can be. This kind of memory is formed easily, particularly when we have strong emotions such as pleasure, love, fun, surprise, fear, anxiety and anger. Sometimes such memories are called “flashbulb memory”. We know from our own experience that a flashbulb memory is very vivid and long-lasting. A famous example of a flashbulb memory, which is often found in textbooks of Neuroscience, is that “many people have strong memories of where they were and what they were doing when they first heard the news of President Kennedy assassination”.

An example of our recent flashbulb memory is the death of Britain's Princess Diana: Some of our friends can recall clearly what they were doing when hearing this news on the radio. Usually we don't keep in mind for such a long time the details of the circumstances surrounding us. Such experiences illustrate the importance of emotion for formation of long-lasting memory. In this paper, we address brain mechanisms that may be involved in such emotion-related memory formation.

Memory that one can recall and declare is called as declarative memory. The hippocampus and the adjacent perirhinal, entorhinal, and parahippocampal cortices are known to play important roles in declarative memory. Memory circuits related to declarative memory can be switched on or enhanced by emotions. How does emotion switch on or enhance memory processing in the hippocampus? Here, we term the state which emotion can enhance memory as the "brain's emotional state". The brain's emotional state is like motivation for memory, because it motivates one to remember things. Neuronal activity associated with the brain's emotional state may greatly change hippocampal information processing and accelerate memory formation.

Synaptic plasticity in the hippocampus, such as long-term potentiation (LTP), is widely assumed to represent the synaptic mechanisms of learning and memory. Because LTP is experimentally reproduced in slice preparations, molecular mechanisms of LTP have been elucidated in great detail. However, the brain's emotional state that we are concerned about here is presumably associated with different mechanisms than those responsible for LTP, because formation of a flashbulb memory does not require repetitive stimulation which is necessary to induce LTP. Are there any experimental models which reflect the enhancement of information processing during the brain's emotional state? In the present paper, we propose an experimental model, using a slice preparation, that allows study of the relationship between emotion and hippocampal information processing.

The fact that certain things can be memorized, out of all the information that constantly bombards us, gives the impression that the brain's emotional state opens a gate to retaining information that will be memorized. If a hippocampal slice preparation could be used to study the mechanisms involved in opening such a gate, this gate could be elucidated rapidly. Here we summarize our efforts to achieve this goal. Using a rat hippocampal slice preparation, we have found gate-like behavior that inhibits the spread of neuronal activity in the hippocampus. We also discuss how such a gating mechanism can enhance hippocampal memory functions.

2 Anatomical Backgrounds for Understanding the Neuronal Circuit in the Hippocampus

We first review classical concepts of the hippocampal network, which are based on anatomical and electrophysiological studies, to lay the foundation for understanding of our new studies.

2.1 Anatomical Studies Using the Golgi Method

The oldest and most popular histological drawing of the hippocampus is a sketch drawn more than 100 years ago by Santiago Ramón y Cajal [1]. In this sketch, pyramidal neurons and axons were drawn, and the hypothetical directions of impulses propagation in the hippocampus were indicated by arrows. Ramón y Cajal largely divided the hippocampal pyramidal cell layer into the regio superior and the regio inferior [1]. Later, Rafael Lorente de Nó named CA1 for the regio superior, and divided the regio inferior into two parts named CA2 and CA3 (CA is derived from Cornu Ammonis) [2]. The nomenclature of Lorente de Nó is widely used today when describing hippocampal cell fields.

2.2 Hippocampal Trisynaptic Circuit and the Lamellar Hypothesis

Our concept of the basic pathway of excitatory synaptic transmission in the hippocampus is largely derived from the histological studies of Ramón y Cajal and Lorente de Nó. In this view, inputs from the entorhinal cortex enter into the hippocampal dental gyrus and project to the CA1 field via the CA3 field. Outputs from the CA1 field then return to the entorhinal cortex. This pathway is termed the “trisynaptic circuit” because it includes three synapses formed by the perforant path, the mossy fibers, and the Schaffer collaterals. Andersen et al. verified this basic synaptic circuit in anesthetized rabbits by electrophysiological recordings [3]. They showed that neuronal activity in the hippocampus is transmitted in one direction along the trisynaptic circuit and that the strongest responses to stimulation of the entorhinal cortex are recorded in the CA1 field at a right angle to the longitudinal axis of the hippocampus. Thus, they proposed “the lamellar hypothesis” that “small strips of the hippocampal cortex may operate as independent functional units” [3].

2.3 Verification of the Lamellar Hypothesis

Detailed investigation of the rat hippocampus, using anterograde and retrograde tracer labeling methods, have revealed that the projections of axons from CA3 pyramidal cells is widely spread, in a fan shape, to cover about two thirds of the hippocampus [4, 5]. In addition, it has been shown that neurons in the entorhinal cortex terminate not only on the dental gyrus granule cells but also on CA1 and CA3 pyramidal cells. Thus, synaptic connections in the intrahippocampal network are more complicated than had been proposed originally. However, the simplicity of the concept still makes the trisynaptic circuit a very attractive model for the basic circuitry of the hippocampal network. Andersen et. al repeatedly verified the lamellar hypothesis by using hippocampal slice preparations that were cut in sheets parallel to the pyramidal cell layers [6]. They confirmed that population field potentials evoked by stimulation of CA3 pyramidal cell axons are largest in lamellar regions almost at a right angle to the longitudinal axis of the hippocampus.

3 Studies Using Hippocampal Slice Preparations

We next will summarize our studies using rat hippocampal slices that highlight a role for the CA2 field in hippocampal function. We propose that the CA2 field has a function that is distinct from that of other hippocampal fields (CA3 and CA1) and that this region functions as a gate that is regulated by the brain's emotional state.

3.1 Usefulness of the Slice Preparation for Studying Hippocampal Networks

The physiological data that serve as the basis for the lamellar hypothesis were obtained under experimental conditions where the hippocampus received no input from other brain regions. Thus, it is possible that signal propagation may be different in the hippocampus of intact animals. This possibility is very likely because CA3 neurons send their projections widely and other brain regions, such as the hypothalamus, the thalamus and the septum, also send projections to the hippocampus. To demonstrate that the excitatory pathway of the hippocampus can be controlled by emotional state, simultaneous recording of neuronal activity at multiple sites of the hippocampus should be performed in awake and freely-moving animals. However, with such approaches, it is difficult to analyze in detail the synaptic pathways involved or the underlying transmission mechanisms. Slice preparations are much more suitable for such detailed analyses. Because long-distance and polysynaptic connections are cut in conventional slice preparations, the nature of the trisynaptic circuit has not been studied in slice preparations. However, if the trisynaptic circuit remains intact to some extent in slice preparations, it should be possible to examine how excitatory synaptic transmission is modified by electrical and pharmacological stimuli that mimic the brain's emotional state.

3.2 Electrophysiological Verification of Synaptic Transmission from CA3 to CA1 in the Slice Preparation

To estimate whether the trisynaptic circuit is intact in a slice preparation, we recorded from the CA1 pyramidal cell layer the field potentials that result from electrical stimulation of mossy fibers [7]. Oblique slices, 450- μm in thickness, were made from the central part of the rat hippocampus, oblique to the longitudinal axis of the hippocampus; that is, at an angle of 30-45 along the alvear fibers running on the ventral surface. A bipolar tungsten electrode was inserted into the hilus to stimulate the entire bundle of mossy fibers coming from the dentate gyrus. Electrical stimulation (150 μs duration and 0.6 mA intensity) was applied to induce maximal amplitude of population spikes (PSs) in CA3b pyramidal cell layer. Latencies of PSs recorded in CA1 pyramidal cell layer were widely distributed - from 7 to 13 ms - with two peaks at 9 ms and 11.5 ms, showing the presence of two groups of synaptic input to CA1 with rapid and slow latencies [7]. Thus, the presence of two peaks indicates that two pathways exist between the

CA3 and CA1 cell layers. Presumably the pathway with slower latency is different from the faster, mono-synaptic pathway arising from the Schaffer collaterals.

3.3 Optical Recording of Membrane Potential Changes

With conventional electrophysiological methods, it is difficult to obtain spatial information, such as the routes through which electrical activity passes. However, optical measurement of membrane potential is well suited to overcome this technical limitation. Due to great improvements in dyes [8] and other sensors [9, 10], optical measurements are now widely used to observe how neuronal excitation spreads in space. For optical measurement of membrane-potential changes, hippocampal slices are stained with a fluorescent dye that is sensitive to membrane potential. Changes in the intensity of transmitted light or fluorescence emission are then measured with an optical recording system which consists of a CCD- or metal oxide (MOS)-imaging sensor mounted on a microscope. The ratio of images obtained before and after stimulation can then be calculated to indicate spread of membrane depolarization, which is usually represented by a pseudo-color display where green to red corresponds to the intensity of the responses.

3.4 Two Distinct Pathways from CA3 to CA1

We then performed optical recordings to visualize the routes of the slower-latency pathway between CA3 and CA1 [7]. The same experimental procedures used for the electrophysiological experiments described above were applied to hippocampal slices stained with a membrane-potential-sensitive dye, RH482 (0.02%). We used a camera system (SD1001, Fujifilm Microdevices) with a MOS imaging sensor of 128 x 128 pixels (9 x 9 mm² size) and acquired images at intervals of 0.6 ms. The slice preparation was mounted on a recording chamber in an inverted microscope (Olympus), and images were collected with a 5X objective lens. Optical signals corresponding to propagation of membrane depolarization from the stimulation site throughout the whole CA3 field were obtained for 4.8 ms after mossy fiber stimulation. The time course of the spread of these optical signals in the CA3 field was consistent between all experiments. Two different patterns were observed later than 6 ms images. In one pattern, termed fast propagation, optical signals were observed in the CA1 field 6 ms after mossy fibers stimulation and propagated to the central part of CA1 7.2 ms after stimulation. In this pattern, optical signals were not detected in the CA2 field. The other pattern was termed slow propagation and in this case optical signals were observed in the CA2 field 6 ms after stimulation. During slow propagation, optical signals in the CA2 field gradually increased in amplitude up to 7.2 ms and then spread to the CA1 field at 8.4 ms. Strong optical signals in the central part of the CA1 layer were observed 10.8 ms after stimulation. Thus, during slow propagation, optical signals were observed in the center part of CA1 field 3 ms later than during fast propagation [7]. The slow propagation observed by the optical recording probably corresponds to the slower-latency PSs recorded in CA1 (see Section 3.2).

3.5 Optical Signals in the CA2 Field

To summarize the studies described above, optical recoding of membrane potential indicated two different pathways for transmission of electrical signals from CA3 to CA1. While the CA2 field does not contribute to fast propagation, it apparently does play a role in slow propagation. Fast propagation is thought to be mediated by the Schaffer collaterals, which are the long axon collaterals of CA3 neurons that terminate on CA1 neurons. Propagation of optical signals from CA3 to CA1 could be observed because many Schaffer collaterals remain intact in the oblique slices used in our experiments. Such propagation is consistent with the trisynaptic circuit classically thought to be the basic hippocampal network.

In contrast to the fast propagation, slow propagation requires the CA2 field. Signal propagation from CA3 to CA2 is probably due to short axon collaterals of CA3 neurons that are different from the Schaffer collaterals. In six of fourteen slices which showed slow propagation, weak optical signals corresponding to fast propagation were observed in the striatum radiatum of CA1. This indicates that an insufficient number of Schaffer collaterals remained intact to produce fast PSs. During slow propagation, optical signals that come from CA2 spread not only along the stratum radiatum but also within the stratum oriens of CA1. The propagation of signals from CA2 to CA1 that were observed during slow propagation is consistent with the projection of CA2 neuron axons [11].

4 Functional Role of the CA2 Field

The contributions of the CA2 field to the intrahippocampal circuit and the hippocampal functions have been ignored [12]. In fact, the CA2 field often is regarded as a distal part of the CA3 field. Our results using hippocampal slices suggest that the CA2 field possesses a functional role that is distinct from that of the other fields, specifically that CA2 contributes to signal propagation from CA3 to CA1. Histological studies have also suggested that the CA2 field has distinct features and we next summarize some of these anatomical observations.

4.1 Anatomical Features of the CA2 Field

The CA2 field is composed of large and small pyramidal cells and classically is called a “Mischzone” (mix zone). Lorente de Nó identified the CA2 field on the basis of several anatomical features, and predicted that the characteristics of the CA2 fields would be clearer in the future [2]. Recently, it has been found that parts of CA3a and CA2 fields receive nerve projections from the supramammillary nucleus (SuM) of the hypothalamus [13,14]. This might provide CA2 neurons with some distinct features.

CA2 neurons are known to be resistant to epileptic seizure and to delayed cell death during ischemia. Thus, in pathological terms, the CA2 field is called a ‘protective zone’ [15, 16, 17]. Immunohistochemistry reveals that several substances are expressed highly in the CA2 field. It is interesting that the hippocampal CA2

field of mature rats highly expresses nerve growth factors, such as NT-3 [18,19] and basic FGF [20,21]. In this regard, CA2 is different from CA3 and CA1.

In hippocampal slice preparations, the CA2 field can be identified easily because of differences in the thickness of its cell layer. Compared with regularly packed small pyramidal cells in the CA1 field, the cell layer becomes wider in CA2. The CA3/CA2 border is the point at which the mossy fiber bundle becomes very thin and fades out. Therefore, unlike the description by Lorente de Nó, the CA2/CA1 border is more readily distinguished than the CA3/CA2 border.

Though the CA2 field can be identified in the slice preparation, this field is very narrow in the rat hippocampus and therefore has been thought to serve as a transitional area between CA1 and CA3. This has caused the CA2 field not to be involved in memory circuitry of the hippocampus. Although the rat hippocampal CA2 field is narrower and smaller in size compared to CA2 of monkey and human, the CA2 field has common features that are shared among animal species. Thus the CA2 field is clearly different from the CA3 and CA1 fields and cannot be ignored when considering the function of hippocampal neuronal networks.

4.2 Adenosine A1 Receptors in the CA2 Field

Recently, we found that the density of adenosine A1 receptors (A1R) is higher in the CA2 field than in other hippocampal fields [22]. Although adenosine is well known as a potent inhibitory modulator, neurons releasing adenosine have not yet been identified. There are two sources of extracellular adenosine: either from metabolism of intracellular adenosine triphosphate (ATP), which is released through cellular membranes after hyperexcitation [23], or from metabolism of extracellular ATP [24] which is co-released with other neurotransmitters and is then metabolized by ectoprotein kinases. Because application of an A1R antagonist enhances neuronal activity, it seems that endogenous adenosine exerts inhibitory influences on neuronal activity [25]. Thus, the dense expression of A1R makes it likely that tonic suppression of neuronal activity by endogenous adenosine is stronger in the CA2 field than in other fields, so that changes in extracellular adenosine levels could control the involvement of the CA2 field in hippocampal neuronal circuitry.

4.3 Neuronal Activity of the CA2 Field Is Suppressed by Endogenous Adenosine

Within a few days after injection of kainic acid into the rat cerebral ventricle, neuronal death occurs but is limited to the hippocampal CA3 field [26]. Because kainic acid receptors are distributed in the CA3 field [27], CA3 neurons are thought to degenerate due to excitotoxicity mediated by these receptors. However, it has not been clear why CA1 neurons - which receive direct input from CA3 neurons - are spared during such excitotoxic hyperexcitation of CA3 neurons. Based on a hint that came from our studies of the distribution

of A1R, Matsuoka et al intraperitoneally administrated an A1R antagonist, 8-cyclopentyl-theophylline (8-CPT), and then infused kainic acid intracerebroventricularly [28]. This treatment expanded the field of neuronal death to include not only CA3 but also CA2 and CA1 neurons. This suggests that A1R activity, which is enriched in the CA2 field, serves to spare CA2 and CA1 neurons from the cell death caused by kainic-acid inducing hyperexcitation of CA3

When we made slice preparations transversely, at a right angle to the long axis of the hippocampus, stimulation of mossy fibers did not evoke PSs in the CA1 field (unpublished data). This result, which differs from what is observed in the oblique slice, has been thought to be due to only a small number of fibers projecting from CA3 neurons in the transverse slice [29]. However, when the same stimulus is given in the presence of 8-CPT, PSs can be evoked in the CA1 field (unpublished data). Thus, these preliminary data show that endogenous adenosine inhibits propagation of neuronal activity from CA3 to CA1. Further experiments to examine the effect of 8-CPT on excitatory transmission between CA3 and CA1 in the transverse slice with the optical recording of membrane potential changes will be required to firmly establish this point. neurons.

4.4 A Gating Function of the CA2 Field

Signal propagation along the CA3-CA2-CA1 pathway probably is controlled by activity of A1R. Because of the dense distribution of A1R in the CA2 field, CA2 neurons might have a higher threshold for firing. Signal propagation along the CA3-CA2-CA1 pathway presumably is regulated by the balance between the strength of synaptic input and the amount of tonic inhibition by adenosine. If inhibition through A1R is stronger than the excitatory inputs from CA3, then the CA3-CA2-CA1 pathway will not be activated. Therefore, CA2 neurons can behave as a gate between the CA3 and CA1 regions. Further, CA2 can also behave as a detector in Ohsuga's theoretical model of human language acquisition [30,31], as will be described elsewhere in this issue.

Because CA2 neurons have short axons and spread in the direction of the long axis of the hippocampus [11], CA2 neurons can propagate signals to the longitudinal direction, across the lamellar organization of the hippocampus. Once signal propagation involves CA2 neurons, trisynaptic circuits in hippocampal layers cannot be independent of each other.

Pharmacological studies show that adenosine or its analogs induce sleep whereas adenosine antagonists stimulate neuronal activity. Interestingly, it is known that the level of extracellular adenosine in the rat hippocampus changes in parallel with behavior, that is, decreases in overall movements and consummatory behavior are observed when adenosine levels reaches their maxima in the hippocampus [32]. Taking all of the evidence together, it seems that the CA3-CA2-CA1 pathway will not function during sleep states. This is consistent with the fact sleepiness usually disrupts formation of memories. When we feel sleepy, the level of emotion declines. However, the CA3-CA2-CA1 pathway might function in awaking brain. Thus, we assume that signal propagation beyond the lamella plays an important role in emotion-related memory formation.

5 How Do Emotions Control the Hippocampal Gating Function?

We will propose one possible mechanism for how emotions regulate hippocampal function. This proposal is based on the supramammillary nucleus (SuM) of the hypothalamus, a brain region that is related to anxiety and projects directly to the dentate gyrus and the CA2 field of the hippocampus.

5.1 The Supramammillary Nucleus of the Hypothalamus and the Hippocampal Activity

Nerve terminals of SuM neurons are found in the CA2 pyramidal cell layer and in the dentate gyrus granule cell layer [13]. These terminals form excitatory synapses with CA2 pyramidal cells and with dentate gyrus granule cells [14]. In addition, it is known that prestimulation of the SUM facilitates perforant path-dentate PSs [33]. Activity of SuM neurons, therefore, is likely to regulate hippocampal activity. However, it has been unclear how and what kinds of hippocampal functions are regulated by SuM activity.

SuM activity has been suggested to be participate in fear or anxiety. It has been shown that the number of SuM neurons that are Fos-positive in response to conditioned stress decreases at the lowest dose of an anxiolytic drug, diazepam [34]. Further, SuM activity is known to be in synchrony with the hippocampal theta rhythm following stimulation of the reticular nucleus pontis oralis (RPO) [35]. The theta rhythm, which is one of the best-characterized oscillatory patterns of the hippocampal formation [36], is a slow rhythmic electroencephalogram (EEG) signal that is recorded when rats explore a novel place.

Taken together, we hypothesize that SuM activity evoked by emotions such as fear and anxiety serves to enhance hippocampal function.

5.2 The SuM Controls a Gating Function of the Dentate Gyrus

Because of the locations of the SuM and the hippocampus, it is very difficult to make a slice preparation that contains both the SuM and the hippocampus or that keeps their connections intact. We examined the role of SuM activity in controlling the hippocampal activity using spread of the epileptic discharges induced by kainic acid [37]. Injection of kainic acid into the lateral and basolateral amygdala of rats induced limbic seizures; subsequent spread of epileptic discharges was then detected using immunohistochemistry to detect c-Fos expression. Fos-positive neurons were observed in the entorhinal cortex and in all fields of the hippocampus (dentate gyrus, as well as the CA3, CA2 and CA1 fields). When muscimol, a GABA agonist, was injected bilaterally into the SuM, to inhibit the SuM activity prior to injection of kainic acid, the hippocampal EEG became flat and limbic seizures were inhibited in conjunction with suppression of hippocampal c-Fos expression. Because c-Fos expression in the entorhinal cortex was not changed by the muscimol injection, we assume that epileptic discharge

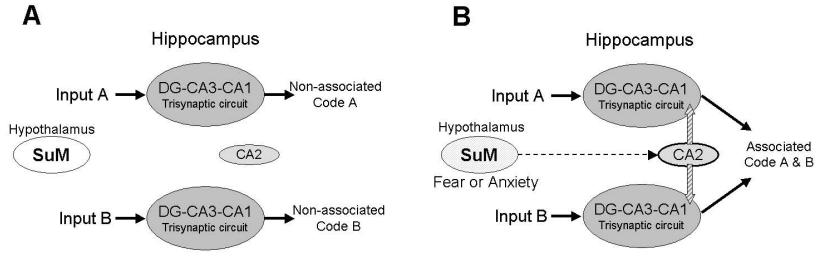


Fig. 1. Schematic drawing of the relation between the activity of the supramammillary nucleus (SuM) and the hippocampal information processing. (A) Information processing through the trisynaptic circuit on the hippocampal lamellar organization without the SuM activity. Input A and B which simultaneously go in to the hippocampus are not associated. (B) When the SuM is activated by fear or anxiety, CA2 neurons associate Input A and B beyond lamellar organizations.

propagation to the hippocampus is inhibited at the site of the dentate gyrus. In other words, an inhibitory mechanism is present at the dentate gyrus which might be released by SuM activity. Thus, we have successfully indicated a gating function of the dentate gyrus which is controlled by the SuM activity. Although it is difficult to show that the CA2 field has a gating mechanism related to that of the dentate gyrus, observations that the threshold for c-Fos expression following limbic seizures is highest in the CA2 field and that suppression of the c-Fos expression in response to muscimol-induced blockade of SuM activity is greatest in the CA2 field supports our hypothesis of a CA2 gating function.

6 Conclusion

We hypothesize that the presence of two distinct intrahippocampal pathways explains how emotions strengthen memory formation. Our proposal is that the CA3-CA1 pathway is used for memory formation in non-emotional state, while the CA3-CA2-CA1 pathway is activated in emotional states. Because A1R are densely expressed in the CA2 field, removal of adenosine inhibition activates the CA3-CA2-CA1 pathway. When adenosine level in the brain increases, for example when we feel sleepy, the CA3-CA2-CA1 pathway would not be activated. This could explain why sleepiness disturbs memory formation. Furthermore, the SuM, which is thought to be activated by fear or anxiety, sends fibers to the CA2 field and to the dentate gyrus. Fear or anxiety awakens the brain and activates the CA3-CA2-CA1 pathway, and would thereby enhance memory formation. Figure 1 summarizes our hypothesis of how the CA2 field regulates hippocampal information processing. When inputs A and B simultaneously go through along the separate trisynaptic circuits within lamella, these inputs are not associated (Fig. 1A). In contrast, these inputs can be associated by CA2 activity, after fear or anxiety activates the SuM (Fig. 1B). In other words, the CA3-CA2-CA1

pathway is switched by emotions. This hypothesis for how the emotional state of the brain enhances memory formation will require further tests in the future.

Acknowledgements

The authors thank Dr. George J. Augustine for critical comments and excellent editing the manuscript, and Ms. Naoko Numata for assistance on creating the figure.

References

1. Ramón y Cajal, S.: *Histology of the Nervous System of Man and Vertebrates* (History of Neuroscience, No 6) Oxford University Press, USA, translated from "Textura del Sistema Nervioso del Hombre y de los Vertebrados (Moya, Madrid, 1899, 1904) (in Spanish)" by Swanson, N. and Swanson, L (1994)
2. de Nó Lorent, R.: Studies on the structure of the cerebral cortex II. Continuation of the study of the ammonic system. *Psychol. Neurol.* 46, 113–177 (1934)
3. Andersen, P., Bliss, T.V., Skrede, K.K.: Lamellar organization of hippocampal pathways. *Exp. Brain Res.* 13, 222–238 (1971)
4. Ishizuka, N., Weber, J., Amaral, D.G.: Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat. *J. Comp. Neurol.* 295, 580–623 (1990)
5. Li, X.G., Somogyi, P., Ylinen, A., Buzsaki, G.: The hippocampal CA3 network: an in vivo intracellular labeling study. *J. Comp. Neurol.* 339, 181–208 (1994)
6. Andersen, P., Soleng, A.F., Raastad, M.: The hippocampal lamella hypothesis revisited. *Brain Res.* 886, 165–171 (2000)
7. Sekino, Y., Obata, K., Tanifuji, M., Mizuno, M., Murayama, J.: Delayed signal propagation via CA2 in rat hippocampal slices revealed by optical recording. *J. Neurophysiol.* 78, 1662–1668 (1997)
8. Grinvald, A., Fine, A., Farber, I.C., Hildesheim, R.: Fluorescence monitoring of electrical responses from small neurons and their processes. *Biophys. J.* 42, 195–198 (1983)
9. Takashima, I., Ichikawa, M., Iijima, T.: High-speed CCD imaging system for monitoring neural activity in vivo and in vitro, using a voltage-sensitive dye. *J. Neurosci. Methods* 91, 147–159 (1999)
10. Vranesic, I., Iijima, T., Ichikawa, M., Matsumoto, G., Knopfel, T.: Signal transmission in the parallel fiber-Purkinje cell system visualized by high-resolution imaging. *Proc. Natl. Acad. Sci. USA* 91, 13014–13017 (1994)
11. Tamamaki, N., Abe, K., Nojyo, Y.: Three-dimensional analysis of the whole axonal arbors originating from single CA2 pyramidal neurons in the rat hippocampus with the aid of a computer graphic technique. *Brain Res.* 452, 255–272 (1988)
12. Lopes da Silva, F.H., Witter, M.P., Boeijinga, P.H., Lohman, A.H.: Anatomic organization and physiology of the limbic cortex. *Physiol. Rev.* 70, 453–511 (1990)
13. Haglund, L., Swanson, L.W., Kohler, C.: The projection of the supramammillary nucleus to the hippocampal formation: an immunohistochemical and anterograde transport study with the lectin PHA-L in the rat. *J. Comp. Neurol.* 229, 171–185 (1984)

14. Magloczky, Z., Acsady, L., Freund, T.F.: Principal cells are the postsynaptic targets of supramammillary afferents in the hippocampus of the rat. *Hippocampus* 4, 322–334 (1994)
15. Cavazos, J.E., Das, I., Sutula, T.P.: Neuronal loss induced in limbic pathways by kindling: evidence for induction of hippocampal sclerosis by repeated brief seizures. *J. Neurosci.* 14, 3106–3121 (1994)
16. Mattson, M.P., Guthrie, P.B., Kater, S.B.: Intrinsic factors in the selective vulnerability of hippocampal pyramidal neurons. *Prog. Clin. Biol. Res.* 317, 333–351 (1989)
17. Williamson, A., Spencer, D.D.: Electrophysiological characterization of CA2 pyramidal cells from epileptic humans. *Hippocampus* 4, 226–237 (1994)
18. Phillips, H.S., Hains, J.M., Laramée, G.R., Rosenthal, A., Winslow, J.W.: Widespread expression of BDNF but not NT3 by target areas of basal forebrain cholinergic neurons. *Science* 250, 290–294 (1990)
19. Wetmore, C., Olson, L.: Expression and regulation of neurotrophins and their receptors in hippocampal systems. *Hippocampus* 3, 171–181 (1993)
20. Matsuyama, A., Iwata, H., Okumura, N., Yoshida, S., Imaizumi, K., Lee, Y., Shiraishi, S., Shiosaka, S.: Localization of basic fibroblast growth factor-like immunoreactivity in the rat brain. *Brain Res.* 587, 49–65 (1992)
21. Woodward, W.R., Nishi, R., Meshul, C.K., Williams, T.E., Coulombe, M., Eckenstein, F.P.: Nuclear and cytoplasmic localization of basic fibroblast growth factor in astrocytes and CA2 hippocampal neurons. *J. Neurosci.* 12, 142–152 (1992)
22. Ochiishi, T., Saitoh, Y., Yukawa, A., Saji, M., Ren, Y., Shirao, T., Miyamoto, H., Nakata, H., Sekino, Y.: High level of adenosine A1 receptor-like immunoreactivity in the CA2/CA3a region of the adult rat hippocampus. *Neuroscience* 93, 955–967 (1999)
23. Mitchell, J.B., Lupica, C.R., Dunwiddie, T.V.: Activity-dependent release of endogenous adenosine modulates synaptic responses in the rat hippocampus. *J. Neurosci.* 13, 3439–3447 (1993)
24. Wieraszko, A., Goldsmith, G., Seyfried, T.N.: Stimulation-dependent release of adenosine triphosphate from hippocampal slices. *Brain Res.* 485, 244–250 (1989)
25. Alzheimer, C., Sutor, B., ten Bruggencate, G.: Disinhibition of hippocampal CA3 neurons induced by suppression of an A1R-mediated inhibitory tonus: pre- and postsynaptic components. *Neuroscience* 57, 565–575 (1993)
26. Nadler, J.V., Perry, B.W., Gentry, C., Cotman, C.W.: Degeneration of hippocampal CA3 pyramidal cells induced by intraventricular kainic acid. *J. Comp. Neurol.* 192, 333–359 (1980)
27. Petralia, R.S., Wang, Y.X., Wenthold, R.J.: Histological and ultrastructural localization of the kainate receptor subunits, KA2 and GluR6/7, in the rat nervous system using selective antipeptide antibodies. *J. Comp. Neurol.* 349, 85–110 (1994)
28. Matsuoka, Y., Okazaki, M., Takata, K., Kitamura, Y., Ohta, S., Sekino, Y., Taniguchi, T.: Endogenous adenosine protects CA1 neurons from kainic acid-induced neuronal cell loss in the rat hippocampus. *Eur. J. Neurosci.* 11, 3617–3625 (1999)
29. Sayer, R.J., Friedlander, M.J., Redman, S.J.: The time course and amplitude of EPSPs evoked at synapses between pairs of CA3/CA1 neurons in the hippocampal slice. *J. Neurosci.* 10, 826–836 (1990)
30. Ohsuga, S.: Symbol Processing by Non-Symbol Processor. In: Foo, N.Y., Göbel, R. (eds.) *PRICAI 1996*. LNCS, vol. 1114, Springer, Heidelberg (1996)
31. Ohsuga, S.: Intelligence for upgrading information (in this book) (2007)

32. Huston, J.P., Haas, H.L., Boix, F., Pfister, M., Decking, U., Schrader, J., Schwartz, R.K.: Extracellular adenosine levels in neostriatum and hippocampus during rest and activity periods of rats. *Neuroscience* 73, 99–107 (1996)
33. Mizumori, S.J., McNaughton, B.L., Barnes, C.A.: A comparison of supramammillary and medial septal influences on hippocampal field potentials and single-unit activity. *J. Neurophysiol.* 61, 15–31 (1989)
34. Beck, C.H., Fibiger, H.C.: Conditioned fear-induced changes in behavior and in the expression of the immediate early gene *c-fos*: with and without diazepam pretreatment. *J. Neurosci.* 15 (1 pt 2), 709–720 (1995)
35. Kirk, I.J.: Supramammillary neural discharge patterns and hippocampal EEG. *Brain Res. Bull.* 42, 23–26 (1997)
36. Vertes, R.P., Kocsis, B.: Brainstem-diencephalo-septohippocampal systems controlling the theta rhythm of the hippocampus. *Neuroscience* 81, 893–926 (1997)
37. Saji, M., Kobayashi, S., Ohno, K., Sekino, Y.: Interruption of supramammillo-hippocampal afferents prevents the genesis and spread of limbic seizures in the hippocampus via a disinhibition mechanism. *Neuroscience* 97, 437–445 (2000)

Genetic Granular Cognitive Fuzzy Neural Networks and Human Brains for Pattern Recognition

Cui Lin¹, Jun Li², Natasha Barrett³,
Yan-Qing Zhang², and David A. Washburn³

¹ Department of Computer Science, Wayne State University
Detroit, MI 48202 USA

cuilin@wayne.edu

² Department of Computer Science, Georgia State University
Atlanta, GA 30302-3994 USA

junlig@gmail.com, yzhang@cs.gsu.edu

³ Department of Psychology, Georgia State University
Atlanta, GA 30303 USA

nbarrett78@yahoo.com, dwashburn@gsu.edu

Abstract. With ever-improving information technologies and high performance computational power, recent techniques in granular computing, soft computing and cognitive science have allowed an increasing understanding of normal and abnormal brain functions, especially in the research of human's pattern recognition by means of computational intelligence. It is well understood that normal brains have high intelligence to recognize different geometrical patterns, but a systematic framework of biological neural network has not yet be established. In this paper, we propose the genetic granular cognitive fuzzy neural networks (GGCFNN) in order to efficiently testify artificial neural networks' learning capability on human's pattern recognition in term of symmetric and similar geometry patterns. In contrast to other information systems, the GGCFNN is a highly hybrid intelligent system integrating the techniques of genetic algorithms, granular computing, and fuzzy neural networks with cognitive science for pattern recognition. Our ability to simulate biological neural networks makes it possible a more comprehensive quantitative analysis on the pattern recognition of human brains, and our preliminary experiment results would shed lights on the future research of cognitive science and brain informatics.

1 Introduction

With ever-improving information technologies and high-performance computational power, complexity and scale of brain informatics have continued to grow at an explosive pace. Recent advances in granular computing, soft computing and cognitive science have allowed an increase understanding of normal and abnormal brain functions, especially in the research of human's pattern recognition by means of computational intelligence. It is well understood that normal

brains have high intelligence to recognize different geometrical patterns in terms of the similarity and symmetry, but a systematic framework of biological neural network has not yet been established. In addition, there is a common problem in cognitive science and computer science: How does the rule-like relational learning emerges (or why does it sometimes fail to emerge?) from the associations yielded by experience? In cognitive science's point of view, this problem is addressed by challenging organisms with novel demands and determining how brains and behaviors are adaptively altered. In computer science, a set of training data is given to a computational model, then the problem is addressed by investigating how the networks glean useful knowledge from the given training data set and then use the discovered knowledge to make right decisions in real applications.

Since human brains consist of biological neural networks, which are known as the major components for performing pattern recognition, it is very important to develop advanced algorithms for exploring the mechanism of biological neural networks and the artificial neural networks when recognizing different geometry patterns.

Recently, hybrid neural networks such as fuzzy neural networks and granular neural networks have been investigated to solve complex application problems with high dimensionality [2,4,5,6,7,8,9,10,14]. Such hybrid neural networks provide remarkable abilities to derive knowledge from complicated or imprecise datasets. Therefore, they have already been applied to extract patterns and detect trends that are too complex to be noticed by either human beings or other computer techniques. In this case, the trained neural networks can be thought of as an "expert" in the category of information that has been given to analyze.

In general, the key technology for solving the curse of dimensionality (including structural dimensionality and parametric dimensionality) of a hybrid soft computing system, such as a Fuzzy Neural Network (FNN), is the divide-and-conquer method. In other words, an original single-stage FNN with high system complexity can be divided into small multi-stage soft computing building blocks with low system complexity so as to reduce both structural dimensionality and parametric dimensionality significantly.

Based on the fuzzy neural network tree [12] and cognitive science, this paper proposes a hybrid intelligent system called Granular Cognitive Fuzzy Neural Network (GCFNN). In order to improve learning performance, genetic algorithms are applied to optimize the GCFNN, hence forms more effective Genetic Granular Cognitive Fuzzy Neural Network (GGCFNN). The GGCFNN provides parameters for deciding the number of iterations and termination criteria in order to control its learning performance. In addition, the hybrid genetic forward-wave-backward-wave learning algorithm is developed to enhance the system's learning quality. Our simulation results generated by the GGCFNN system are compared to human beings and are analyzed in terms of computer science and cognitive science.

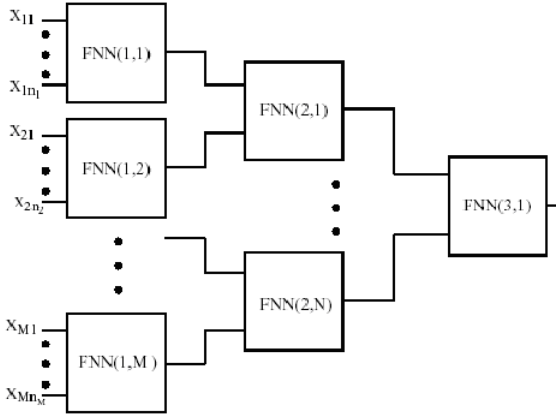


Fig. 1. Architecture of a 3-stage GCFNN

2 Granular Cognitive Fuzzy Neural Networks

Similar to our human brains, which have a variety of hierarchically and discretely distributed components to perform different functions, the GCFNN consists of multiple sub-stages, and each sub-stage is built up by many single-stage FNNs using the normal fuzzy reasoning.

2.1 Architecture of the GCFNN

In general, the k_{th} -stage FNN uses outputs generated from the relevant $k - 1_{th}$ -stage FNNs as inputs, and then generates an output that becomes an input of relevant $k + 1_{th}$ -stage FNNs. The final stage FNN will generate a final output. For example, Figure 1 shows the 3-stage GCFNN. The $FNN(i, j)$ is the i_{th} -stage- j_{th} FNN. More specifically, $FNN(1, j)$ is used as input, where $j = 1, 2, \dots, M$; similarly, $FNN(2, l)$ for $l = 1, 2, \dots, N$ uses the outputs generated from relevant 1st-stage FNNs as inputs, and finally $FNN(3, 1)$ generates the final output y . Therefore, the final output y is a function of all original inputs $x_{j1}, x_{j2}, \dots$, for $j = 1, 2, \dots, M$.

2.2 GCFNN Algorithms

Suppose that the GCFNN has K inputs and one output. Given input data vectors X^p (i.e., $X^p = (x_1^p, x_2^p, \dots, x_K^p)$) and an output data vector for Y^p for $p = 1, 2, \dots, L$. The global energy function is defined by

$$GE^p = \frac{1}{2} [f(x_1^p, \dots, x_K^p) - Y^p]^2$$

The basic learning algorithm for GCFNN was described in [12]. Here, for clarity and convenience, the relevant algorithms are introduced as follows. A

Step 1: Heuristic Initialization of Parameters.

Step 2: Gradient Descending Learning from Data.

Then we can get the following learning algorithms
for $i = 1, 2, \dots, N, k = 1, 2, \dots, M, p = 1, 2, \dots, N,$
 $t = 0, 1, 2, \dots,$ and a learning rate $\lambda > 0$.

Step 2.1: Train b^k .

$$b^k(t+1) = b^k(t) - \lambda \left. \frac{\partial E}{\partial b^k} \right| t.$$

Step 2.2: Train η^k .

$$\eta^k(t+1) = \eta^k(t) - \lambda \left. \frac{\partial E}{\partial \eta^k} \right| t.$$

Step 2.3: Train a_i^k .

$$a_i^k(t+1) = a_i^k(t) - \lambda \left. \frac{\partial E}{\partial a_i^k} \right| t.$$

Step 2.4: Train σ_i^k .

$$\sigma_i^k(t+1) = \sigma_i^k(t) - \lambda \left. \frac{\partial E}{\partial \sigma_i^k} \right| t.$$

Step 2.5: Train $wleft_i^k$ or $wright_i^k$.

IF $x_i^k \leq a_i^k$
THEN

$$wleft_i^k(t+1) = wleft_i^k(t) - \lambda \left. \frac{\partial E}{\partial wleft_i^k} \right| t$$

ELSE

$$wright_i^k(t+1) = wright_i^k(t) - \lambda \left. \frac{\partial E}{\partial wright_i^k} \right| t.$$

Fig. 2. The Procedure of Local Forward-Wave Learning Algorithm

new hybrid genetic-algorithms-based learning algorithm will be proposed in subsection C.

A. Local Forward-Wave Learning

In general, suppose that a FNN has N inputs and 1 output. Given input data vectors X^p (i.e., $X^p = (x_1^p, x_2^p, \dots, x_N^p)$) and 1-dimensional output data vector y^p for $p = 1, 2, \dots, N$. The local energy function is defined by

$$E^p = \frac{1}{2} [f(x_1^p, \dots, x_N^p) - y^p]^2$$

For simplicity, let E denotes E^p . Based on the learning algorithm in [10][11], the basic steps of the local forward-wave learning algorithm for a FNN are given in Figure 2.

B. Global Backward-Wave Learning

Based on the traditional back-propagation learning method, the global backward-wave learning algorithm is proposed to train all local FNNs in the

Step 1: Calculate a back-propagation error δ^1 to the relevant $(l-1)th$ stage FNNs based on an error $\delta^{(l+1)}$ from the $(l+1)th$ stage FNN

$$\delta^l = \frac{\partial E}{\partial y^{l-1}} = \delta^{l+1} \frac{\partial y^l}{\partial y^{l-1}}.$$

where y^l and $y^{(l-1)}$ are outputs of the lth stage FNN and the $(l-1)th$ stage FNN, respectively.

Step 2: Update parameters using the back-propagation error δ^l .

Step 3: Train $wleft_i^k$ or $wright_i^k$.

Step 4: Discover Fuzzy Knowledge.

Fig. 3. The Procedure of Global Backward-Wave Learning Algorithm

back-propagation manner. The following procedure shown in Figure 3 is a general algorithm for updating parameters of any FNN in GCFNN.

C. Genetic Forward-Wave-Backward-Wave Learning

The new hybrid learning algorithm for training the GCFNN is called genetic forward-wave-backward-wave learning algorithm, which combines the techniques of genetic algorithms, local forward-wave learning and global backward-wave learning, as shown in Figure 4.

Step 1: Partition the original training data sets input data vectors X^p and an output data vector Y^p for $p = 1, 2, \dots, N$ into M local training data sets where M is the number of 1st-stage FNNs.

Step 2: Initial the Parameter Optimization Using Genetic Algorithms. All parameters of the genetic granular cognitive fuzzy neural network are optimized initially by using genetic algorithms.

Step 3: Call the Local Forward-Wave Learning.

Step 4: Call the Global Backward-Wave Learning.

Fig. 4. The Procedure of Genetic Forward-Wave-Backward-Wave Learning Algorithm

2.3 The GGCFNN Algorithm

Based on the parameters and energy function defined in Section 2.2, a genetic algorithm is designed to search the minimum output of the energy function. The procedure of our genetic algorithm starts with an initialization phase, which randomly generates initial populations. During each generation, the GGCFNN produces the next population from the current population by applying genetic operators sequentially, viz., selection, crossover, mutation, evaluation and elitist. The sequence of operators is run iteratively to produce one generation after

another until a termination condition is reached. Each solution in the population is associated with a figure of merit depending on the fitness function. A gene in this genetic algorithm is defined as a data structure covering all parameters that should be optimized. The best population in each generation will be selected, and all its corresponding parameters will be saved.

3 Simulations and Analysis

The simulations of the GGCFNN system for pattern recognition are divided into two categories: One is designed for pattern similarity, which tests whether the noised pattern images can still be recognized as the similar to its original version, after horizontal or vertical transformation; Another category targets to symmetry tests: whether the GGCFNN can identify vertically or horizontally reversed pattern pairs with noise disturbance.

During each test, specific pattern samples are used for training and other groups of samples are served for testing. Each pattern sample is depicted by 7×9 pixel matrix, so totally 63 inputs are required in order to present one pattern, as shown in Figure 5. For each sample group, multiple pairs of patterns are randomly chosen within varied noise ranges (100% noise denotes that half points of a standard sample are reversed).

To accommodate the input requirement, the GGCFNN's granular layer has 9 local GCFNNs with 7 inputs, so that 63 pixel inputs are satisfied, as demonstrated in Figure 5: in the second hidden layer, each GCFNN has one output; the output layer has 9 inputs and 1 output. Therefore, the total number of inputs of the system is 126.

Finally, the system will make a final decision whether or not two images in each testing sample are similar or symmetric based on the maximum output of the three GGCFNNs. Meanwhile, human beings are chosen to make their subjective judgment on pattern samples, by answering what degree the two patterns are similar or symmetric to each other.

3.1 Simulations for Similar Patterns

With regard to the similarity, we conduct two scenarios for simulations. Each scenario presents a typical case when people recognize patterns, so that their corresponding results are used to train and testify GGCFNN's recognition performance.

3.1.1 Scenario 1

The Scenario 1 includes two pattern groups: Group A contains 50 pairs of arrow pattern samples within the following five noise levels, which are 0%-10%, 10%-20%, 20%-30%, 30%-40%, 40%-50% respectively. In each pair of arrow pattern sample, as the Figure 6-A, the second pattern is transformed from the first pattern after horizontal reversal at the first, and then added into the same level of noises.

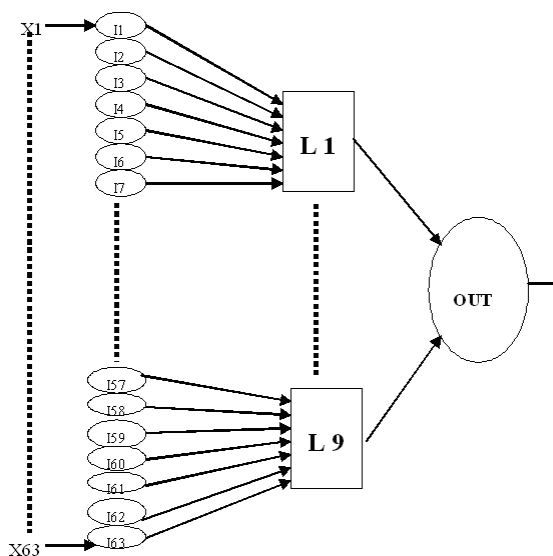


Fig. 5. The GGCFNN structure for pattern recognition

Similarly, in Figure 6-B, Group B includes 50 pairs of cup-like pattern samples within the same five noise levels as Group A. In each pair of cup-like-pattern sample, the second pattern is transformed from the first pattern after vertical reversal and then added into the same level of noises.

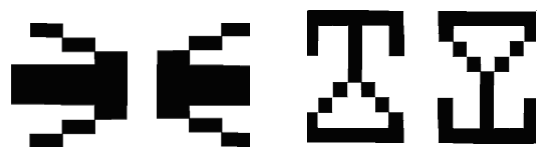


Fig. 6. A: in Group A, the second arrow pattern is transformed from the first after horizontal reversal; B: in Group B, the second cup-like pattern is transformed from the first after vertical reversal

Table 1 shows the comparisons of decisions made by human beings and the GGCFNN in Scenario 1. As can be seen, the GGCFNN and the human brain

Table 1. Comparisons of GGCFNN simulation results with human beings

Group name	Different answers	Total samples	Agreement percentage
Group A	5	50	90%
Group B	22	50	56%
Total	27	100	73%

Table 2. Different answers to vertical reversal within 5 noise levels

Noise level	Different answers	Total samples	Disagreement percentage
40%-50% noise 0		10	0%
30%-40% noise 0		10	10%
20%-30% noise 1		10	10%
10%-20% noise 1		10	10%
0%-10% noise 3		10	30%

Table 3. Different answers to horizontal reversal within 5 noise levels

Noise level	Different answers	Total samples	Disagreement percentage
40%-50% noise 3		10	30%
30%-40% noise 4		10	40%
20%-30% noise 4		10	40%
10%-20% noise 4		10	40%
0%-10% noise 7		10	70%

make 73% identical choices with total 100 testing data. In Group A, Only 10% of 50 testing samples are different; while, in Group B, almost half of the testing samples, which are more than four times higher than Group A, introduce contradictive decisions to its counterpart. Thus, human and computer system in Group B make much more different answers than in Group A. It might indicate that the horizontal reversal causes more confusions than vertical reversal to humans or to the GGCFNN system.

According to Tables 2 and 3, we discovered an interesting phenomenon: as noise level increases, the disagreement on pattern’s similarity between human and the GGCFNN tends to decrease. More specifically, as shown in Table 2, disagreement percentage is exactly 0% in [40%, 50%] while it increases to 16.67% in [0%, 30%]. In Table 3, disagreement percentage is increased from 36% within the noise level [20%, 50%] to 55% in [0%, 20%]. Additionally, the disagreement percentages experience prominent increases within low noise levels.

3.1.2 Scenario 2

Scenario 2 introduces more complicated patterns for further investigation. Two



Fig. 7. A: in Group A, the second pattern is transformed from the first pattern after horizontal reversal; B: in Group B, the second pattern is transformed from the first pattern after vertical reversal

Table 4. Different answers to the pattern similarity

Group name	Different answers	Total samples	Disagreement percentage
Group A	24	100	24%
Group B	19	100	19%
Total	43	200	21.5%



Fig. 8. A: in Group A, the second hemicycle pattern is transformed from the first after horizontal reversal; B: in Group B, the second pentagon pattern is transformed from the first after vertical reversal

groups of patterns are testified with one common characteristic, as showed in Figure 7, each pattern is similar to its counterpart at the same position. Our intention is to investigate whether the GGCFNN can still identify the second pattern similar to another in each pair, after adding varied levels of noise interference.

The testing results in Table 4 demonstrate the increase of disagreement percentage, compared with scenario1. But once again, the GGCFNN exhibits better learning performance when dealing with vertical transformation than horizontal reversal.

3.2 Simulations for Symmetrical Patterns

As for the symmetry, we provided 50 pairs of hemicycle patterns with five noise ranges to test horizontal reversal; 50 pairs of pentagon patterns are given to conduct vertical reversal test, as showed in Figure 8.

From Table 5 to Table 6, the testing results illustrate again that the GGCFNN system has higher recognition capability on vertical symmetry than on

Table 5. Different answers to the symmetry of hemicycle patterns within 5 noise levels

Noise level	Different answers	Total samples	Disagreement percentage
40%-50% noise	1	10	10%
30%-40% noise	1	10	10%
20%-30% noise	1	10	10%
10%-20% noise	1	10	10%
0%-10% noise	1	10	10%

Table 6. Different answers to the symmetry of pentagon patterns within 5 noise levels

Noise level	Different answers	Total samples	Disagreement percentage
40%-50% noise	0	10	0%
30%-40% noise	0	10	0%
20%-30% noise	0	10	0%
10%-20% noise	0	10	0%
0%-10% noise	0	10	0%

horizontal symmetry. In addition, the GGCFNN shows much better recognition performance when identifying the symmetry than the similarity of the patterns, which coincides to the human’s cognitive faculty: Once a normal adult gets the rule of how to identify symmetrical patterns, usually he can easily recognize different symmetrical patterns even with noise interruption, but human beings have difficulty in providing with precise ratio on similar patterns, especially when dealing with noised images.

4 Conclusions

The GGCFNN has exhibited good learning performance when stimulating human’s pattern recognition in term of symmetry and similarity. It achieved higher agreement ratios when recognizing symmetrical patterns than similar patterns, and vertical transformation than horizontal transformation, even within high noise levels. The preliminary experiments indicated that the GGCFNN is able to simulate the human’s recognition ability and adapt its learning experience to new patterns and achieve quite similar results as human beings.

In the future, more training samples and testing samples will be used and more statistical analysis will be applied in upcoming simulations. Professional psychological surveys will be conducted to guarantee the generalization in human’s point of views. Additionally, more representative patterns will be got involved in the experiments.

Besides, other approaches will be attempted for studying the differences between Artificial Neural Network and Biological Neural Network, e.g. breaking one network unit in the system. Some of advanced data mining techniques will be applied for new discoveries.

In addition, the simplest but most straightforward examples of relational or associative learning come from comparative-cognition research with animals (e.g., the monkeys or chimpanzees in our research program). In this comparative-cognition research, specific types of problems have already been addressed to monkeys, and thus data are available both on exact training stimuli and conditions that animals had experienced, and also on the outcomes of this training in terms of the amount and the nature (relational and general or associative and specific) of the learning.

Acknowledgment

We would like to appreciate supports from GSU Research Program Enhancement Grant and GSU Brain and Behavior Research Grant.

References

1. Chen, H.P., Parng, T.M.: A new approach of multi-stage fuzzy logic inference. *Fuzzy Sets and Systems* 78, 51–72 (1996)
2. Chung, F.-L., Duan, J.-C.: On multistage fuzzy neural network modeling. *IEEE Trans. on Fuzzy Systems* 8(2), 125–142 (2000)
3. Furuhashi, T., Matsushita, S., Tsutsui, H., Uchikawa, Y.: Knowledge extraction from hierarchical fuzzy model obtained by fuzzy neural networks and genetic algorithms. In: *Proc. IEEE Int. Conf. on Neural Networks (ICNN 1997)*, Houston, pp. 2374–2379 (1997)
4. Gupta, M.M., Rao, D.H.: On the principles of fuzzy neural networks. *Fuzzy Sets and Systems* 61, 1–18 (1994)
5. Jang, J.-S.R.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Trans. on Systems, Man and Cybernetics* 23(3), 665–685 (1993)
6. Lin, C.-T.: *Neural Fuzzy Control Systems with Structure and Parameter Learning*. World Scientific, Singapore (1994)
7. Lin, C.-T., Lee, C.S.G.: ‘Neural-network-based fuzzy logic control and decision system. *IEEE Trans. on Computers* 40(12), 1320–1336 (1991)
8. Pedrycz, W., Kandel, A., Zhang, Y.-Q.: Neurofuzzy Systems. In: Dubois, D., Prade, H. (eds.) *Fuzzy Systems: Modeling and Control*, pp. 311–380. Kluwer Academic A Publishers, Dordrecht (1997)
9. Wang, L.-X.: *Adaptive Fuzzy Systems and Control Design and Stability Analysis*, PTR Prentice Hall (1994)
10. Zhang, Y.-Q., Fraser, M.D., Gagliano, R.A., Kandel, A.: Granular neural networks for numerical-linguistic data fusion and knowledge discovery, Special Issue on Neural Networks for Data Mining and Knowledge Discovery. *IEEE Trans. on Neural Networks* 11(3), 658–667 (2000)
11. Zhang, Y.-Q., Kandel, A.: *Compensatory Genetic Fuzzy Neural Networks and Their Applications*. Series in Machine Perception Artificial Intelligence, vol. 30. World Scientific, Singapore (1998)
12. Zhang, Y.-Q., Chung, F.: Fuzzy Neural Network Tree with Heuristic Back-propagation Learning. In: *Proc. of IJCNN of World Congress on Computational Intelligence 2002*, pp. 553–558, Honolulu (2002)
13. Zhang, Y.-Q., Kandel, A.: Compensatory Neurofuzzy Systems with Fast Learning Algorithms. *IEEE Trans. on Neural Networks* 9(1), 83–105 (1998)
14. Zhang, Y.-Q.: Constructive Granular Systems with Universal Approximation and Fast Knowledge Discovery. *IEEE Trans. on Fuzzy Systems* 13(1), 48–57 (2005)

Domain-Oriented Data-Driven Data Mining (3DM): Simulation of Human Knowledge Understanding

Guoyin Wang

Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, 400065, P.R. China
wanggy@ieee.org

Abstract. Recent advances in computing, communications, digital storage technologies, and high-throughput data-acquisition technologies, make it possible to gather and store incredible volumes of data. It creates unprecedented opportunities for large-scale knowledge discovery from database. Data mining (DM) technology has emerged as a means of performing this discovery. It is a useful tool in many fields such as marketing, decision making, etc. There are countless researchers working on designing efficient data mining techniques, methods, and algorithms. Unfortunately, most data mining researchers pay much attention to technique problems for developing data mining models and methods, while little to basic issues of data mining. What is data mining? What is the product of a data mining process? What are we doing in a data mining process? What is the rule we should obey in a data mining process? In this paper, we will address these questions and propose our answers based on a conceptual data mining model. Our answer would be “data mining is a process of knowledge transformation”. It is consistent with the process of human knowledge understanding. Based on analysis of the user-driven and “data-driven” data mining approaches proposed by many other researchers, a conceptual knowledge transformation model and a conceptual domain-oriented data-driven data mining (3DM) model are proposed. It integrates user-driven data mining and data-driven data mining into one system. Some future works for developing such a 3DM data mining system are proposed.

1 Introduction

Recent advances in computing, communications, digital storage technologies, and high-throughput data-acquisition technologies, make it possible to gather and store incredible volumes of data. One example is the hundreds of terabytes of DNA, protein-sequence, and gene-expression data that biological sciences researchers have gathered at steadily increasing rates. Similarly, data warehouses store massive quantities of information about various business operation aspects. Complex distributed systems (computer systems, communication networks, and

power systems, for example) are equipped with sensors and measurement devices that gather and store a variety of data for use in monitoring, controlling, and improving their operations also [1].

Data mining (also known as Knowledge Discovery in Databases - KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [2]. It uses machine learning, statistical and visualization techniques to discover knowledge from data and represent it in a form that is easily comprehensible for humans. Data mining has become a hot field in artificial intelligence. Over 10 million messages will be resulted by searching “data mining” on GOOGLE. There are numerous researchers and data miners working on designing efficient data mining techniques, methods, and algorithms every day. Data mining has become a useful and efficient technique in many domains like marketing, decision making, etc.

There are many commonly used techniques in data mining like artificial neural networks, fuzzy sets, rough sets, decision trees, genetic algorithms, nearest neighbor method, statistics based rule induction, linear regression and linear predictive coding, et al. Unfortunately, most data mining researchers pay much attention to technique problems for developing data mining models and methods, while little to basic issues of data mining. What is data mining? What is the product of a data mining process? What are we doing in a data mining process? What is the rule we should obey in a data mining process? In this paper, we will address these questions and propose our answers based on a conceptual data mining model. Our answer would be “data mining is a process of knowledge transformation”. It is consistent with the process of human knowledge understanding.

In a data mining process, we are transforming knowledge from a data format, which is not understandable for human, into another understandable symbolic format like rule, formula, theorem, etc. No new knowledge will be generated in a data mining process. That is, we are transforming knowledge from one format into another format while not producing new knowledge. Data is also a format for knowledge representation. The knowledge we mined from data was originally stored in data. Unfortunately, we can not read, understand, or use it, since we can not understand data.

Using the conceptual data mining model of this paper, a data-driven data mining model is further proposed. Some data-driven data mining algorithms are discussed [3, 4], e.g., the data driven default rule generation algorithm and decision tree pre-pruning algorithm.

Data-driven is an idea that could be applied in many fields. For example, in the linked environments for atmospheric discovery (LEAD) project [15], Plale, Huang, Kandaswamy, Pallickara, and Slominski addressed the limitations of current weather forecast frameworks through a new, service-oriented architecture capable of responding to unpredicted weather events and response patterns in real time. They proposed three services — the metadata catalog service, notification service, and workflow service — that together form the core services for managing complex experimental meteorological investigations and managing

the data products used in and generated during the computational experimentation. These services work together on the user's behalf, easing the technological burden on the scientists and freeing them to focus on more of the science that compels them.

The concept of "data-driven data mining" or "data-driven knowledge acquisition" has also been used by many researchers before [5-7]. For example, Erdem and Zimmermann adopted a data-driven analysis method to reduce the initial dimension of the phonetic and linguistic input space of a neural network [5]. With the assistance of the knowledge about concept hierarchies, data relevance, expected rule forms, and attribute-oriented induction, Han, Cai, and Cercone developed an attribute-oriented induction method for learning quantitative rules in relational databases [6]. Kusiak and Shah developed a data-driven and modular alarm-system architecture for prediction of water chemistry faults [7]. Verlinde, Cock and Boute proposed a data-driven approach with fuzzy clustering for partitioning the range of attribute values in large databases in intervals [16]. However, their methods or models are not really data-driven methods defined in this paper, while just generate or learn knowledge from data and use the generated knowledge to solve some problems. It is the normal style of data mining. In this sense, all data mining methods are "data-driven".

Many real world data mining tasks, for instance financial data mining in capital markets, are highly constraint-based and domain-oriented. Thus, it targets actionable knowledge discovery, which can afford important grounds for performing appropriate actions. Many data mining researchers proposed domain-driven or user-driven data mining methods for such tasks [8-14]. The basic ideas of user-driven data mining models are also analyzed in this paper.

At last, an integrated data mining model of user-driven and data-driven, domain-oriented data-driven data mining (3DM) is proposed according to our understanding of data mining.

The following sections of this paper are organized as followings. In Section 2, the concepts of knowledge, knowledge representation and storage are discussed. In Section 3, the basic idea of data-driven data mining is proposed. Some data-driven knowledge acquisition methods are also discussed in this section. The idea of user-driven data mining is discussed in Section 4. In Section 5, an integration of data-driven and user-driven data mining is further proposed. At last, in Section 6, we conclude this paper.

2 Knowledge, Knowledge Representation and Knowledge Storage

Knowledge could be explained as the psychological result of perception, learning and reasoning [15]. It could also be explained more specifically as follows [16]:

- It is the act or state of knowing; clear perception of fact, truth, or duty; certain apprehension; familiar cognizance; cognition.
- It is or may be known; the object of an act of knowing; cognition.

- It is gained and preserved by knowing; instruction; acquaintance; enlightenment; learning; scholarship; erudition.
- It is the familiarity which is gained by actual experience; practical skill; as a knowledge of life.
- It is the scope of information; cognizance; notice.

In artificial intelligence and information science [17], knowledge could also be explained as the objects, concepts and relationships that are assumed to exist in some area of interest. A collection of knowledge represented using some knowledge representation language is known as a knowledge base and a program for extending and/or querying a knowledge base is a knowledge-based system. Knowledge differs from data or information in that new knowledge may be created from existing knowledge using logical inference. If information is data plus meaning then knowledge is information plus processing. A common form of knowledge, e.g. in a Prolog program, is a collection of facts and rules about some subject. For example, a knowledge base about a family might contain the facts that John is David's son and Tom is John's son and the rule that the son of someone's son is their grandson. From this knowledge it could infer the new fact that Tom is David's grandson.

Knowledge exists everywhere and it is very important for our daily life and work. Knowledge could be in our brain, book, natural real world, database, etc. However, does anyone know what the knowledge is in his/her brain although everyone knows that there is much knowledge in his/her own brain? Unfortunately, no one could list all the knowledge in his/her brain. In other word, we do not know what knowledge or how much knowledge we have exactly in our brain. We could know the exact knowledge a book has. All knowledge in a book is illustrated by its text, pictures, and formula. We know some knowledge about the natural real world, for example, the law of universal gravitation. However, we may not know all the knowledge about the natural real world. Many scientists are working hard to know more about it. That is, they are "mining" the natural real world.

Knowledge could be expressed in many different ways. There are many formats to encode knowledge. The easiest format might be such symbolic formats as formula, equation, rule, and theorem. It is very easy for people to understand and use it. These formats are often used in books, documents, and even expert systems.

Natural phenomenon is another format for encoding knowledge about the natural real world. It is a kind of expression of the knowledge of the law of universal gravitation that we will drop down to the ground after jumping up.

Biological neural network (neural link) system is another natural knowledge encoding format. Everybody has much knowledge in his/her brain. There is even some knowledge in the brain of many lower animals. With such knowledge, humans and animals are intelligent and smart. The knowledge of a human or animal is encoded in the biological neural network system, or neural links, in its brain. One can not understand (or read) the knowledge in any other's brain. A channel is needed to transform knowledge between two persons. For example, one can

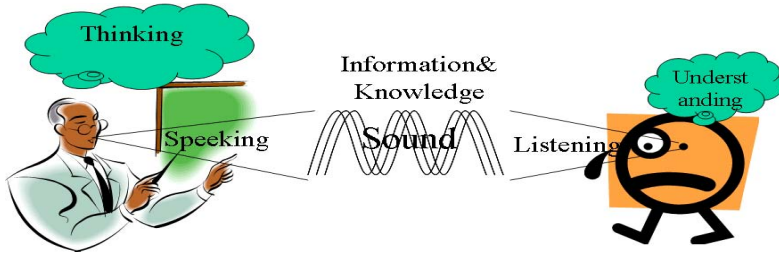


Fig. 1. Sound channel for transforming knowledge

output his knowledge by thinking and speaking and the other can accept his knowledge by listening and understanding. In this case, there is a sound channel for transforming knowledge, as shown in Figure 1. This is also a knowledge mining process. That is, we are mining knowledge from a biological neural network system, a brain. There are many other channels like visual channel, tactile channel, etc., for transforming knowledge between brains.

Data is a manmade format for encoding knowledge. There are numerals data records generated in each field. Many natural phenomenon, rules, and even human experience are recorded into data everyday. Much useful information could be concluded into data. In this sense, data mining is defined as “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” [2]. People could not read, understand, or use the knowledge expressed in data format. In the data mining process, the knowledge will be transformed into some other format understandable for human so that it could be usable.

3 Data-Driven Data Mining

Data mining tools can scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining techniques are the result of a long process of research and product development. This evolution began when data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time.

Now, let us discuss the knowledge transformation process of data mining. To understand this problem better, we would be better to have a look at the process of the knowledge transformation process between different systems.

The knowledge transformation process between different systems could be completed in many different ways. Reading and understanding are simple ways to transform knowledge from symbolic format into biological neural link format, while speaking and writing are converse processes to transform knowledge from biological neural link format into symbolic format. Translating a book in one language into another language is also a process of transforming knowledge from one symbolic format into another symbolic format. We are learning from and studying on the natural real world everyday. This is a process of transforming

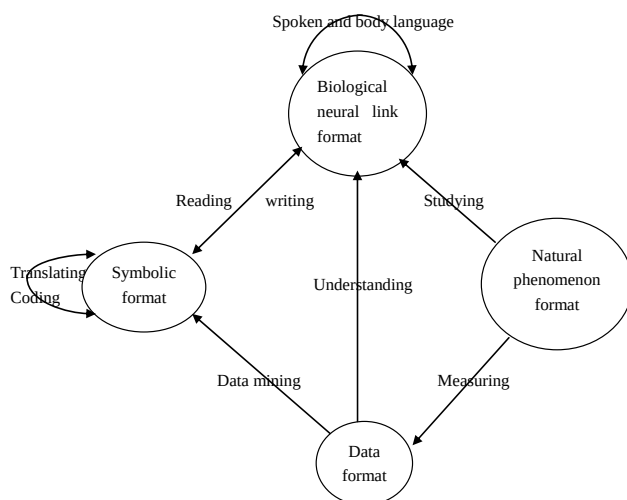


Fig. 2. Knowledge transformation among different formats

knowledge from natural phenomenon format into biological neural link format. People exchange their knowledge through spoken language and body language also. In a data mining process, we are transforming knowledge from data format into symbolic format. Data could also be taken as a measure result of the natural real world. Thus, there are many channels and ways for knowledge transformation between different systems. Figure 2 is an illustration for knowledge transformation.

From Figure 2, one can find that data mining is a kind of knowledge transformation process to transform knowledge from data format into symbolic format. Thus, no new knowledge will be generated in a data mining process. In a data mining process, knowledge is just transformed from data format, which is not understandable for human, into symbolic format, which is understandable for human and easy for application. It is similar to the process of translating a book from English into Chinese. In this translating process, the knowledge itself in the book should remain unchanged. What will be changed is just the coding format of the knowledge. That is, the knowledge of the Chinese book should be the same as the knowledge in the English one. Otherwise, there must be some mistake in the translating process. Figure 3 is an illustration for this case. Following this understanding of data mining, we could have the knowledge transformation framework for data mining as shown in Figure 4.

From Figure 4, one can find that any knowledge could be encoded into natural format, data format, symbolic format, and neural link format. That is, knowledge could be stored in a natural real world system, a data system, a symbol system, or a biological neural network system. The knowledge expressed in each format should have some properties, that is, Pi's. There should be some relationship between the different formats of the same knowledge. In order to keep the knowledge unchanged in a data mining process, properties of the knowledge

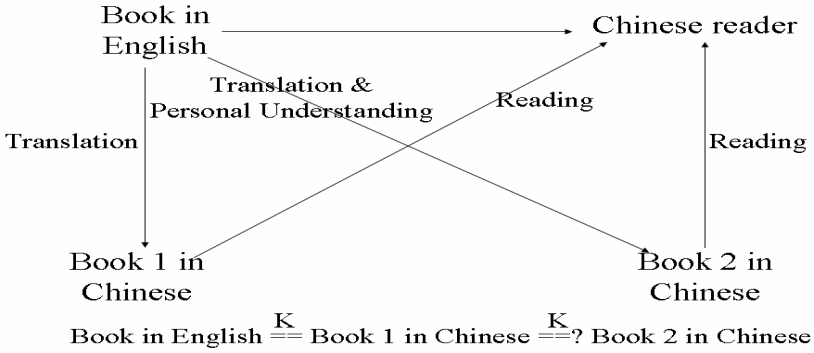


Fig. 3. Translating a book in English into Chinese

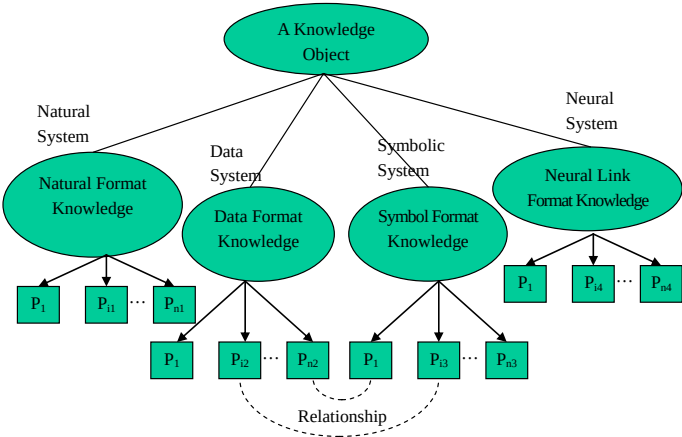


Fig. 4. Knowledge transformation framework for data mining

should remain unchanged during the knowledge transformation process. Otherwise, there should be some mistake in the knowledge transformation process.

The relationship between knowledge in natural format and data format, natural format and neural link format, symbolic format and neural link format is omitted in Figure 4. It is just like the relationship between the knowledge in data format and symbolic format.

In a data mining process, the properties of knowledge in the data format should remain unchanged. This information could provide some guideline for designing data mining algorithms. It would also be helpful for us to keep the knowledge in the data format unchanged in a data mining process. Unfortunately, the representation of knowledge is an unsolved problem in artificial intelligence. We do not know all the properties of knowledge. It is still not known how much properties are enough or needed for knowledge representation. So, how could we keep the knowledge unchanged in a data mining process? Fortunately,

we know some properties of knowledge representation, for example uncertainty of knowledge. These properties should not be changed in a data mining process in order to keep the knowledge unchanged. Thus, in order to keep the knowledge unchanged in a data mining process, we need to know some properties of the knowledge in data format, and use it to control the data mining process and keep it unchanged. This is the key idea of a data-driven data mining model. There would be three steps for designing a data-driven data mining method.

1. Select a property of knowledge which could be measured in both data format and the symbolic format for encoding knowledge generated from data.
2. Measure the property of the knowledge in the data format and the symbolic format.
3. Use the property to control the data mining process and keep it unchanged.

The knowledge property is measured in two different systems, data system and symbolic system. There might be a problem. Is the measured result of the knowledge property in data format comparable to the result from a symbolic format? If not, how could we know whether it is unchanged in the data mining process? So, we need to design a comparable measuring method for the selected property. That is, we need to establish some relationship between the knowledge property in data format and symbolic format.

In [3], we use the certainty degree of knowledge to control the data mining process. The certainty degree of knowledge in data format is measured at first. Then, propositional default decision rules are generated one after another in the order of certainty degree. The algorithm stops when all propositional default decision rules which certainty degrees are greater than the certainty degree of the data system (decision information system) are generated. With this data-driven data mining algorithm, the certainty degree of knowledge remains unchanged in the data mining process. Readers could refer to [3] for detailed information about the algorithm and experiment results. It is omitted here due to page limit.

In [4], we use another kind of certainty degree of knowledge, the global certainty of condition attribute, to control a decision tree pre-pruning process. We decide whether a tree could grow by comparing the certainty of a condition attribute class with the global certainty of the current splitting-attribute or the next splitting-attribute. The performance of our data-driven decision tree pre-pruning method based on rough set is even comparable to the post-pruning method reduced-error pruning. The detailed algorithm and experiment results are omitted here too. Readers could refer to [4] about it.

A few similar works on data-driven data mining have also been done in several application fields in recent years. For example, Frigui proposed a data-driven data preprocessing approach [17]. It took into account the intricate nature of each attribute's distribution independently of the others and then aggregated the cluster labels to create a new membership space. It allowed the complete freedom to adapt different ways to compute the concept membership functions for different attributes depending on prior knowledge and attribute type such as categorical versus numerical.

There are still many unsolved problems in data-driven data mining studies. At first, the representation of knowledge should be studied. We need to know more about the properties of different knowledge representation formats. Then, the relationship of these properties between different knowledge presentation formats should be established. At last, how to use these properties and relationship to automatically control the knowledge mining process will be another problem.

4 User-Driven Data Mining

Many real world data mining tasks, for instance financial data mining in capital markets, are highly constraint-based and domain-oriented. Thus, it targets actionable knowledge discovery, which can afford important grounds for performing appropriate actions. Some domain-driven or user-driven data mining methods for such tasks have also been developed in recent years [8-14]. In this section, we will analyze the basic ideas of these user-driven data mining methods.

Zhang, Cao and Lin proposed the domain-driven in-depth pattern discovery framework shown in Figure 5 for financial data mining in capital markets [8, 9]. Their key ideas are as follows.

1. Dealing with constraint-based context.
 - Data constraints.
 - Domain constraints.
 - Interestingness constraints.
 - Rule constraints.
2. Mining in-depth patterns.
 - In-depth patterns refer to patterns which are highly interesting and actionable in business decision-making.
 - In-depth patterns are not only interesting to data miners, but also to business decision-makers. Actionable trading strategies can be found via model refinement or parameter tuning.
3. Supporting human-machine-cooperated interactive knowledge discovery.
 - The in-depth pattern discovery is conducted under the cooperation of business analysts and data analysts.
4. Viewing data mining as a loop-closed iterative refinement process.
 - The data mining process is closed with iterative refinement and feedbacks of hypotheses, features, models, evaluation and explanations in the human-involved or centered context.

Yao and Zhao proposed an interactive user-driven classification method using a granule network also [14]. The key ideas of their method are:

1. It allows users to suggest preferred classifiers and structures.
2. It is an interactive manner between users and machines.

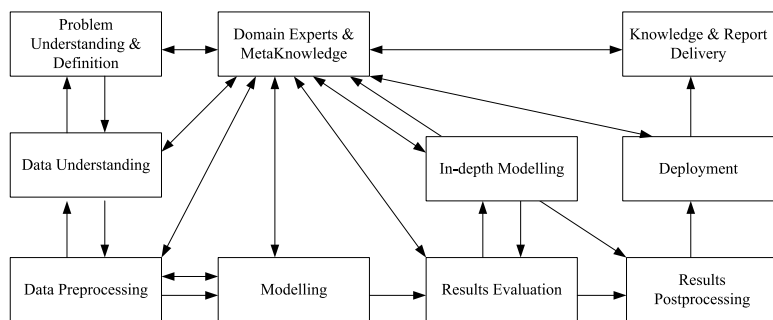


Fig. 5. Domain-driven in-depth pattern discovery framework

3. Its input and output are interleaved, like a conversation.
4. A user can freely explore the dataset according to his/her preference and priority, ensure that each classification stage and the corresponding results are all understandable and comprehensible.

Kuntz, Guillet, Lehn, and Briand also developed a human-centered process for discovering association rules where the user is considered as a heuristic which drives the mining algorithms via a well-adapted interface [10]. Han and Lakshmanan integrated both constraint-based and multidimensional mining into one framework that provided an interactive, exploratory environment for effective and efficient data analysis and mining [11]. For creating lexical knowledge bases, Patrick, Palko, Munro and Zappavigna proposed a semi-automatic approach that exploits training from a knowledgeable user to identify structural elements in the dictionary's stream of text. Once learnt from the user the structures are then applied automatically to other text streams in the same document or to other documents [12]. In semantic image classification, Dorado, Pedrycz and Izquierdo used some domain knowledge about the classification problem as part of the training procedures [13].

Through analyzing the above user-driven or domain-driven data mining methods, we find that there are some common basic ideas in these methods.

1. A user-driven data mining process is constraint based.
2. User's interesting is considered in a user-driven data mining process.
3. Prior domain knowledge is required in a user-driven data mining process.
4. Interaction between user and machine is required in a user-driven data mining process.

5 Domain-Oriented, Data-Driven Data Mining (3DM)

Does data-driven data mining conflicts with user-driven data mining? Could we integrate them into one system? We will discuss on this problem in this section.

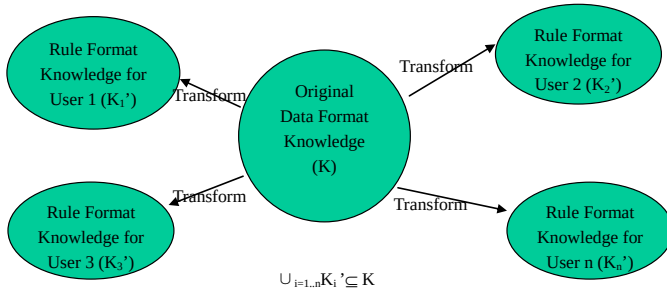


Fig. 6. Users access different knowledge from a data format knowledge base

Different users could access different data of a database from their own view. If we take data as a format of knowledge representation, a database (data set) could be also taken as a knowledge base. Therefore, the different user could find and use a different subset of the whole knowledge base for his/her task. That is, through his/her view, a user could access a subset of knowledge in the data format and transform it from data format into the format he/she required. The knowledge transformation process for each user could still be done in a data-driven manner. Figure 6 is an illustration of this understanding.

In a user-driven data mining process, user's interesting, constraint, and prior domain knowledge are very important. An interaction between user and machine is needed. The data mining process might be controlled by a user. In this case, the knowledge source of this mining process includes data and the user, while not just data. Therefore, the prior domain knowledge is also a source for the data mining process. The control of a user to the data mining process could be taken as additional input of the data mining process. It is just like the data generation process in an incremental dynamic data mining process. Furthermore, we may also deal with the user's control using incremental data-driven data mining methods. Figure 7 is an illustration of this idea.

From the discussion above, we know that a user-driven data mining does not conflict with a data-driven data mining. They could be integrated into one system. There are still a lot of works to be done to implement such a domain-oriented data-driven data mining process.

1. Designing a format for encoding prior domain knowledge.
2. Designing a format for encoding user's interestingness and constraint for a specific task.
3. Designing a format for encoding user's control.
4. Designing an incremental data-driven data mining method that could take data, prior domain knowledge, user's interestingness, user's constraint, user's control together as its input.

Therefore, there would be a long way to implement domain-oriented data-driven data mining.

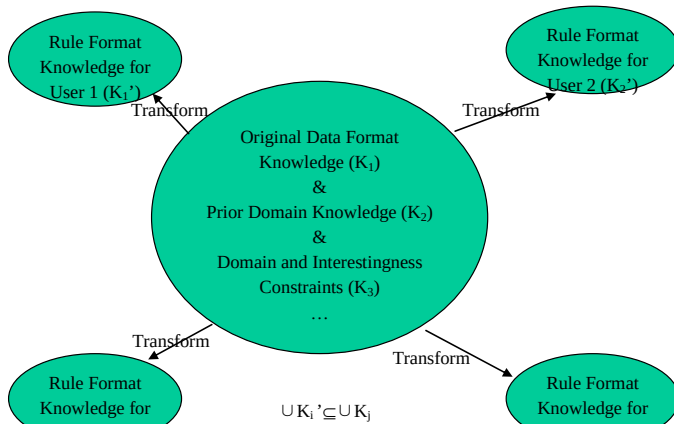


Fig. 7. User's interesting, constraint, and prior domain knowledge are all input of a data mining process (Domain-oriented data-driven data mining, 3DM)

6 Conclusion

In this paper, the essence of data mining is examined. Data mining is defined as a knowledge transformation process to transform knowledge from data format, which is not understandable for human, into symbolic format, which is understandable for human and easy to be used. We could not generate new knowledge from data base. Knowledge should not be changed in a data mining process. That is, knowledge should not be increased in the data mining process while its format could be changed. Using the relationship between the knowledge in data format and symbolic format, a data-driven data mining model is discussed. User-driven data mining methods are also discussed. A new model for domain-oriented data-driven data mining model is proposed at last.

Acknowledgments

This paper is partially supported by National Natural Science Foundation of P.R. China under Grant No.60573068, Program for New Century Excellent Talents in University (NCET), Natural Science Foundation of Chongqing under Grant No.2005BA2003, and Science & Technology Research Program of the Municipal Education Committee of Chongqing of China (No.060517).

References

1. Kumar, A., Kantardzic, M., Madden, S.: Distributed Data Mining – Framework and Implementations. *IEEE Internet Computing* 10(4), 15–17 (2006)
2. Frawley, W., Piatetsky-Shapiro, G., Matheus, C.: Knowledge Discovery in Databases: An Overview. *AI Magazine* 13, 213–228 (1992)

3. Wang, G.Y., He, X.: A Self-Learning Model under Uncertain Condition. *Journal of Software* 14(6), 1096–1102 (2003)
4. Yin, D.S., Wang, G.Y., Wu, Y.: Data-Driven Decision Tree Learning Algorithm Based on Rough Set Theory. In: Tarumi, H., Li, Y., Yoshida, T. (eds.) *Proc. of the 2005 International Conference on Active Media Technology (AMT2005)*, Takamatsu, Kagawa, Japan, pp. 579–584 (2005)
5. Erdem, C., Zimmermann, H.G.: A Data-Driven Method for Input Feature Selection within Neural Prosody Generation. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 477–480 (2002)
6. Han, J., Cai, Y., Cercone, N.: Data-Driven Discovery of Quantitative Rules in Relational Databases. *IEEE Transactions on Knowledge and Data Engineering* 5(1), 29–40 (1993)
7. Kusiak, A., Shah, S.: Data-Mining-Based System for Prediction of Water Chemistry Faults. *IEEE Transactions on Industrial Electronics* 53(2), 593–603 (2006)
8. Cao, L., Lin, L., Zhang, C.: Domain-driven In-depth Pattern Discovery: A practical methodology, [Research Report], Faculty of Information Technology, University of Technology, Sydney, Australia (June 2005)
9. Zhang, C., Cao, L.: Domain-Driven Data Mining: Methodologies and Applications. In: Li, Y.F., Looi, M., Zhong, N. (eds.) *Advances in Intelligent IT - Active Media Technology 2006*, pp. 13–16. IOS Press, Amsterdam (2006)
10. Kuntz, P., Guillet, F., Lehn, R., Briand, H., User-Driven, A.: Process for Mining Association Rules. In: Zighed, A.D.A., Komorowski, J., Żytkow, J.M. (eds.) *PKDD 2000. LNCS (LNAI)*, vol. 1910, pp. 483–489. Springer, Heidelberg (2000)
11. Han, J., Lakshmanan, L., Ng, R.: Constraint-Based, Multidimensional Data Mining. *IEEE Computer* 32(8), 46–50 (1999)
12. Patrick, J., Palko, D., Munro, R., Zappavigna, M.: User Driven Example-Based Training for Creating Lexical Knowledgebases. In: *Australasian Natural Language Processing Workshop*, Canberra, Australia, pp. 17–24 (2002)
13. Dorado, A., Pedrycz, W., Izquierdo, E.: User-Driven Fuzzy Clustering: On the Road to Semantic Classification. In: Ślkezak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) *RSFDGrC 2005. LNCS (LNAI)*, vol. 3641, pp. 421–430. Springer, Heidelberg (2005)
14. Zhao, Y., Yao, Y.Y.: Interactive Classification Using a Granule Network. In: *Proc. of the 4th IEEE Int. Conf. on Cognitive Informatics*, Irvine, USA, 2005, pp. 250–259 (2005)
15. Plale, B., Gannon, D., Huang, Y., Kandaswamy, G., Pallickara, S.L., Slominski, A.: Cooperating Services for Data-Driven Computational Experimentation. *Computing in Science & Engineering* 7(5), 34–43 (2005)
16. Verlinde, H., Cock, M.D., Boute, R.: Fuzzy Versus Quantitative Association Rules: A Fair Data-Driven Comparison. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics* 36(3), 679–684 (2006)
17. Frigui, H.: MembershipMap: Data Transformation Based on Granulation and Fuzzy Membership Aggregation. *IEEE Transactions on Fuzzy Systems* 14(6), 885–896 (2006)

An Ontology-Based Mining System for Competitive Intelligence in Neuroscience

Jiao Li, Minlie Huang, and Xiaoyan Zhu

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China
jiao-li04@mails.tsinghua.edu.cn,
{aihuang, zxy-dcs}@tsinghua.edu.cn

Abstract. Neuroscience becomes a strategic growing field for both academic institutions and industrial companies because of their profound impact on human health, clinical therapy, and basic research such as brain informatics and cognitive science. To track activities of the rapidly developing field and observe new trends, neuroscientists must keep up-to-date with all the relevant information on the internet. In this paper, we present an ontology-based mining system, which is able to probe Competitive Intelligence in Neuroscience by using a well-defined ontology. It is able to support decision making by searching neuroscientific discoveries semantically and tracking new trends statistically. The experiments performed on 15,433,668 MEDLINE articles yield evidence for the feasibility and validity of our system.

1 Introduction

In the past decades, experimental advances in neuroscience, from image of specific issue by Magnetic Resonance Imaging (MRI) to localization of specific macromolecules by microscopies, have given individual neuroscientist an increasingly powerful arsenal for obtaining data across multiple scales, from the level of molecules to entire nervous systems [1]. Unlike other academic fields, the objects studied in neuroscience are considerably large, complex and rapidly growing, including cellular architectures, connectivity, physiological and behavioral data. The challenges in the above aspects inspire experts from multi-disciplines to engage in related research. To record new discoveries and advances in neuroscience, peer-reviewed literature has exploded to an unpredicted vast volume. It's extremely difficult for any individual to connect with all the related fields altogether and discover scientific trends manually through reading neuroscientific papers [2].

Thus, an automatic tool is needed to track activities in a wide range of the field and transform information into exploitable intelligence, which can be used by decision makers, named as *competitive intelligence* [3]. As a sub-area of knowledge management, competitive intelligence is aiming to monitor the external environment of a firm to obtain information relevant to its decision-making process [4].

It has been applied in many fields (e.g., military [5] and business [6] [7]) and been successful in practice, however, competitive intelligence applied in academic field is still in its infancy [8].

In this paper, we proposed an Ontology-based Mining System for Competitive Intelligence in Neuroscience (OMSCIN). It can automatically crawl the up-to-date documents about neuroscience from the public websites, classify them to different categories, annotate them within the controlled vocabularies, index them in the proper structure, and map them into the designed ontology. The most significant contribution of OMSCIN is that it integrates the formalism of ontology into competitive intelligence system, which supports insightful comprehension for the information by defining a ‘specification of conceptualization’ in neuroscience.

The rest of the paper is organized as follows: in Section 2, we review related work in competitive intelligence, related tools, and ontology. The system architecture is presented in Section 3. Section 4 describes ontology modeling for OMSCIN and interprets its roles in the system. Section 5 exemplarily shows experiments on more than 15 million MEDLINE records. After concluding our work by far, we discuss the open problems and future work in Section 6.

2 Related Work

Traditionally, competitive intelligence relied upon published company reports and other kinds of printed information. Instead of the printed reports, information acquired from the internet rapidly becomes one of dominant sources for competitive intelligence professionals [9]. With the rapid development of applicable disciplines and the tight collaboration between industry and academia, competitive intelligence should have a much broader definition than ever before. Additionally, for domain-specific studies (e.g. neuroscience) in information science, ontology becomes more mature and more institutionalized than the classic one in philosophy [10]. Thus, it’s necessary to review competitive intelligence studies, Web-based competitive intelligence tools as well as ontology modeling in neuroscience.

2.1 Competitive Intelligence

The Society of Competitive Intelligence Professionals (SCIP) defines competitive intelligence as ‘the process of ethically collecting, analyzing and disseminating accurate, relevant, specific, timely, foresighted and actionable intelligence regarding the implications of the business environment, competitors and the organization itself’ [11]. It is often mistakenly thought of in a narrow way, as a means of gathering ‘secret’ information like an espionage to gain advantage over competitors restrained in business or market. We adopt a much broader view of competitive intelligence, defining it as the use of public sources to explore information about the competition in the concerned environment [12]. It is not only

the collection of information from variety sources, but also analysis and synthesis of such information, which could help users to make decisions by providing comprehensive and intuitive evidences.

Competitive Intelligence is an ongoing process comprised with four stages: (1) direction: determining strategic information needed; (2) collection: collecting of environmental data; (3) analysis: interpreting data and producing intelligence; (4) dissemination: distributing intelligence, known as intelligence cycle [13] [14]. Well-done competitive intelligence can bring benefits at the aspects of risk estimate, investment decision and organization survival etc. [7].

Different from competitive intelligence in business environment, our OMSCIN allows experts to know the development of a specific domain, helps entrants to identify the famous scientists and their work, assists group managers to evaluate the performances of research employees, provides company managers the new trends, and reports the distribution of supported funds for foundation manager.

2.2 Web-Based Competitive Intelligence Tools

With the tremendous growth of the internet, most commercial and academic information is easily accessible publicly. It enables organizations to monitor and search the Web sites of their competitors, alliances, and possible collaborators. Web-based information sources are becoming increasingly important in competitive intelligence processes. Current researches mainly encompass the following two issues: collecting relevant information and analyzing Web link.

Collecting relevant information

A survey of over 300 competitive intelligence professionals shows that data collection is the most time-consuming task in typical competitive intelligence projects, accounting for more than 30% of the total time spent [13]. To address the information needs of competitive intelligence professionals in strategic decision making, a large number of commercial systems have been developed. Convera RetrievalWare and Internet Spider collect, monitor and index information from textual files as well as graphic files. However, they do not automatically categorize documents into different groups [15]. Automated tools support a wide range of information collection which includes automatic searching and monitoring information sources in the internet and corporate intranet, and categorizing documents into categories predefined by users or domain experts [16]. Competitive intelligence Spider searches for relevant Web pages based on keywords and other criteria specified by users. The returned documents are indexed and clustered into different groups, where categories are not predefined [17].

Analyzing Web link

The hyperlinks among Web pages represent an underlying social structure of linked communities which are implicit and 'natural' in that they are self-organizing [18]. They provide new opportunities for analyzing hidden Web communities. Fuld & Company, a competitive intelligence consultancy company, identified competitors' relationship by hypertext link analysis [19]. Borgman &

Furner (2002) used link analysis to measure the level of importance/influence of persons, organizations or nations [20]. Reid (2004) proposed a framework for detecting and analyzing relationships and activities of a company's implicit Web communities [21].

Our proposed OMSCIN system architecture depicted in Section 3 is also Web-based. However, we employ ontology to organize the information collected from the internet. Moreover, semantic relationships between objects represented by ontology are considered instead of the hyperlinks. Ontology will be briefly introduced in Section 2.3 and specifically interpreted in Section 4.

2.3 Ontology

Unlike data models, which are focused on structural association of attributes, ontology conveys much more abundant semantics (meaning) of data. Ontology is used to model domain concepts, their relationships, and the association with data. An widely accepted definition for the ontology is 'the specification of a conceptualization' of a domain [22].

Bio-ontology, taken up by biologists as a means to consistently annotate features from genotype to phenotype, has changed markedly since 2000. It has moved from a niche activity to one that is, in all aspects, a mainstream activity [10] [23]. The widely used bio-ontologies include: GO developed by Gene Ontology Consortium [24] [25]; OBO developed by Open Biomedical Ontology Consortium [26]; as well as MeSH (Medical Subject Headings) and UMLS (Unified Medical Language System) developed by National Library of Medicine [27] [28] etc.

Relevant but not equal to bio-ontology, ontology in neuroscience is also under developing. Based on the extension of Yale Sense Lab [29] [30] ontology, the Neuroscience Semantic Web [31] uses a common RDF to mediate queries across nine neuroscience related databases, and provides semantic retrieval and navigation in an integrated fashion. Moreover, the W3C (World Wide Web Consortium) created the HCLSIG (Health Care and Life Sciences Interest Group) to develop and support the use of Semantic Web technologies to improve collaboration, research and development, and innovation adoption in the health care and life science domains [32].

Referring to MeSH and ontology in neuroscience, we design the ontology model of the OMSCIN. The detailed description about our ontology as well as its roles in the whole system will be discussed in Section 4.

3 System Architecture

From the perspective of system architecture, the proposed OMSCIN is a classic client-server system, with its semantic search and statistics view at the client side, and its data process at the server side (see Fig. 1). In this section, the components of OMSCIN prototype system are presented in detail.

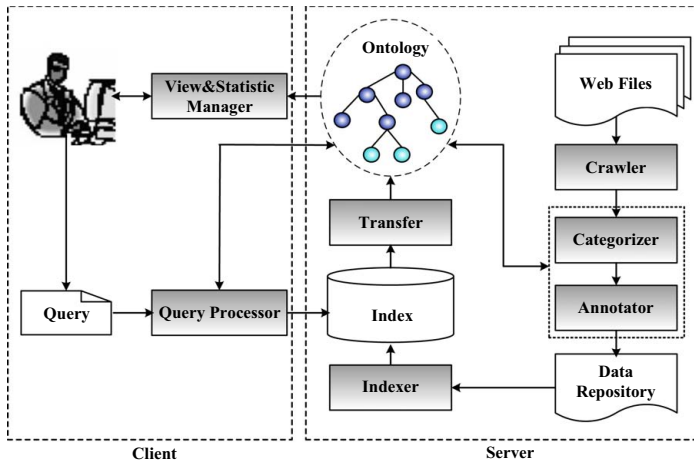


Fig. 1. The System Architecture of OMSCIN

3.1 Crawler

The Web-based workflow usually begin with a crawler, which downloads documents from websites specified in the configuration designed by the administrator [33]. Moreover, the crawled documents within unrelated contents need to be filtered out. Subsequently, documents in different formats need to be converted into XML (eXtensible Markup Language format).

3.2 Categorizer and Annotator

The categorizer and annotator attempt to mimic two activities carried by human annotators to assign concepts of an ontology (e.g., Medical Subject Headings) to the XML formatted documents as major topics, where, the categorizer is used for a classification task, and the annotator assigns the proper subcategory concepts to a document.

Take a document whose title is ‘*Elevated plasma noradrenaline concentration in untreated patients with early-stage Parkinson disease*’ [34] for instance. Ideally, it would be assigned with *Parkinson Disease* (a neurodegenerative disease), which is a sub-concept of *Nervous System Diseases* in the ontology automatically. From previous studies, it’s feasible to apply Support Vector Machines (SVM) [35] and extended bi-gram features [36] for solving the key technical issues in this step.

3.3 Indexer

Ontology offers powerful data representation and semantic search, however, these advantages are limited with ontology scale increasing [37]. For a large amount of annotated documents, such as data repository including millions of records,

ontology cannot afford all of them. Referring to modern information retrieval techniques, data repository can be pushed into index within its structured information remained [38]. Relevant information is retrieved from the index, before being imported into the designed ontology as the instances. Thus, the indexer component is used to release ontology's scale pressure to a certain extent.

3.4 Query Processor and Transfer

The initial retrieval, based on Information Retrieval (IR) techniques, focuses on providing as many relevant documents as possible for the ontology instances. Meanwhile, ontology which organizes domain concepts taxonomically can improve the correct understanding of query terms through query expansion. The query processor component takes advantage of taxonomic concepts of the ontology to expand queries with their synonyms as well as the sub-concepts in a certain range. Subsequently, the original query term and the expanded ones are expressed in the query language, and fed into the IR system. Take another neurodegenerative disease, *Alzheimer Disease*, for an original query example. Its synonym attribute value, *Senile Dementia*, is expanded according to the OMSCIN ontology.

The transfer component does not work until the relevant documents are retrieved by the IR module. It extracts objects in the documents and imports them into the ontology model as instances, which enriches the knowledge network.

About the roles of ontology in the OMSCIN system, it will be further discussed in Section 4.

3.5 View and Statistic Manager

The up-to-date information related with neuroscience in the internet is processed by the crawler, categorizer, annotator, indexer and transfer, and finally, is stored in the ontology as instances. The above series of document processing at the server side are the groundwork of the OMSCIN system, making the precise, timely and reliable service available at the client side. The view & statistic manager is designed at the client side to generate insightful and significant data layouts for further competitive analysis and decision making.

4 Ontology Modeling

In information science, ontology is a description framework to formulate an exhaustive and rigorous conceptual schema about a domain. This domain does not have to be the complete knowledge of that topic, but purely a domain of interest decided upon by the creator of the ontology [22]. The OMSCIN focuses on a neuroscience domain, in which conceptual specification is crucial for the semantics understanding and knowledge network constructing. The two roles of ontology, adding semantics and bracing knowledge network, are considered in the OMSCIN ontology modeling (see Fig. 2).

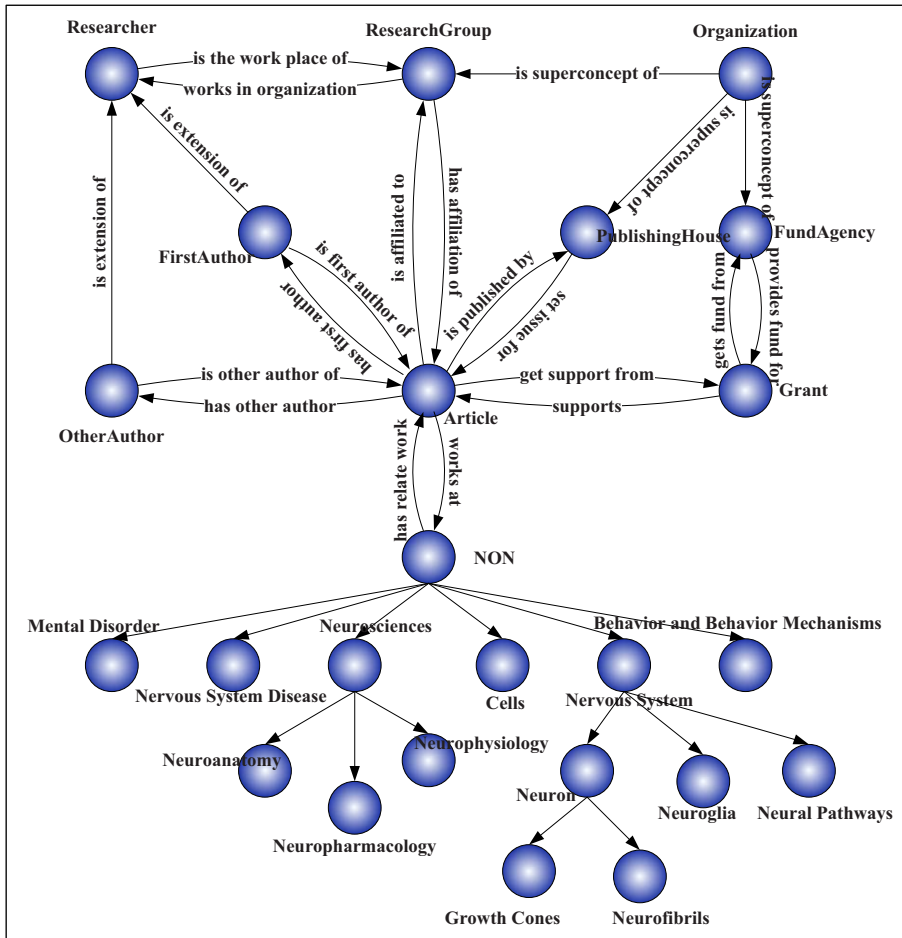


Fig. 2. The Concept-level Ontology in OMSCIN

4.1 Adding Semantics

In linguistics, semantics is defined as the meaning of (part of) words, phrases, sentences, and texts, which is opposed to the formal expression. Take the following question for example, “Who engaged in ‘Brain Disease’ in 2000?”. It is difficult to answer this question if there is no correct understanding of ‘*Brain Disease*’ although ‘*Brain*’ and ‘*Disease*’ are two simple words in English. This is due to the gap between semantics and expressions, therefore, ontology is employed in OMSCIN to bridge the gap.

Ontology uses standard components (e.g., concept, relation, instance, and extension etc.) to describe the meaning of data. We take advantage of the Yale Sense Lab’s Neuro Semantic Web Ontology and National Library of Medicine’s Medical Subject Headings to design a Neuroscience Ontology (NON for short).

As shown in the bottom part of Fig. 2, all concepts in NON are organized in a hierarchical structure. It totally consists of 1983 concepts falling into 6 categories: *Mental Disorder*, *Nervous System Disease*, *Neurosciences*, *Cells*, *Nervous System*, as well as *Behavior* and *Behavior Mechanisms*. Each category is further divided into subcategories. Within each subcategory, concepts are organized hierarchically from most general (e.g., ‘*Nervous System*’) to most specific (e.g., ‘*Caudate Nucleus*’) levels, amounting to 11 levels. The relation between the above concepts is subsumption. The attributes assigned to each concept include *ScopeNote*, *Synonym*, *TreeNumber*, *DateCreated* and so on, used for identifying the concept as the unique one in the whole knowledge network.

Also take ‘*Brain Disease*’ for example to explain how NON adds semantics to it. An article **A** whose title is ‘*Senile dementia associated with amyloid beta protein angiopathy and tau perivascular pathology but not neuritic plaques in patients homozygous for the APOE-epsilon4 allele*’ was published in the journal of ‘*Acta neuropathologica*’ in 2000 [39]. Readers can easily identify **A** as the ‘*Brain Disease*’ related one, depending on their knowledge that ‘*Senile Dementia*’ is a kind of ‘*Brain Disease*’, although they failed to find the phrase ‘*Brain Disease*’ in any section of **A**. In our NON, one *Synonym* attribute of the concept ‘*Alzheimer Disease*’ is ‘*Senile Dementia*’, moreover, ‘*Alzheimer Disease*’ is a sub-concept of ‘*Dementia*’ which is a sub-concept of ‘*Brain Disease*’. Because of the transitivity of the NON’s subsumption relationship, we can infer that ‘*Senile Dementia*’ or ‘*Alzheimer Disease*’ is a subconcept of ‘*Brain Disease*’. Both terms appear in **A** frequently. By this ways, **A** could be judged as the relevant article of ‘*Brain Disease*’ automatically.

The above example illuminates how the hierarchical ontology NON adds semantics to domain-specific terms through standard components description and logic inference.

4.2 Bracing Knowledge Network

In our OMSCIN, ontology not only enhances the correct understanding for domain specific terms, but also braces the global knowledge network as a backbone of the system.

Compared with the typical one in a hierarchical structure, the ontology modeled in this part uses more complex relations to link the concepts (see the upper part of Fig. 2). *Researcher* is an extensible concept with two extensions, *FirstAuthor* and *OtherAuthor*, who may play different roles in *Article*. Three *Organizations* have different relations with *Article* respectively, i.e., *Research-Group* has the affiliation of *Article*; *FundAgency* provides supporting *Grant*; *PublishingHouse* publishes *Article*. NON concepts assigned to *Article* represent the major topics of *Article* in neuroscience. In order to support more flexible semantic search, reverse relations are adopted in the knowledge network. For example, *Grant* supports *Article*, reversely, *Article* gets support from *Grant*. Furthermore, a relation and its reverse one may have distinct attributes, e.g., a *Researcher* can be the *FirstAuthor* of multiple *Articles*, but an *Article* can merely have one *FirstAuthor*.

Besides the various relations between concepts, the corresponding attributes are designed elaborately including attribute names, types, whether indexed for further retrieval, whether multiple appearance and so on. For example, a query constructed for *Article* can involve the published date (*PubDate*) before, after, or between some specific ones, because the attribute type of *Article PubDate* is *Date (dd.mm.yyyy)* which supports date processing in the query. All these ingenuity designs make the network much better functionality.

Based on the concept-level ontology mentioned above, the respective instances can be imported into the knowledge network automatically, and inherit the relations between their instance concepts. For example, article **A**, ‘*Senile dementia associated with amyloid beta protein angiopathy and tau perivascular pathology but not neuritic plaques in patients homozygous for the APOE-epsilon4 allele*’ [39], is an instance of concept *Article*, and its *PubDate* attribute value is (09.11.2000). *Vidal* is an instance of *FirstAuthor*. ‘*Department of Pathology, New York University School of Medicine, NY 10016, USA*’ is an instance of *Research-Group*. ‘*National Institute on Aging*’ is an instance of *FundAgency*. ‘*AG08721*’ is an instance of *Grant*. These instances inherit the concept-level relations shown in Fig. 2, i.e., *Vidal* is the first author of **A** which is affiliated to ‘*Department of Pathology, New York University School of Medicine, NY 10016, USA*’, and ‘*National Institute on Aging*’ provides funds for grant ‘*AG08721*’ which supports **A** working at ‘*Alzheimer Disease*’.

To demonstrate the feasibility of our ontology model and its contribution to mining for competitive intelligence, we apply it to MEDLINE resource in life science.

5 Competitive Intelligence in Neuroscience

Our ontology-based mining system facilitates analysis for competitive intelligence in neuroscience through semantic search and statistics view. These two fundamental functions could provide evidence for competitive intelligence professionals to make decisions.

In our experiments, a literature resource named MEDLINE, which is powered by National Library of Medicine, is used as our data repository. Released in 2005, MEDLINE database contains 15,433,668 documents, tracing back to 1950s in life science (completed data to Dec. 31st 2004). Literature from approximately 4,800 worldwide leading journals covers biomedicine and health, broadly defined to encompass those areas of the life science, behavioral science, chemical science, and bioengineering by professionals engaged in basic research and clinical care, public health, health policy development, or related educational activities [40].

The following two subsections describe the semantic search and statistics view functions that benefit from our ontology modeling, in the scenarios that OM-SCIN responds to the questions: “Who engaged in ‘Alzheimer Disease’ research in 2000?” and “How about ‘Magnetic Resonance Imaging (MRI)’ and Electroencephalography (EEG) development from 1960?”, and some competition-related questions.

5.1 Semantic Search

The user query, for instance, “Who engaged in ‘Alzheimer Disease’ research in 2000?”, is translated into an ontology-guided query (see Fig. 2) as follows: (1) *Researcher* has an extension of *FirstAuthor*; (2) *FirstAuthor* is the first author of *Article*; the *PubDate* attribute value of *Article* is between 01.01.2000 and 31.12.2000; (3) *Article* works at *NON*; (4) the *Name* attribute value of *NON* equals to *Alzheimer Disease*. The query translation for the semantic search is the compilation of concepts, instances, relations and attributes to nested and static search queries, whose results vary only when the knowledge network is changed.

5.2 Statistics View

In the second scenario of “How about ‘Magnetic Resonance Imaging (MRI)’ and Electroencephalography (EEG) development from 1960?”, OMSCIN responds to it through tracking all the concepts that connect with MRI annually. It includes annual research group number engaged in MRI, annual article number on MRI, and annual grant number provided for MRI. These cases also hold for EEG. Furthermore, OMSCIN provides the intuitive statistics view for the tracking results.

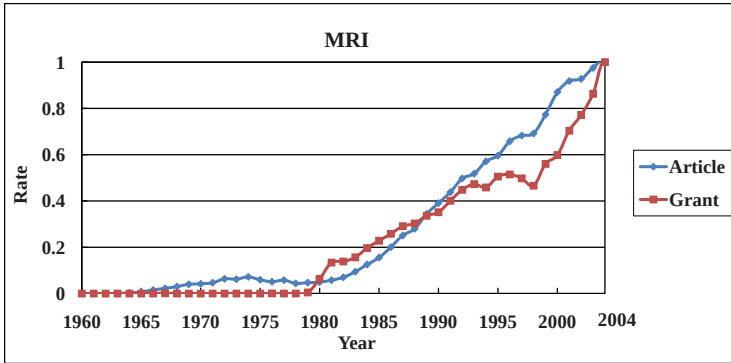


Fig. 3. Tracking MRI development from 1960 to 2004

Figure 3 shows the normalized annual article number on MRI and annual grant number provided for MRI. Generally speaking, MRI research has become increasingly hot since 1990s. In 2004, there were 16,072 articles and 3,195 grants on MRI in MEDLINE totally. The more grants support, the more achievements win. In 1980s, the curve of the grant rate crossed with that of the article rate. It states the interactive and hysteretic nature of this phenomenon, i.e., the research achievements inspire the grant distributions, meanwhile, the grants encourage the research, but the research achievements do not instantly follow the supports applied to them.

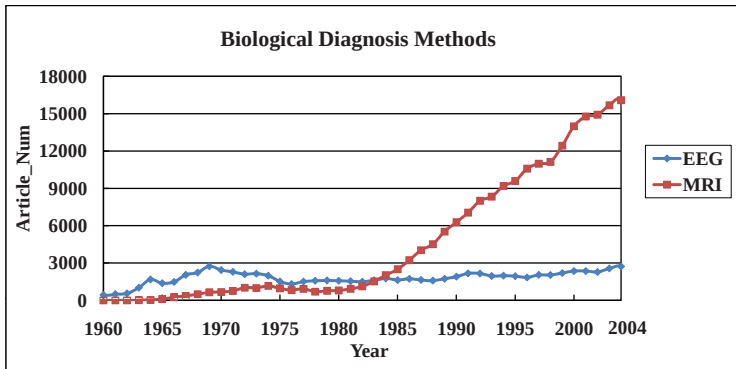


Fig. 4. Tracking MRI and EEG development from 1960 to 2004

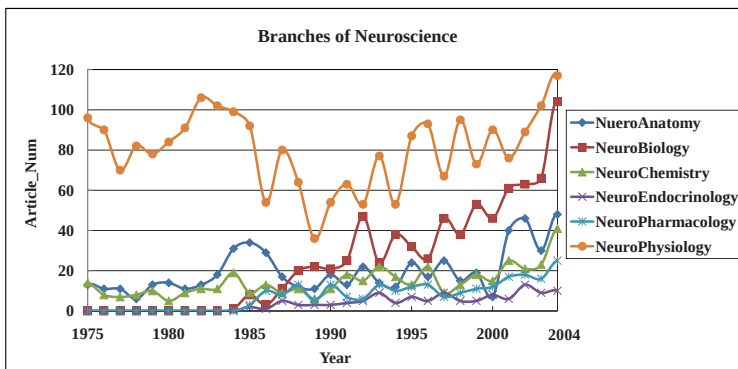


Fig. 5. Tracking research branches of neuroscience from 1975 to 2004

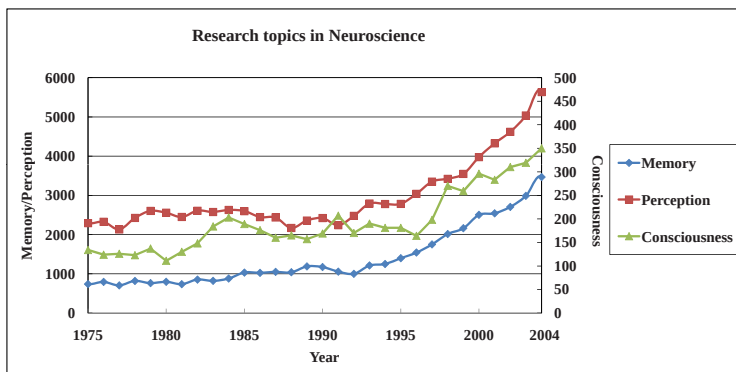


Fig. 6. Tracking research topics in neuroscience from 1975 to 2004

Figure 4 shows the number of articles has been published annually on MRI and EEG since 1960. We can see that MRI and EEG have different trends, although both of them are methods of neurological diagnosis. The two curves are consistent with the history of technique advances. The electroencephalograph was invented in the 1920s by Hans Berger, and applied to clinical fields in 1935 [41]. The development of MRI began with the discovery of nuclear magnetic resonance (NMR) which was used to investigate biological tissues in the mid of 1960s. As MRI offers increased-contrast resolution, enabling better visualizations of soft tissues, and provides highly detailed information without exposing the body to radiation, it's primarily used to examine the internal organs for abnormalities [42]. The MRI related research (e.g., image analysis for brain diseases) becomes a hot topic in neuroscience. Many new discoveries in neuroscience are found as the MRI equipments become more and more sophisticated, and as a result, the number of papers published is increasing growing. However, EEG plays few roles in most of these discoveries, and consequently, much less articles is published.

Figure 5 illustrates the number of published articles within six branches of neuroscience. These curves share similar trends, implying that these branches are interweaved, and that advances from one branch propel and influence the development of others. The most active branches are neurobiology and neurophysiology these years. Figure 6 shows the number of annually published articles related to three research topics in neuroscience: *memory*, *perception*, and *consciousness*. It can be observed: perception is the most widely studied among the three topics, and there has been more than 3000 papers published on the topic annually since 1995; all the three topics have gained increasingly attention since 1995, because of advances of experimental techniques (e.g. fMRI) and other related disciplines; there are few paper (less than 300 papers in average per year) focusing on consciousness, which is the most difficult topic in neuroscience. These facts lead to a conclusion that the topic of consciousness will be the hotspot of neuroscience research for the next ten, or maybe five years.

6 Conclusions and Future Work

An ontology-based mining system for Competitive Intelligence in neuroscience (OMSCIN) is presented in this paper. Its most significant feature is that a well defined ontology becomes the backbone of the system, which makes the deep semantics behind data be easily understood. The proposed system transfers and maps Web-based information into concepts defined by our ontology, and provides semantic search and statistics view to help competitive intelligence professionals to make decisions. Experiments carried on MEDLINE database, have shown promising results, which affirms us that such an ontology-based mining system will be greatly helpful for competitive intelligence analysis.

The research is still ongoing, and our future work mainly includes: (1) In-depth evaluation approaches: more experiments are needed to validate the reliability of the system; (2) Deep mining buried knowledge in the ontology: for example, the information that two experts have similar research profiles in neuroscience

but never cooperate with each other, will be quite important and potentially useful for competitive intelligence. There are many other challenges including Web-based text incompatibility among complex data in neuroscience, interoperability caused by heterogeneous and distributed data sources, as well as ontology annotation etc. All of these inspire us to make further advances in our OMSCIN.

Acknowledgments. The work was supported by Natural Science Foundation of China under grant No. 60572084, 863 Program under No. 2006AA02Z321, and Tsinghua Basic Research Foundation under grant No. 052220205 and No.053220002. We also would like to thank Dr. Thomas Kamps in Fraunhofer IPSI for his advice on our work.

References

1. Martone, M.E., Gupta, A., Ellisman, M.H.: e-Neuroscience: Challenges and Triumphs in Integrating Distributed Data from Molecules to Brains. *Nature Neuroscience* 7(5), 467–472 (2004)
2. Jensen, L.J., Saric, J., Bork, P.: Literature Mining for the Biologist: From Information Retrieval to Biological Discovery. *Nature Reviews Genetics* 7, 119–129 (2006)
3. Rouach, D., Santi, P.: Competitive Intelligence Adds Value: Five Intelligence Attitudes. *European Management Journal* 19(5), 552–559 (2001)
4. Gilad, B., Gilad, T.: Business Intelligence System: A New Tool for Competitive Advantage: AMACOM/American Management Association (1988)
5. Kahaner, L.: Competitive Intelligence: How to Gather, Analyze, and Use Information to Move Your Business to the Top: Touchstone (1998)
6. Drucker, P.: The Coming of the New Organization. *Harvard Business Review on Knowledge Management* (1998)
7. Gates, B.: *Business at the Speed of Thought*. Penguin Books, London (2002)
8. Carlucci, S., Page, A., Finegold, D.: The Role of Competitive Intelligence in Biotech Startups. *Nature Biotechnology*, 525–527 (2005)
9. Group, Futures, Ostriches & Eagles 1997. The Futures Group Articles (1998)
10. Bodenreider, O., Stevens, R.: Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics* 7, 256–274 (2006)
11. SCIP: Society of Competitive Intelligence Professionals (access 2006), <http://www.scip.org/>
12. McGonagle, J.J., Vella, C.M.: *Outsmarting the Competition: Practical Approaches to Finding and Using Competitive Information*, Sourcebooks Inc. (1990)
13. Prescott, J.E., Fleisher, C.S.: SCIP: Who We Are, What We Do. *Competitive Intelligence Review* 2(11), 22–26 (1991)
14. Vriens, D.: The Role of Information and Communication Technology in Competitive Intelligence. In: Vriens, D. (ed.) *Information and Communication Technology for Competitive Intelligence*, pp. 1–33. Idea Group Pub. Hershey, PA (2004)
15. Convera (access 2006), <http://www.excalib.com/>
16. Autonomy (access 2006), <http://autonomy.com/content/home/>
17. Chen, H., Chau, M., Zeng, D.: CI Spider: A Tool for Competitive Intelligence on the Web. *Decision Support Systems* 34(1), 1–17 (2002)
18. Flake, G.W., et al.: Self-Organization and Identification of Web Communities. *IEEE Computer* 35(3), 66–71 (2002)

19. Fuld & Company (access 2006), <http://www.fuld.com/>
20. Borgman, C.L., Furner, J.: Scholarly Communication and Bibliometrics. In: Borgman, C.L. (ed.) *Annual Review of Information Science and Technology*, Information Today, Inc. (2002)
21. Reid, E.: Chapter III: Using Web Link Analysis to Detect and Analyze Hidden Web Communities. In: Vriens, D. (ed.) *Information and Communication Technology for Competitive*, pp. 57–84. Idea Group Pub, Hershey, PA (2004)
22. Gruber, T.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies* 43, 907–928 (1993)
23. Stevens, R., Goble, C.A., Bechhofer, S.: Ontology-based Knowledge Representation for Bioinformatics Briefings in Bioinformatics 1(4) 398–414 (2000)
24. Ashburner, M., et al.: Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* 25(1), 25–29 (2000)
25. Gene Ontology (access 2006), <http://www.geneontology.org/>
26. Open Biomedical Ontologies (access 2006), <http://obo.sourceforge.net/>
27. MeSH: Medical Subject Headings (access 2006), <http://www.nlm.nih.gov/mesh/>
28. UMLS: Unified Medical Language System (access 2006), <http://umlsinfo.nlm.nih.gov/>
29. Miller, P.L., et al.: Integration of Multidisciplinary Sensory Data: A Pilot Model of the Human Brain Project Approach. *Journal of the American Medical Informatics Association* 8, 34–48 (2001)
30. Yale SenseLab (access 2006), <http://senselab.med.yale.edu/senselab/>
31. Neuroscience Semantic Web (access 2006), <http://neuroweb.med.yale.edu/neurosw/>
32. HCLSIG: W3C Health Care and Life Sciences Interest Group (access 2006), <http://www.w3.org/2001/sw/hcls/charter>
33. Su, Z., et al.: Market Intelligence Portal: An Entity-based System for Managing Market Intelligence. *IBM Systems Journal* 43(3), 534–554 (2004)
34. Naoko, S., et al.: Elevated Plasma Noradrenaline Concentration in Untreated Patients with Early-stage Parkinson Disease. *Autonomic Neuroscience* 130(1-2), 65–65 (2006)
35. Cristianini, N., Taylor, J.S.: *An Introduction to Support Vector Machines: and Other Kernel-based Learning*. Cambridge University Press, New York (2000)
36. Zhang, X., Zhu, X.: Extended Bi-gram Features in Text Categorization. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) *IBPRIA 2005*. LNCS, vol. 3523, Springer, Heidelberg (2005)
37. Zhang, Y., Vasconcelos, W., Sleeman, D.: OntoSearch: An Ontology Search Engine. In: Webb, G.I., Yu, X. (eds.) *AI 2004*. LNCS (LNAI), vol. 3339, Springer, Heidelberg (2004)
38. Metzler, D., Croft, W.B.: Combining the Language Model and Inference Network Approaches to Retrieval. *Information Processing and Management. Special Issue on Bayesian Networks and Information Retrieval* 40(5), 735–750 (2004)
39. Vidal, R., et al.: Senile Dementia Associated with Amyloid Beta Protein Angiopathy and tau Perivascular Pathology but not Neuritic Plaques in Patients Homozygous for the APOEepsilon4 Allele. *Acta Neuropathologica* 100(1), 1–12 (2000)
40. Fact Sheet of Medline (access 2006), <http://www.nlm.nih.gov/pubs/factsheets/medline.html>
41. Ebersole, J.S.: *Current Practice of Clinical Electroencephalography*: Lippincott Williams & Wilkins (2002)
42. Jacques, B., Vlaardingerbroek, M.: *Magnetic Resonance Imaging Theory and Practice*. Springer, Heidelberg (1996)

Supervised Web Document Classification Using Discrete Transforms, Active Hypercontours and Expert Knowledge

P.S. Szczepaniak^{1,2}, A. Tomczyk¹, and M. Pryczek¹

¹ Institute of Computer Science, Technical University of Lodz
Wolczanska 215, 93-005 Lodz, Poland

tomczyk@ics.p.lodz.pl, michalp@stud.ics.p.lodz.pl

² Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland

Abstract. In this paper, a new method of *supervised classification* of documents is proposed. It utilizes *discrete transforms* to extract *features* from *classified* objects and adopts *adaptive potential active hypercontours* (APAH) for document *classification*. The idea of APAH generalizes classic contour methods of image segmentation. It has two main advantages: it can use almost any knowledge during the search for an optimal *classification* function and it can operate in a *feature* space where only metric is defined. Here, both of them are utilized - the first one by using expert knowledge about significance of *documents* from training set and the second one by inducing new metrics in *feature* spaces. The method has been evaluated on the subset of *open directory project* (ODP) database and compared with *k*-NN, the well known *classification* technique.

1 Introduction

The rapid development of *Web Intelligence* (WI) [1,2,3,4,5,6,7,8,9] technologies leads to the growth of the amount of reliable knowledge that can be used for the efficiency improvement of many standard tasks in artificial intelligence, which in turn WI can benefit from. This imposes the necessity to either create new methods that are able to effectively adapt knowledge coming from different sources or modify the existing techniques in order to satisfy *Web Intelligence* requirements. The presented approach joins experiences gained from the domains that have been considered separately so far, giving mechanisms capable of utilizing external knowledge in an efficient and flexible way.

The paper is organized as follows: in section 2 the problem of *classification* of documents is stated, in section 3 integral spatial transformations using kernel methods for *feature* extraction are described and in section 4 the *adaptive potential active hypercontour* algorithm used for construction of an optimal *classifier* is presented. The next two sections focus on the presentation of data used in the experiments and the discussion of obtained results respectively. The paper concludes with the summary of the proposed method.

2 Supervised and Unsupervised Document Classification

2.1 Classification

The *classification* problem can be formulated as the task of assigning a proper label l from the finite set of labels $\mathcal{L}$ (where e.g. $\mathcal{L} = \{1, \dots, L\}$ and L is a number of classes) to each object o from the given set of objects $\mathcal{O}$. Such an assignment can formally be described as a *classification* function (*classifier*) $k : \mathcal{O} \rightarrow \mathcal{L}$ (each object $o \in \mathcal{O}$ receives a unique label $l \in \mathcal{L}$). Because there are many functions $k \in \mathcal{K}$ that map $\mathcal{O}$ into $\mathcal{L}$ (where $\mathcal{K}$ denotes a set of all possible *classifiers* in a given problem) the problem of construction of an optimal *classifier* emerges. Optimality can be reflected in the form of *performance index* $Q : \mathcal{K} \rightarrow \mathbb{R}$, which is capable of evaluation of the usefulness of each function k . Thus, the problem of construction of optimal *classifier* is optimization of *performance index* Q [23,15,11,16].

Feature Space. The objects from the real world cannot be used by *classification* algorithms directly. They can operate only on the data extracted from the objects (i.e. sensed, measured, etc.) called *features* of the objects. The extraction process can be described as a function $\phi : \mathcal{O} \rightarrow \mathcal{X}$. In practice, a sequence of M subsequent transformations $\phi^0, \dots, \phi^{M-1}$ such that $\mathcal{X}_0 = \mathcal{O}$, $\mathcal{X}_M \subseteq \mathcal{X}$ and $\phi^{i-1} : \mathcal{X}_{i-1} \rightarrow \mathcal{X}_i$ for all $i = 1, \dots, M$ is applied. It is important to mention that the structure of the resulting sets $\mathcal{X}_i$ as well as definitions of ϕ^i may not be known as long as we can satisfy specific algorithm restrictions. Usually this means that the final space is an inner product or a metric one. In order to simplify the notation, the objects and their *feature* descriptors will be used interchangeably when the *feature* extraction method ϕ is either precisely known or insignificant.

Unsupervised Classification. When *a priori* knowledge used for construction of the *classifier* does not contain any information about the kind of object labels, one can speak about *unsupervised classification* (*clustering*). Here, the construction of the *classifier* is based only on inner similarities or dissimilarities of the objects. Consequently, the goal of the *clustering* algorithm is to divide the given set of objects $\mathcal{O}$ into groups such that homogeneity inside the groups and the heterogeneity between them are fulfilled (the number of groups L can be unknown). In this case, labels are used only to discriminate between different groups. Obviously, the division depends on the type of object *features* and the given similarity measures. There exist many approaches to *clustering*, such as: *single-link*, *complete-link* and *average-link clustering*, *Hard C-Means* (HCM), *Fuzzy C-Means* (FCM), *Possibilistic C-Means* (PCM) *clustering*, *Kohonen neural network*, knowledge-based *clustering* [20], etc.

Supervised Classification. Sometimes, the knowledge used for construction of *classification* algorithm contains information about the expected assignments of concrete labels to the objects and the number of classes L is then precisely known. Similarities and dissimilarities between objects can also be used here to

divide the given set of objects $\mathcal{O}$. This type of approach is called a *supervised classification* (or simply *classification* as opposed to *clustering*). There are many techniques used for *supervised classification*, for example: *k*-NN, *Multi-layer Perceptron* (MLP) ([11,16]), *probabilistic classification* ([15]), *Support Vector Machines* (SVM) [12], etc.

2.2 Document Classification

As stated before, *classification algorithms* usually do not depend on the type of the *classified* objects $o \in \mathcal{O}$, as long as the space $\mathcal{X}$ possesses the required properties. This is ensured by a proper, even not given explicitly, transformation ϕ . This *feature* is advantageous when *classification* of textual documents needs to be performed.

Unsupervised classification of documents can be used to enhance information retrieval e.g. in web browsers. It may also be used in automatic feature induction, data mining and data structure reckoning [27]. *Supervised classification* of documents can be used in diagnostic systems where a part of information about objects is in the form of textual notes, reports, etc. Information retrieval systems adopting search results for determination of the user profile may profit from this type of *classification*. Moreover, the enhancement of search parameters by category selection results in the reduction of document search space dimension. This paper focuses on *supervised classification* of documents. However, the methods and ideas presented here can be used in an *unsupervised* case.

3 Transformations

Transformation of objects $o \in \mathcal{O}$ into the *feature* space $\phi(\mathcal{O}) \subseteq \mathcal{X}$ is a crucial point, as further processing is performed only in the resulting $\mathcal{X}$ space. In this section, the space $\mathcal{D}$ of documents is considered.

3.1 Vector Space Model

One of the most flexible ways of defining $\mathcal{X}$ is the transformation of $\mathcal{D}$ into a vector space; the notion *vector space model* (VSM) is used. For this purpose, a set of *feature* premises (descriptions) is needed. In this particular application, i.e. document *classification*, a finite set (in general it can be infinite) of terms $\mathcal{T} = \{\tau_0, \tau_1, \dots, \tau_{T-1}\}$, i.e. words or word stems is used (if stemming algorithm was applied). Consequently, ϕ is considered as a vector valued function $\phi = (\phi_0, \dots, \phi_{T-1})$ where $\phi_i : \mathcal{D} \rightarrow \mathcal{F}_i$ for $i = 0, \dots, T-1$, T - number of terms. Of course, $\phi : \mathcal{D} \rightarrow \mathcal{F}_0 \times \dots \times \mathcal{F}_{T-1} \subseteq \mathcal{X}$. In the following sections, functions ϕ_i related to term τ_i will also be denoted as ϕ_{τ_i} and both of the symbols will be used interchangeably. Moreover, terms and phrases will be considered as equivalent, even though there is an important logical difference between them.

3.2 Document Transformation

The classic VSM uses term counts as *feature* values. In this case, $\phi_\tau : \mathcal{D} \rightarrow \mathbb{R}$ for every $\tau \in \mathcal{T}$. Real codomain has been chosen because various extensions including vector normalization and weighting of elements may be applied. However, this model neglects the relations between occurrences of terms and spatial information about word distribution throughout the document which has been proven to be of importance in document retrieval and *classification* tasks [17,18,22]. The method adopted here was first introduced in *Fourier domain scoring* (FDS) [17], a novel document retrieval method. As it is the heart of the presented considerations, its basic concepts are shortly presented below.

The presented document indexing process is a sequence of vector space transformations $\phi^i = (\phi_0^i, \dots, \phi_{T-1}^i)$ where $i = 0, 1, 2$. To avoid misunderstandings and to be consistent with symbols used in document *classification* literature, d is used to denote a document from $\mathcal{D}$ and t denotes the index of term τ in $\mathcal{T}$. Moreover, the notation presented in Table 1 is applied. All the functions and transformations presented there are parametrized. However, for reasons of clarity, this fact has no reflection in the notation.

Table 1. Notation

Description	Functional notation	Vector/matrix notation
Term frequency signal	$\phi_\tau^0(d)$	f_t^d
Weighted term signal	$\phi_\tau^1(d)$	w_t^d
Final term signal	$\phi_\tau^2(d)$	v_t^d
Object in the <i>feature</i> space	$\phi(d)$	V_d

Term Signals (ϕ_0). Instead of counting the number of occurrences of the term τ in the whole document another strategy is used. The document d is divided into B sections (bins) of equal length. Terms are then counted in bins and a vector $f_t^d = (f_t^d(0), f_t^d(1), \dots, f_t^d(B-1))$ is created. These vectors are considered discrete signals in spatial or time domain and they are subject of further processing.

Weighting (ϕ_1). The application of significance weighting is a common technique used to improve representation of a document and make it more meaningful. There exist numerous techniques for doing this. The most widely used and flexible is the *term document frequency* $\times$ *inverse document frequency* (TDF $\times$ IDF) [17]. As we are operating on vectors rather than numbers, the TDF $\times$ IDF is adopted to fit vector structure of *feature* value domain. The respective weighting methods called *term bin frequency* TBF

$$w_t^d(b) = 1 + \ln f_t^d(b) \quad (1)$$

and *proportional term frequency* PTF:

$$w_t^d(b) = f_t^d(b) \left(1 + \ln \sum_{b=0}^{B-1} f_t^d(b) \right) \left(\sum_{b=0}^{B-1} f_t^d(b) \right)^{-1} \quad (2)$$

were proposed by Park [17]. Here $f_t^d(b)$ and $w_t^d(b)$ are count of term t in the spatial bin b in document d and its weighted value, respectively. As long as discrete transform described in the next section is a linear transform, IDF weighting (of standard form) can be applied after transformation.

Discrete Integral Transform (ϕ_2). A variety of well known transforms may be used e.g. *Fourier*, *cosine II*, *Hartley*. In general, discrete integral transform is of the form:

$$v_t^d(k) = \sum_{b=0}^{B-1} w_t^d(b) K_B(k, b) \quad (3)$$

where function K_B determines its type:

- *Fourier* transform (DFT)

$$K_B(k, b) = e^{-\frac{j2\pi kb}{B}} \quad (4)$$

- *Cosine II* transform (DCT2)

$$K_B(k, b) = \cos \left(\frac{\pi k(2b+1)}{2B} \right) \quad (5)$$

- *Hartley* transform (DHT)

$$K_B(k, b) = \cos \left(\frac{2\pi kb}{B} \right) + \sin \left(\frac{2\pi kb}{B} \right) \quad (6)$$

The resulting vector of complex or real numbers, i.e. term (word) signal in frequency domain, is taken as representation of the relation between document d and term t . To standardize representation of all documents, vector of zeros is taken (when some term was not found in the indexed document). It is also convenient to define a final index of a document as matrix:

$$V^d = [v_{t,k}^d] = [v_t^d(k)] \quad (7)$$

in which rows correspond to terms ($t = 0, \dots, T-1$) and columns reflect spatial kernel components ($k = 0, \dots, B-1$).

3.3 Inner Product, Norm and Metric

As stated before, general purpose classification algorithms do not depend on space structure as long as specific operations are defined in it, e.g. inner product or metric. According to Hilbert induction scheme, it is easy to induce metric space from inner product space. In this paper, the space $\mathcal{X}$ is in fact a finite product space, $\mathcal{X} = \mathcal{F}_0 \times \dots \times \mathcal{F}_{T-1}$. Assuming that the addition and scalar

multiplication in this space are defined and in each space $\mathcal{F}_i$ a norm is defined, $\|\cdot\|_{\mathcal{F}_i} : \mathcal{F}_i \rightarrow \mathbb{R}$, another function $\|\cdot\| : \mathcal{X} \rightarrow \mathbb{R}$ can be defined in the following way:

$$\|x\| = \sum_{i=0}^{T-1} \|x_i\|_{\mathcal{F}_i} \quad (8)$$

which is a norm in $\mathcal{X}$. Moreover:

- In every inner product space $\|\cdot\| : x \rightarrow \sqrt{\langle x, x \rangle}$, where $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ denotes inner product. Other induction schemes for norm are also possible (e.g. *Minkowski distance (p-norm distance)* definition gives a family of possible formulas).
- There exist numerous formulas of inducing a metric $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ in a normed or inner product space, provided that a subtraction operator is defined, e.g.:

$$\rho(x, y) = \|x - y\| \quad (9)$$

or

$$\rho(x, y) = 1 - \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad (10)$$

This gives the flexibility of enabling a different kind of algorithms having only an inner product or a norm defined.

3.4 Standard Operations

Metrics. In the experiments presented here, the following standard metrics are used:

- *Euclidean* metric

$$\rho_1(d_1, d_2) = \sqrt{\sum_{t=0}^{T-1} \sum_{k=0}^{B-1} (v_t^{d_1}(k) - v_t^{d_2}(k))^2} \quad (11)$$

- *Manhattan* metric

$$\rho_2(d_1, d_2) = \sum_{t=0}^{T-1} \sum_{k=0}^{B-1} |v_t^{d_1}(k) - v_t^{d_2}(k)| \quad (12)$$

- *Chebyshev* metric

$$\rho_3(d_1, d_2) = \max_{t=0, \dots, T-1, k=0 \dots B-1} |v_t^{d_1}(k) - v_t^{d_2}(k)| \quad (13)$$

However, to adapt them directly to the *DFT* transform, one must operate on half-complex encoded vectors that represent transformation results. To overcome this inconvenience and use the phase information, other metrics based on the idea presented in section 3.3 can be applied.

Inner Products. In this paper, two inner products are proposed for induction of metrics applied to vectors of complex numbers obtained from *DFT* transform:

$$\langle d_1, d_2 \rangle_1 = \sum_{t=0}^{T-1} \sum_{k=0}^{B-1} \left| v_t^{d_1}(k) \overline{v_t^{d_2}(k)} \right| \quad (14)$$

and

$$\langle d_1, d_2 \rangle_2 = \sum_{t=0}^{T-1} \sum_{k=0}^{B-1} \Re(v_t^{d_1}(k)) \Re(v_t^{d_2}(k)) + \Im(v_t^{d_1}(k)) \Im(v_t^{d_2}(k)) \quad (15)$$

where $\Re(z)$ and $\Im(z)$ denote the real and imaginary part of complex number $z \in \mathbb{C}$ while $\bar{z}$ denotes the complex conjugate. Using the inner product (14) and induction schemes (9) and (10) the metrics ρ_4 and ρ_5 are created. The inner product (15) was used only for the scheme (10) which yields metric ρ_6 , while scheme (9) leads to the *Euclidean* metric presented above:

$$\rho_4(d_1, d_2) = \sqrt{\langle d_1 - d_2, d_1 - d_2 \rangle_1} \quad (16)$$

$$\rho_5(x, y) = 1 - \frac{\langle d_1, d_2 \rangle_1}{\|d_1\|_1 \|d_2\|_1}, \text{ where } \|d\|_1 = \langle d, d \rangle_1 \quad (17)$$

$$\rho_6(x, y) = 1 - \frac{\langle d_1, d_2 \rangle_2}{\|d_1\|_2 \|d_2\|_2}, \text{ where } \|d\|_2 = \langle d, d \rangle_2 \quad (18)$$

Norm. To induce metric ρ_7 by means of scheme (9), the following norm is proposed:

$$K(d_1, d_2) = \sum_{t=0}^{T-1} \sum_{b=0}^{B-1} \left| v_t^{d_1}(b) + v_t^{d_2}(b) \right| \quad (19)$$

$$\|d\|_3 = K(d, d) = 2 \sum_{t=0}^{T-1} \sum_{k=0}^{B-1} |v_t^d(k)| \quad (20)$$

$$\rho_7(d_1, d_2) = \|d_1 - d_2, d_1 - d_2\|_3 \quad (21)$$

4 Adaptive Potential Active Hypercontours

4.1 Active Hypercontours

The concept of *active hypercontours* (AH) was first introduced in [25] as a generalization of the *active contour* (AC) techniques that are widely used in image analysis ([13,10]). It was shown in [24] that *active contour* method indirectly constructs function k used for the *classification* of image elements. The advantage of this segmentation method is its ability to use any *a priori* knowledge for determination of an optimal *classifier*. Moreover, it can be used in any metric

space $\mathcal{X}$, unlike the majority of traditional methods which usually assume that $\mathcal{X} = \mathbb{R}^n$, where $n \in \mathbb{N}$ (the novelty of the presented approach). The idea of *active hypercontours* has arisen to exchange experience between the domains that so far have been developing separately.

To enable this exchange, the definition of *hypercontour* has been proposed as a generalization of the *contour* which is defined in $\mathbb{R}^2$ and separates only the types of image regions: object and background. *Hypercontour*, however, can separate any given set of objects, described by *features* in metric space $\mathcal{X}$, into an arbitrarily chosen number of classes (regions) L .

Definition 1. Let ρ denote any metric in $\mathcal{X}$, $\mathcal{L} = \{1, \dots, L\}$ denote the set of labels and let $K(x_0, \varepsilon) = \{x \in \mathcal{X} : \rho(x_0, x) < \varepsilon\}$ denote the sphere with center $x_0 \in \mathcal{X}$ and radius $\varepsilon > 0$. The set $h \subseteq \mathcal{X}$ with information about labels of regions it surrounds, is called a *hypercontour* if and only if there exists a function $f : \mathcal{X} \rightarrow \mathbb{R}$ and $p_0 = -\infty, p_1 \in \mathbb{R}, \dots, p_{L-1} \in \mathbb{R}, p_L = \infty$ ($p_1 < p_2 < \dots < p_{L-1}$) such that:

$$h = \{x \in \mathcal{X} : \exists_{l_1, l_2 \in \mathcal{L}, l_1 \neq l_2} \forall_{\varepsilon > 0} \exists_{x_1, x_2 \in K(x, \varepsilon)} \omega(x_1, l_1) \wedge \omega(x_2, l_2)\} \quad (22)$$

where condition $\omega(x, l)$ is true only when $p_{l-1} \leq f(x) < p_l$ and the region $\{x \in \mathcal{X} : \omega(x, l)\}$ represents class $l \in \mathcal{L}$.

It has been proven in [25] that *hypercontour* is equivalent to *classifier* if $\mathcal{X} = \mathbb{R}^n$ and $n \in \mathbb{N}$, but this statement is still true if any other metric space $\mathcal{X}$ is considered (the proof is almost identical). This means that each *classifier* generates a *hypercontour* in each metric space $\mathcal{X}$ which has a sufficient discriminative power to distinguish *classified* objects and conversely, each *hypercontour* unambiguously generates the corresponding *classification* function. The name *hypercontour* is used only to emphasize the relationship of the proposed technique with *active contour* methods.

As mentioned in section 2.1, in order to correctly *classify* objects the optimal *classifier* in $\mathcal{K}$ must be found. In the *active contour* technique, the optimal *contour* in the sens of energy function E is determined. *Energy* function is a kind of *performance index* Q used in optimization methods. The notion *active contour* reflects the ability of the evolution of the *contour* during search of the object in the image. The same idea is used in the *active hypercontour* approach. The search of an optimal *hypercontour* (*classifier*) is performed by means of optimization of the *energy* function $E : \mathcal{H} \rightarrow \mathbb{R}$ where $\mathcal{H}$ denotes the space of all available *hypercontours*.

4.2 Adaptive Potential Active Hypercontours

The idea of *active hypercontours* presented above is convenient for theoretical considerations. For practical use, however, those concepts must be implemented in some way. One possible approach, namely *potential active hypercontour* (PAH) is introduced in [26]. Its generalization for any metric space is presented here.

Potential Hypercontour. Let *feature space* $\mathcal{X}$ be a metric space with a metric $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. The *potential hypercontour* is defined by means of a set of labeled control points: $D^c = \{(\mathbf{x}_1^c, l_1^c), \dots, (\mathbf{x}_{N^c}^c, l_{N^c}^c)\}$ where $\mathbf{x}_i^c \in \mathcal{X}$ and $l_i^c \in \mathcal{L}$ for $i = 1, \dots, N^c$. Each point is a source of potential the value of which decreases with the increase of distance from the source point. The *classifier* k (and consequently the corresponding *hypercontour* h which it generates) is defined in the following way:

$$\forall \mathbf{x} \in \mathcal{X} \quad k(\mathbf{x}) = \arg \max_{l \in \mathcal{L}} \sum_{i=1}^{N^c} P_{\Psi, \mu_i}(\mathbf{x}_i^c, \mathbf{x}) \delta(l_i^c, l) \quad (23)$$

where $\delta : \mathcal{L} \times \mathcal{L} \rightarrow \{0, 1\}$, $l_1 \neq l_2 \Rightarrow \delta(l_1, l_2) = 0$, $l_1 = l_2 \Rightarrow \delta(l_1, l_2) = 1$ and $P : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a *potential function* e.g. *exponential potential* function:

$$P_1 \Psi_\mu(\mathbf{x}_0, \mathbf{x}) = \Psi e^{-\mu \rho^2(\mathbf{x}_0, \mathbf{x})} \quad (24)$$

or *inverse potential* function:

$$P_2 \Psi_\mu(\mathbf{x}_0, \mathbf{x}) = \frac{\Psi}{1 + \mu \rho^2(\mathbf{x}_0, \mathbf{x})} \quad (25)$$

In both cases $\Psi \in \mathbb{R}$ and $\mu \in \mathbb{R}$ are parameters characterizing the potential field. Those parameters and the distribution of control points fully describe the *classifier*.

Energy of the Potential Hypercontour. As it was mentioned at the beginning of this section, the main advantage of *active hypercontour* method is its ability to define *energy* (*objective function*) in an almost arbitrary way. Its choice is crucial for optimal *classifier* construction so the information used to create it and its form should be chosen very carefully. Additionally, the type of that information determines also the type of the *classification* i.e. whether it is *supervised* or *unsupervised*. The types of *energy* functions used in this article are presented separately in section 4.3.

Evolution of the Potential Hypercontour. The form of the *potential hypercontour* and consequently its *classification* ability depend on the distribution of control points in $\mathcal{X}$ and on the parameters Ψ and μ characterizing *potential* functions. In the presented implementation of *potential active hypercontour*, the *simulated annealing* (SA) is used as an optimization algorithm ([14]). The main advantage of this method is that it does not require any information about gradient. Only values of *objective function* are required. Moreover, it is the global optimum search. Any other optimization techniques (e.g. *genetic algorithm*, etc.) can be applied here as well.

Adaptive Potential Active Hypercontours. As described in [26], in some situations, the large number of randomly chosen initial control points can cause problems during the optimization process. For example, if there is no *a priori* information about some fragments of *feature space* $\mathcal{X}$, it can be hard to get out of the local minima. Another problem is the proper choice of the number of control points needed for a satisfactory description of a desired *classifier*. That choice

determines the generalization abilities of *hyperconotur*. To avoid this problem and to improve the performance of the proposed algorithm, an adaptation mechanism can be added to PAH, which results in an *adaptive potential active hyperconotur* (APAH) algorithm. This adaptation allows to start with a small number of control points in D^c and add some new ones, if necessary (see section 4.4). Diverse adaptation algorithms can be applied.

4.3 Energy Functions

The *energy* function mentioned in section 4.2 is used as an objective function in the process of optimization. Its form depends on the *a priori* knowledge used. Below, two *energy* functions are defined.

Energy Basing on the Training Set (E_1). Here, the number of incorrect *classifications* in the given training set of correctly labeled objects is calculated. Let $D^{tr} = \{(\mathbf{x}_1^{tr}, l_1^{tr}), \dots, (\mathbf{x}_{N^{tr}}^{tr}, l_{N^{tr}}^{tr})\}$ where $\mathbf{x}_i^{tr} \in \mathcal{X} \subseteq \mathbf{R}^n$ and $l_i^{tr} \in \mathcal{L}$ for $i = 1, \dots, N^{tr}$ denote a sample training set of correctly labeled vectors. The *energy* E_1 is defined as follows:

$$\forall_{h \in \mathcal{H}} E_1(h) = \sum_{i=1}^{N^{tr}} (1 - \delta(l_i^{tr}, k(x_i^{tr}))) \quad (26)$$

Energy Basing on the Training Set and Expert Knowledge (E_2). As described in section 4.2, the main advantage of *active hypercontours* is their ability to utilize any expert knowledge when constructing the optimal *classifier*. The *energy* function E_2 proposed here is similar to the previous one. The advantage is that additional information about significance of each object from training set can be introduced by the expert. The higher the value of $u_i \in \mathbb{R}$ (where $i = 1, \dots, N^{tr}$) the more significant (he more representative for the considered class) the objects. This *energy* is defined in the following way:

$$\forall_{h \in \mathcal{H}} E_2(h) = \sum_{i=1}^{N^{tr}} u_i (1 - \delta(l_i^{tr}, k(x_i^{tr}))) \quad (27)$$

A similar concept of using experts knowledge was used earlier in the *unsupervised classification* problem presented in [19] where an expert was supposed to evaluate the proximity between pairs of objects from the training set.

4.4 Adaptation Algorithms

Adaptation algorithms are supposed to enhance the performance of the PAH as described in section 4.2. The only adaptation algorithm used in this work is described below:

Adaptation Algorithm Using the Worst LPoints (A_1). This method of adaptation takes each point from the training set and checks whether it is correctly *classified* or not. Then only misclassified objects are considered. The neighborhood of each object given as vector x (in this work it is a sphere with a

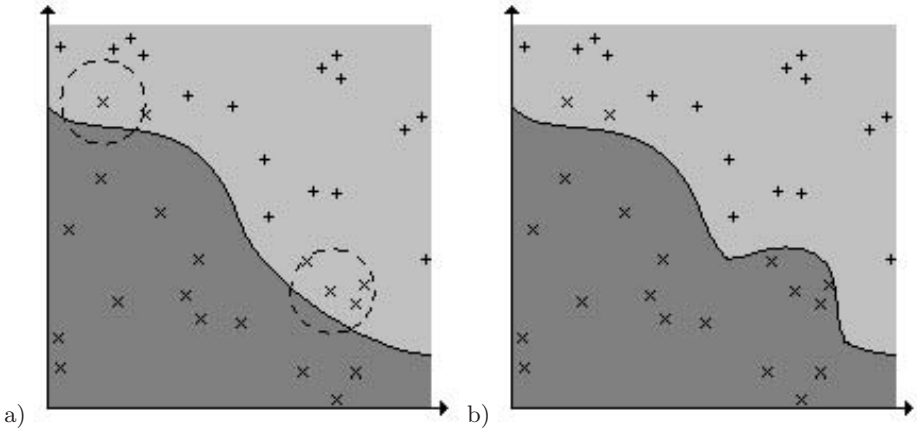


Fig. 1. Adaptation algorithm (A_1)

center in the considered point x and radius equal to r of the maximum distance between objects from the training set) is scanned through to find out how many of other objects from this neighborhood are misclassified in the same way. For each given class l , a new control point (x, l) was added to D^c . The new control point is placed in those of the neighborhoods where the number of incorrect *classifications* to that class was maximum (Fig. 1). This scheme is performed for each class. In the experiments $r = 0.2$ proved to give satisfactory results.

5 Data Used in the Experiments

During the experiments the ODP (*open directory project*) [38] database was used. It is a collection of references to Internet-deployed web sites labeled with a category set having hierarchical structure.

5.1 Category Structure

The categories in which documents are grouped in ODP database have a hierarchical, tree structure. Most of them have additional textual descriptions. Moreover, a lot of documents have additional abstracts. As most of the categories do not contain a satisfying amount of documents, a specific nodes from the tree have been selected, and all documents from subcategories have been joined recursively. The resulting categories labels may be specified by path expressions. Basing on research presented in [19], the categories chosen were:

- *Top/Computers/Software/Graphics**
- *Top/Shopping/Gifts/Personalized**
- *Top/News/Media/Journalism/Photojournalism**

Table 2. Results of k -NN method applied to the data generated using DCT2 transform (percentage of correctly *classified* documents)

$k =$	1	2	3	4	5	6	7	8	9	10
DCT2- $\mathcal{T}_0, \rho_1$	70.5	70.5	72.8	71.5	72.3	71.8	72.0	71.0	71.0	71.0
DCT2- $\mathcal{T}_0, \rho_2$	71.3	71.3	72.8	72.8	73.6	73.8	71.8	74.1	72.3	73.3
DCT2- $\mathcal{T}_0, \rho_3$	69.2	69.2	69.5	69.0	69.5	69.7	69.2	68.2	67.1	67.1
DCT2- $\mathcal{T}_1, \rho_1$	51.2	51.2	56.1	55.3	58.2	60.0	59.7	57.9	58.4	60.0
DCT2- $\mathcal{T}_1, \rho_2$	49.7	49.7	53.3	55.1	56.6	57.4	56.9	58.4	58.4	57.9
DCT2- $\mathcal{T}_1, \rho_3$	50.7	50.7	53.3	54.8	53.8	54.1	54.8	56.6	56.6	57.9

Table 3. Results of APAH method applied to the data generated using DCT2 transform (percentage of correctly *classified* documents)

	E_1, A_1	E_2, A_1
DCT2- $\mathcal{T}_0, \rho_1$	73.1	73.1
DCT2- $\mathcal{T}_0, \rho_2$	64.4	60.5
DCT2- $\mathcal{T}_0, \rho_3$	71.0	71.0
DCT2- $\mathcal{T}_1, \rho_1$	60.8	61.0
DCT2- $\mathcal{T}_1, \rho_2$	61.3	55.9
DCT2- $\mathcal{T}_1, \rho_3$	56.9	60.5

The presented technique supplied data sets of desired size (more than 600 documents) without sacrificing category specificity. Selection of more top-level categories results in a bigger set, but document *classification* becomes doubtful and disjoint.

5.2 Term Selection

The effectiveness of the method depends on the selection of *feature* description because document index is created on this basis. To avoid excessive growing of vector dimension only a small subset of the terms occurring in documents were considered. The first set contains terms arbitrarily selected from descriptions of categories: $\mathcal{T}_0 = \{image, software, digital, manipulate, transfer, gift, photo, logo, concert, music, journal, promotion, portrait\}$. For this set no stemming algorithm was applied. The values $f_t^d(0), \dots, f_t^d(B-1)$ were incremented if the word contained specific term $\tau \in \mathcal{T}_0$. An additional data set was also prepared with letters as premises: $\mathcal{T}_1 = \{a, b, c \dots z\}$. Similarly, counters were incremented if a specific word contained a given letter. For clarity of presentation, let us consider the word *photojournalism* and premises *photo*, *journal*, *logo* and from the other set *o* and *z*. When the analyzed word appears in some bin, it increments counters for premises *photo*, *journal* and *o* by 1, as these premises are sub-words of *photojournalism* and the do not modify the remaining counters for *logo* and *z*.

5.3 Preparation of Data Sets and Parameters

All documents used in the experiments have been divided into $B = 8$ bins. The set of documents was randomly split into training and test sets containing 500 and 130 documents respectively. The weighting scheme $\text{TF} \times \text{IDF}$ was applied and *Fourier*, *Hartley* and *cosine II* discrete transforms were used to generate *features*. As it was done for both $\mathcal{T}_0$ and $\mathcal{T}_1$ sets, six experimental data sets were prepared: DFT- $\mathcal{T}_0$, DFT- $\mathcal{T}_1$, DCT2- $\mathcal{T}_0$, DCT2- $\mathcal{T}_1$, DHT- $\mathcal{T}_0$ and DHT- $\mathcal{T}_1$. Special attention was paid to the elimination of *zero* elements before generation of data sets.

Additionally, information about significance of documents (used in *energy* function E_2) was prepared by an expert who looked through the documents and selected the most representative ones for each class, assigning them proper significance coefficients.

6 Discussion of the Results

The experiments were performed using k -NN method and APAH technique where *inverse potential* function P_2 was chosen.

6.1 k -NN Results

Experiments presented in Table 2, Table 4 and Table 6 reveal similarities between results obtained by the use of discrete transforms with standard metrics ρ_1 , ρ_2 , ρ_3 . Here, *Manhattan metric* seems to perform best. It is interesting to discuss the results obtained by the use of DFT transform with consideration of phase information. Test results obtained with metric ρ_5 show that the use of spatial information may improve the *classification* performance. However, incorporation of expert knowledge, even in the simple form [21], is difficult in both methods ϕ and k -NN.

6.2 APAH Results

Active hypercontours are flexible in adopting external knowledge during classifier preparation. This is important, as it may be used without modification of document indexes. The method is also much more flexible because diverse requirements can be considered by proper definition of the *energy* function.

The results presented in Table 5 show that metrics ρ_4 , ρ_5 , ρ_6 and ρ_7 used with APAH similarly to k -NN give better results than classic metrics. Moreover, also in this case the use of *phase* information improves classification results.

It is worth mentioning that the construction of a *classifier* can be time-consuming. Computational effort depends on the structure of *feature* space, the type of *energy* function, etc. However, it is not a crucial disadvantage because the construction of the *classifier* is performed only once (unless new information is provided). The *classification* itself is very quick, which in the case of k -NN applied for a huge training set may not be true. Secondly, as a consequence of

using *simulated annealing* as the optimization algorithm, the finally obtained *classifiers* can be different (algorithm is not deterministic). But as mentioned above it is not a problem either, because many *hypercontours* can evolve and the best are chosen. The results presented here are generated by the best obtained *classifier*.

Moreover, all the presented results are obtained in the same way. The APAH algorithm starts with one control point for each class and in each successive adaptation step one additional point for each class is added (4.4). The experiments show that 5 adaptation steps suffice (at least for the data used). Consequently, the final *hypercontour* was determined by 18 control points, i.e. 6 points for each class.

Table 4. Results of k -NN method applied to the data generated using DFT transform (percentage of correctly *classified* documents)

$k =$	1	2	3	4	5	6	7	8	9	10
DFT- $\mathcal{T}_0, \rho_1$	71.0	71.0	71.8	71.8	72.8	71.8	72.0	71.8	70.5	70.5
DFT- $\mathcal{T}_0, \rho_2$	70.5	70.5	72.8	74.6	72.0	74.8	73.3	74.1	73.3	73.0
DFT- $\mathcal{T}_0, \rho_3$	68.4	68.4	67.7	69.2	69.2	69.2	67.7	66.6	65.9	64.6
DFT- $\mathcal{T}_0, \rho_4$	71.0	71.0	71.8	71.8	72.8	71.8	72.0	71.8	70.5	70.5
DFT- $\mathcal{T}_0, \rho_5$	73.0	73.0	76.2	76.2	76.7	76.7	76.9	77.2	77.2	77.2
DFT- $\mathcal{T}_0, \rho_6$	72.6	72.6	73.6	73.8	74.1	73.3	74.6	74.4	74.9	74.4
DFT- $\mathcal{T}_0, \rho_7$	69.0	69.0	72.6	73.6	73.8	73.6	73.8	73.8	73.6	73.1
DFT- $\mathcal{T}_1, \rho_1$	48.5	48.5	54.1	56.1	57.4	55.1	58.2	58.7	58.2	58.5
DFT- $\mathcal{T}_1, \rho_2$	49.4	49.4	54.8	53.3	56.9	55.6	58.9	57.4	59.2	56.9
DFT- $\mathcal{T}_1, \rho_3$	49.4	49.4	54.6	56.4	54.6	55.1	57.1	58.9	56.1	58.4
DFT- $\mathcal{T}_1, \rho_4$	48.5	48.5	54.1	56.1	57.4	55.1	58.2	58.7	58.2	58.5
DFT- $\mathcal{T}_1, \rho_5$	50.5	50.5	55.1	56.7	58.2	57.4	59.5	59.2	59.5	59.7
DFT- $\mathcal{T}_1, \rho_6$	53.8	53.8	56.7	54.4	56.4	54.6	55.6	56.9	54.6	52.5
DFT- $\mathcal{T}_1, \rho_7$	51.5	51.5	54.9	56.1	57.4	59.7	59.5	59.5	57.2	57.9

6.3 Comparison

Both methods work better for the set of terms $\mathcal{T}_0$. It seems to be obvious because the second set allows to extract less significant semantic information. However, even such a simple set of terms as $\mathcal{T}_1$, allows to recognize correctly about 60 per cent of documents. In the second case, the APAH method usually gives better results than k -NN (especially for DHT). Moreover, in the majority of the experiments performed for $\mathcal{T}_1$, the expert knowledge encoded in *energy* function E_2 improves *classification*. It happens more frequently with $\mathcal{T}_1$ than $\mathcal{T}_0$. The form in which expert knowledge was incorporated into the *energy* function proved to be sufficient when working with $\mathcal{T}_1$ whereas in the case of $\mathcal{T}_0$ the distribution of training objects was good enough and additional knowledge in this particular form did not help significantly.

Analyzing the results it can be also noticed that sometimes the APAH method gives worse results than k -NN. The reason of that may be fact that, as it was

Table 5. Results of APAH method applied to the data generated using DFT transform (percentage of correctly *classified* documents)

	E_1, A_1	E_2, A_1
DFT- $\mathcal{T}_0, \rho_1$	70.0	70.0
DFT- $\mathcal{T}_0, \rho_2$	62.3	61.3
DFT- $\mathcal{T}_0, \rho_3$	70.5	72.1
DFT- $\mathcal{T}_0, \rho_4$	73.8	69.7
DFT- $\mathcal{T}_0, \rho_5$	75.3	77.9
DFT- $\mathcal{T}_0, \rho_6$	76.4	76.9
DFT- $\mathcal{T}_0, \rho_7$	72.8	72.6
DFT- $\mathcal{T}_1, \rho_1$	62.3	62.6
DFT- $\mathcal{T}_1, \rho_2$	57.7	57.7
DFT- $\mathcal{T}_1, \rho_3$	56.0	57.7
DFT- $\mathcal{T}_1, \rho_4$	60.5	61.0
DFT- $\mathcal{T}_1, \rho_5$	53.1	63.3
DFT- $\mathcal{T}_1, \rho_6$	60.8	58.2
DFT- $\mathcal{T}_1, \rho_7$	62.8	63.1

mentioned in previous section, the evolution process is not deterministic and during experiments only such classifiers were found. Moreover, in the presented experiments the number of control points were constant. It allows to achieve only a limited class of *hypercotours* and can be not able to separate objects satisfactorily. Moreover, improvement of the adaptation mechanism can increase the classification quality. In general however, the idea of using *adaptive potential active hypercontours* with additional expert knowledge for document *classification* task proved to be very promising because it enabled the construction of *classifiers* that were better than the classic methods.

Table 6. Results of k -NN method applied to the data generated using DHT transform (percentage of correctly *classified* documents)

$k =$	1	2	3	4	5	6	7	8	9	10
DHT- $\mathcal{T}_0 \rho_1$	70.0	70.1	71.8	72.0	71.5	72.5	71.0	70.7	70.5	70.5
DHT- $\mathcal{T}_0 \rho_2$	71.8	71.8	73.0	73.8	72.8	74.3	73.5	74.3	74.3	75.8
DHT- $\mathcal{T}_0 \rho_3$	68.9	68.9	67.9	68.2	68.4	68.2	67.9	67.4	65.9	66.4
DHT- $\mathcal{T}_1 \rho_1$	47.9	47.9	53.5	53.8	56.4	58.4	58.7	58.2	59.2	59.4
DHT- $\mathcal{T}_1 \rho_2$	51.2	51.2	54.1	53.5	56.9	59.2	59.7	56.4	57.1	57.4
DHT- $\mathcal{T}_1 \rho_3$	49.2	49.2	54.1	53.3	55.9	56.6	55.1	57.1	57.4	56.1

7 Summary

The Web is an enormous, unlimited and dynamically changing source of various kinds of useful information. The aspiration to meet an obvious need for effective information retrieval by making improvements in the solutions offered by the

Table 7. Results of APAH method applied to the data generated using DHT transform (percentage of correctly *classified* documents)

	E_1, A_1	E_2, A_1
DHT- $\mathcal{T}_0 \rho_1$	69.2	68.7
DHT- $\mathcal{T}_0 \rho_2$	63.8	63.1
DHT- $\mathcal{T}_0 \rho_3$	70.0	69.0
DHT- $\mathcal{T}_1 \rho_1$	51.3	57.7
DHT- $\mathcal{T}_1 \rho_2$	51.5	57.0
DHT- $\mathcal{T}_1 \rho_3$	51.5	55.6

popular indexing engines is observed. It is also expected that the resources will become well structured and the structure (clusters, trees) will reflect the content of documents. Here, the following tasks must be considered: representation of documents (models), *clustering* and *classification*.

Following the taxonomy proposed in [28] the models may be divided into two groups, namely classic and alternative. The former include vector-space, probabilistic and Boolean models. The latter are remarkably numerous and varied; they cover: fuzzy set models, extended Boolean models, generalized *vector space models*, neural network models, etc. For classic representation, the *vector space model* (VSM) is commonly used [28,29]. VSM is, in a nutshell, the representation of textual document by a vector; within the vector, whose elements are numbers characterizing the *weight* (significance) of term τ in document d . These so-called indexing weights can be computed from frequencies of occurrence of the terms in a given document [33,34]. Document vectors can then be easily compared to each other, or to query vectors. However, the information about the spatial location of words is not considered. The hypothesis that the information about the spatial location of words is an important factor which can improve the document search and ranking, has been recently confirmed in practice [17,18]. The spatial information was preserved by application of *Fourier* and *cosine* discrete transformations, which allows to store a term signal, and therefore to follow the spreading of the term throughout the document. The magnitude and phase are the *features* used for comparison of query and document terms.

Classification of documents is the process of assigning a proper label l to the given document, and *classifier* is a function which realizes this assignment. The *active contour* methods originating from image analysis can be adopted to solve the multidimensional classification task [25], since each *hypercontour* generates a *classification* function. Having any vector representation of documents and an implementation of the *active hypercontour* concept (here the novel *adaptive potential active hypercontour*), the *supervised classification* of documents can be performed. Moreover, because of existence of the *energy* function, the user obtains the possibility to influence the *classifier* construction process

by incorporation of human knowledge about the problem, for example about the level of relevance of the *classified* documents.

Representation of documents with the use of discrete transformations, application of *adaptive hypercontours*, and incorporation of expert knowledge are pillars of the method proposed and evaluated in practice. The synergy of these three components leads to a novel flexible approach for document *classification* which is able to give better *classification* results than the other classic methods.

References

1. Zhong, N., Liu, J., Yao, Y.Y.: Web intelligence. Springer, Heidelberg (2003)
2. Zhong, N.: Developing Intelligent Portals by Using WI Technologies. In: Li, J.P., et al. (eds.) Wavelet Analysis and Its Applications, and Active Media Technology, vol. 2, pp. 555–567. World Scientific, Singapore (2004)
3. Zhong, N., Liu, J. (eds.): Intelligent Technologies for Information Analysis. Springer, Heidelberg (2004)
4. Szczepaniak, P.S., Segovia, J., Kacprzyk, J., Zadeh, L.A. (eds.): Intelligent Exploration of the Web, 2nd edn. Physica Verlag, Heidelberg (2003)
5. Menasalvas, E., Segovia, J., Szczepaniak, P.S. (eds.): AWIC 2003. LNCS (LNAI), vol. 2663. Springer, Heidelberg (2003)
6. Favela, J., Menasalvas, E., Chávez, E. (eds.): AWIC 2004. LNCS (LNAI), vol. 3034. Springer, Heidelberg (2004)
7. Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.): AWIC 2005. LNCS (LNAI), vol. 3528. Springer, Heidelberg (2005)
8. Last, M., Szczepaniak, P.S., Volkovich, Z., Kandel, A. (eds.): Advances in Web Intelligence and Data Mining. Springer, Heidelberg (2006)
9. Wegrzyn-Wolska, K., Szczepaniak, P.S. (eds.): Advances in Intelligent Web Mastering. Springer, Heidelberg (2007)
10. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. International Journal of Computer Vision 22(1), 61–79 (1997)
11. Bishop, C.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1993)
12. Joachims, T.: Text Categorization using Support Vector Machines: Learning with Many Relevant Features. Research Reports of the unit nr VIII. Computer Science Department of the University of Dortmund. Dortmund (1997)
13. Kass, M., Witkin, W., Terzopoulos, D.: Snakes: Active Contour Models. International Journal of Computer Vision, 321–331 (1988)
14. Kirkpatrick, S., Gerlatt, C.D., Vecchi Jr., M.P.: Optimization by Simulated Annealing. Science 220, 671–680 (1983)
15. Kwiatkowski, W.: Methods of Automatic Pattern Recognition. WAT. Warsaw (in Polish) (2001)
16. Looney, C.: Pattern Recognition Using Neural Networks. Theory and Algorithms for Engineers and Scientists. Oxford University Press, New York (1997)
17. Park, L.A.F., Palaniswami, M., Ramamohanarao, K.: Fourier Domain Scoring: A Novel Document Ranking Method. IEEE Trans. on Knowledge and Data Engineering 16(5), 529–539 (2004)
18. Park, L.A.F., Ramamohanarao, K., Palaniswami, M.: A Novel Web Text Mining Method Using the Discrete Cosine Transform. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 385–396. Springer, Heidelberg (2002)

19. Pedrycz, W., Loia, V., Senatore, S.: P-FCM: A proximity-based fuzzy clustering. *Fuzzy Sets and Systems* 128, 21–41 (2004)
20. Pedrycz, W.: *Knowledge-Based Clustering*. Wiley-Interscience, Hoboken, New Jersey (2005)
21. Szczepaniak, P.S., Pryczek, M.: Web Textual Documents Scoring Based on Discrete Transforms and Fuzzy Weighting. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) *AWIC 2005. LNCS (LNAI)*, vol. 3528, pp. 415–420. Springer, Heidelberg (2005)
22. Szczepaniak, P.S., Pryczek, M.: On Textual Documents Classification Using Fourier Domain Scoring. In: *Proceedings of 2006 IEEE /WIC/ACM International Conference on Web Intelligence (WI 2006)*, Hong Kong, IEEE Computer Society Press, Los Alamitos (2006)
23. Tadeusiewicz, R., Flasiński, M.: *Pattern Recognition*. PWN. Warsaw (in Polish) (1991)
24. Tomczyk, A., Szczepaniak, P.S.: On the Relationship between Active Contours and Contextual Classification. In: Kurzynski, 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)
25. Tomczyk, A.: Active Hypercontours and Contextual Classification. In: *5th International Conference on Intelligent Systems Design and Applications (ISDA)*, Wrocław, Polska, pp. 256–261. IEEE Computer Society Press, Los Alamitos (2005)
26. Tomczyk, A., Szczepaniak, P.S.: Adaptive Potential Active Hypercontours. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L., Zurada, J. (eds.) *ICAISC 2006. LNCS (LNAI)*, vol. 4029, pp. 692–701. Springer, Heidelberg (2006)
27. Węgrzyn-Wolska, K., Szczepaniak, P.S.: Classification of RSS-formatted Documents using Full Text Similarity Measures. In: Lowe, D.G., Gaedke, M. (eds.) *ICWE 2005. LNCS*, vol. 3579, pp. 400–405. Springer, Heidelberg (2005)
28. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison Wesley, New York (1999)
29. Witten, I.H., Moffat, A., Bell, T.C.: *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, San Francisco (1999)
30. Pal, S.K., Talwar, V., Mitra, P.: Web Mining in Soft Computing Frameworks: Relevance, State of the Art and Future Directions. *IEEE Trans. on Neural Networks* 13(5) (2002)
31. Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function*. Plenum Press, New York (1981)
32. Jain, A.K., Dubes, R.C.: *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs (1988)
33. Kraft, D.H., Chen, J.: Integrating and Extending Fuzzy Clustering and Inferencing to Improve Text Retrieval Performance. In: Larsen, H.L., et al. (eds.) *Flexible Query Answering Systems*, Springer, Heidelberg (2001)
34. Kraft, D.H., Chen, J., Martin-Bautista, M.J., Amparo-Vila, M.: Textual Information Retrieval with User Profiles using Fuzzy Clustering and Inferencing. In: Szczepaniak, P.S., Segovia, J., Kacprzyk, J., Zadeh, L. (eds.) *Intelligent Exploration of the Web*, Springer, Heidelberg (2003)
35. Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function*. Plenum Press, New York (1981)

36. Bezdek, J.C.: A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 2, 1–8 (1980)
37. Bezdek, J.C., Hathaway, R.J., Sabin, M.J., Tucker, W.T.: Convergence Theory for Fuzzy c-Means: Counterexamples and Repairs. *IEEE Trans. on Systems, Man, and Cybernetics* 17, 873–877 (1987)
38. ODP, <http://www.dmoz.org/>

Fuzzy Web Surfer Models: Theory and Experiments

Narayan L. Bhamidipati and Sankar K. Pal

Machine Intelligence Unit,
Indian Statistical Institute,
203, B. T. Road,
Calcutta - 700108, India
{bln-r, sankar}@isical.ac.in

Abstract. A novel web surfer model, where the transition probabilities are fuzzy quantities, is proposed in this article. Based on the theory of Fuzzy Markov Chains, we introduce FuzzRank, which is the counterpart of PageRank. Apart from discussing the theoretical aspects of fuzzy surfer models and FuzzRank, we have also compared its ranking, convergence and robustness properties with PageRank. Extensive experimental results and a detailed example depict the advantages of FuzzRank over PageRank.

1 Introduction

The World Wide Web is a complex entity, with billions of web pages and many more links between them. That there are new pages constantly being added, some old pages being deleted, and existing pages being changed all the time, adds to the complexity of the situation. Web Intelligence [1] exploits Artificial Intelligence and advanced Information Technology on this complex and heterogeneous Web made up of vastly diverse content.

As the web consists of pages created by millions of individuals, there is a wide variety of authoring styles. Most present day content and link analysis algorithms are robust against differences in fonts, colors, *etc.*, which are mostly ornamental. Some others can withstand, to some extent, malicious manipulation of content and links. However, they are sensitive to whether the information is contained in a single document or is spread out in a collection of documents. For the sake of uniformity in comparison during content and link analysis, information present in a single web page may be artificially divided into a collection of web pages. This division introduces an uncertainty in the page boundaries as well as the targets of hyperlinks.

A variety of web surfer models exist which model the sequence of web pages a surfer follows as a Markov process. The transition probabilities are obtained by considering the number of links in each page. Here, it is assumed that there is no uncertainty in the given web pages or the transition probabilities. In practice, this is not the case. This imprecision may be modeled with the help of fuzzy sets, or in particular, by fuzzy numbers. This forms the basis for the present investigation, where we extend existing surfer models to fuzzy surfer models (Fuzzy web surfer models were first introduced by the authors in [2], and the present article is an extended version of it). Since we now deal with fuzzy numbers, Markov chain theory is replaced by fuzzy Markov chain theory,

which employs the max-min (or fuzzy) algebra instead of the classical algebra with multiplication and addition operations. These models may be employed, among other things, to compute ranks of web pages, which we call FuzzRanks. We believe that these models add to the set of tools needed for the development of intelligent information technologies [3] to applied in the areas of web intelligence [4].

Fuzzy web surfer models described in this chapter, apart from being able to handle fuzziness in various aspects, inherit the advantages of fuzzy Markov models, namely, robustness and finite convergence. Robustness is a very important aspect because it implies that small changes in the transition matrix would not change the results drastically. Its significance arises from the fact that the transition matrices are not known beforehand and are estimated during the analysis phase, and so, (slightly) different methods of estimation, may lead to immensely dissimilar results. As a consequence, FuzzRank is more stable as compared to PageRank. FuzzRank reflects the belief of a surfer being on a page, and cannot fluctuate to extreme cases as in the case of probabilistic models.

This chapter is organized as follows. Section 2 discusses the preliminaries such as fuzzy sets, Markov chains, fuzzy Markov chains and web surfer models. We make use of these components to describe fuzzy web surfer models in Section 3, which we begin with a few motivational examples, and also define FuzzRank, which is the fuzzy-equivalent of PageRank. Section 4 consists of an illustrated example, and several experimental results, which convincingly demonstrate the advantages of FuzzRank over PageRank. Section 5, concludes the chapter and mentions some future directions of research on this topic.

2 Preliminaries and Background

We now provide the background as well as the notation on fuzzy sets, Markov chains, and fuzzy Markov chains.

2.1 Fuzzy Sets

Conventional sets consist of a group of elements. An element of the universe (Ω) may or may not belong to a given set, and only one of these two possibilities may happen. However, for the sake of situations where it is not clear if an element belongs to a set or not, the concept of fuzzy sets was proposed [5]. A fuzzy set is a generalization of the conventional set, where there is some measure of uncertainty of membership in the set. For a fuzzy set S , there is a membership function associated with it which provides a membership value for each element in Ω .

$$\mu_S : \Omega \rightarrow [0, 1] \quad (1)$$

Generally, μ is so chosen that $\max_{x \in \Omega} \mu_S(x) = 1$, in which case, it is said to be normalized.

The union and intersection operations of the classical sets is extended to the fuzzy sets, by taking the max and min, respectively, of the corresponding membership values of each element.

2.2 Markov Chains

A (first order) Markov chain [6] is a sequence $\{X_n\}_{n \in \mathbb{N}}$ of random variables where each random variable, X_i , takes a value from a state space S , and the sequence satisfies

$$P(X_{n+1}|X_0, X_1, \dots, X_n) = P(X_{n+1}|X_n). \quad (2)$$

$\{X_n\}$ is called homogeneous if $P(X_{n+1}|X_n)$ is independent of n . In this chapter, we shall deal with only discrete, homogeneous Markov chains, with finite state space $S = \{1, 2, \dots, N\}$. Let p_{ij} denote $P(X_{n+1} = i|X_n = j)$, which is the one step transition probability from state i to state j . $P = ((p_{ij}))_{i,j \in S}$ is called the (one-step) transition probability matrix. The probability of X_{n+1} assuming a state j is given by

$$\begin{aligned} P(X_{n+1} = j) &= \sum_{i=1}^N P(X_{n+1} = j|X_n = i)P(X_n = i) \\ &= \sum_{i=1}^N p_{ij}P(X_n = i) \end{aligned} \quad (3)$$

Now, the m -step transition probability from i to j , denoted by $p_{ij}^{(m)}$ may be expressed in terms of p_{ij} (which is the same as $p_{ij}^{(1)}$) as:

$$p_{ij}^{(m)} = P(X_{n+m} = i|X_n = j) \quad (4)$$

$$\begin{aligned} &= \sum_{k=1}^m \prod_{k=1}^m P(X_{n+k} = x_{n+k}|X_{n+k-1} = x_{n+k-1}) \\ &= \sum_{x_{n+1}, \dots, x_{n+m-1} \in S} \prod_{k=1}^m p_{x_{n+k}, x_{n+k-1}} \end{aligned} \quad (5)$$

From this expression, it may be observed that the m -step transition probability matrix $P^{(m)}$ is the same as P^m , the m -th power of P .

A state i is called aperiodic if $\gcd\{n : p_{ii}^{(n)} > 0\}$ is 1. A Markov chain is called aperiodic if all the states in S are aperiodic. It is called irreducible if every pair of states in S are reachable from each other. A finite, aperiodic, irreducible Markov chain is called regular, and $P^n > 0$ for some $n \geq 1$ for regular Markov chains.

For a regular Markov chain, $p_{ij}^{(n)} \rightarrow \pi_j \forall i, j \in S$. $\pi = (\pi_1, \pi_2, \dots, \pi_N)$ is called the stationary distribution of the Markov chain. This property is termed ergodicity, and means that, regardless of its initial state, $P(X_n = j)$ converges to a unique π_j . The convergence and uniqueness of the chain are guaranteed only if the chain is aperiodic and irreducible, respectively.

2.3 Fuzzy Markov Chains

The probabilities in the previous Subsection are real numbers and are all assumed to be known. In practice, they are estimated, and there are errors associated with the estimation procedure, which in turn, may again be estimated under suitable assumptions. The

uncertainty in the transition probabilities may sometimes be better modeled in terms of fuzzy numbers.

In order to define a fuzzy Markov chain, we first define a fuzzy distribution and a fuzzy transition matrix.

A fuzzy distribution on S is defined by a mapping $\mu_x : S \rightarrow [0, 1]$, and is represented by a vector $\mathbf{x} = (\mu_x(1), \dots, \mu_x(N))$.

A fuzzy transition matrix P is defined as a fuzzy distribution on the Cartesian product $S \times S$. P is represented by a matrix $((p_{ij}))_{i,j \in S}$ [7]. With this notation, a fuzzy Markov chain is defined as a sequence of random variables, where the transitions are determined by the fuzzy relation P and satisfy

$$\mu_{x^{(n+1)}}(j) = \max_{i \in S} \{ \mu_{x^{(n)}}(i) \wedge p_{ij} \}, j \in S \quad (6)$$

Equation 6 is the fuzzy algebraic equivalent of the transition law of classical Markov chains provided in Eq. 3. The multiplication and addition operations in Eq. 3 have been replaced by the min and max operations, respectively. Naturally, the powers of the matrix P may be defined in the same manner as earlier. The interesting result is that, unlike the case of classical Markov chains, whenever the sequence of matrices P^n converges, it does so in finitely many steps to a matrix P^τ . If it does not converge, it oscillates with a finite period ν starting from some finite power. The above statements are proved rather easily [7].

When the powers of P converge to a non-periodic solution P^τ , the associated fuzzy Markov chain is called aperiodic and P^τ is called a limiting fuzzy transition matrix. A fuzzy Markov chain is called ergodic if the rows of P^τ are identical. This definition is again similar to that of classical Markov chains, but the necessary and sufficient conditions for ergodicity are not known in this case [7].

2.4 Surfer Models

A variety of surfer models, such as *random surfer* [8], *HITS (Hypertext Induced Topic Selection)* [9], *directed surfer* [10], *topic-sensitive pagerank* [11], *topic continuity based surfer model* [12], etc., are available in the literature. More recently, another model called WPSS (*Web Page Scoring Systems*), which generalizes several of the above mentioned models, has been proposed in [13]. We describe the random surfer model and HITS, which have been studied in most detail. The other models are generalizations, extensions, or combinations of these two.

The random surfer model models a user who keeps visiting new pages by clicking, at random, links available on the current page. Thus, given that the surfer is on page v at time t , the probability of his being on page u at time $t + 1$, $P(X_{t+1} = u | X_t = v)$, is assumed to be $\frac{1}{|F_v|}$, where F_v is the set of forward links from v . Therefore, the probability of the surfer being on page u at time $t + 1$, $P(X_{t+1} = u)$, is computed as $\sum_{v=1}^N P(X_{t+1} = u | X_t = v) P(X_t = v)$, where N is the total number of pages. Counting only those pages v for which $P(X_{t+1} = u | X_t = v) > 0$ (i.e., those pages which have a link to u), we have $P(X_{t+1} = u) = \sum_{v \in B_u} \frac{P(X_t = v)}{|F_v|}$, where B_u denotes the set of backlinks of u .

Let the transition matrix for the stochastic process $\{X_t\}$ be denoted by

$$P = ((p_{uv}))_{u,v \in \{1,2,\dots,N\}}.$$

We then have, $p_{uv} = \frac{l_{uv}}{\sum_{w=1}^N l_{uw}}$, where l_{uv} denotes the (u, v) th element of the link matrix of the web graph, and is defined as being equal to 1 if and only if v has a link to u , for $u, v \in \{1, 2, \dots, N\}$. Let R_u^t denote $P(X_t = u)$. Then, the probability distribution of X_{t+1} , $\mathbf{R}^{(t+1)} = (R_1^{t+1}, R_2^{t+1}, \dots, R_N^{t+1})^T$, may be recursively defined as, $\mathbf{R}^{(t+1)} = P\mathbf{R}^{(t)}$.

If this stochastic process has a stationary distribution, it would satisfy $\mathbf{R} = P\mathbf{R}$, for some vector $\mathbf{R}$. The u th element of the vector $\mathbf{R}$ is the unconditional probability of the surfer being on page u , and may be considered to be the rank of page u , and $\mathbf{R}$ may be called the rank vector. To compute $\mathbf{R}$, which is nothing but the dominant (or principal) eigenvector of P , the power method is employed, whereby $\mathbf{R}^t$ converges to $\mathbf{R}$ as $t \rightarrow \infty$. This is the basic idea behind the PageRank algorithm suggested by Brin and Page [8] and relies on the theory of Markov chains.

The HITS algorithm [9] proposed by Jon Kleinberg has a slightly different approach for ranking web pages. Here, each page has a hub score and an authority score. The hub score is proportional to the authority scores of the pages that the page leads to. Likewise, the authority score of a page is defined in terms of the hub scores of the pages leading to it. This too may be studied as a surfer model, where a surfer alternately chooses one of the forward and backward links to jump to a new page. It turns out that the stationary vectors are the dominant eigenvectors in this case too, for suitably chosen matrices. Let $G = ((a_{ij}))$ denote the adjacency matrix, with $G_{ij} = 1$ if page j has a link to page i . Then, the hub and authority values may be obtained as the principal eigenvectors of GG^T and G^TG , respectively [9]. This model is not as realistic as the random surfer model because, in practice, a surfer would not be able to know the set of pages leading to the page he currently is on. Theoretically, however, it is equally interesting to study the long term behaviour of such a surfer.

We now propose a web surfer model that relies on the theory of fuzzy Markov chains. Since web surfer models and fuzzy Markov chains have been described in detail already, to avoid duplication, we describe the proposed methodology in a concise manner, making use of the notations and notions of this section.

3 Fuzzy Web Surfer Models

3.1 Motivation

We look at a few examples which demonstrate the need for new surfer models to deal with various kinds of uncertainty on the web. Authoring styles on the web vary widely and this results in the same kind of content being displayed in various formats. For example, the same content may be packed in one (possibly, big) document, or may be spread out across several linked list of documents. The process of retrieval and ranking are sensitive to such differences, which are usually a simple consequence of contrasting tastes or conveniences.

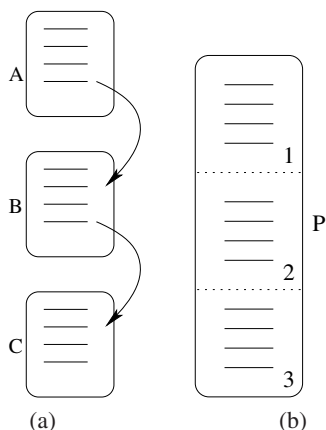


Fig. 1. Equivalent information in (a) HTML and (b) PDF

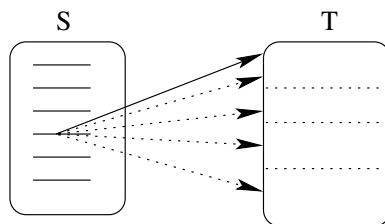


Fig. 2. Which section is being pointed to? Actual target is fuzzy

With most search engines indexing an increasing number of documents in PDF, PS and other formats, this situation is encountered all the more often. Fig. 1 shows a set of HTML pages, and a PDF document, both containing equivalent information. However, when the web is treated as consisting of individual atomic documents, it results in unfair comparison, as the PDF document has more content compared to each of the individual pages. One way to improve the level of fairness during such comparisons is to detect equivalent information, even if it is split across documents in one place and not the other, as performed in [14].

Fig. 2 shows a link from a source document S to a target document T. Now, T contains a lot of information, but the link is just for a particular portion of the page T. The question is which is the portion of T being implicitly referred to by this link. The relevance of this question lies in the fact that the link points to the web page T as a whole and so the weight being transferred through this link spills over to all of T instead of being restricted to the intended portion only.

To strengthen this argument, we provide a real life example.

Fig. 3 shows a portion of the web page located at [www.webmasterworld.com](http://www.webmasterworld.com/forum10003/428.htm)¹ which contains a link to Jon Kleinberg's home page². The link that leads to Kleinberg's home page provides no more information than its URL. The home page under consideration has two named sections, namely *Papers* and *Links*, and there is an introduction above it. It is clear from the context that the above mentioned link indeed refers to the *Papers* portion of the page. In addition, the *Papers* section is further subdivided according to the topics of the papers, but the subsections are not named. Had they been named, we can once again conclude that the link in question actually leads to the *Web Analysis and Search: Hubs and Authorities* subsection.

¹ <http://www.webmasterworld.com/forum10003/428.htm>

² <http://www.cs.cornell.edu/home/kleinber/index.html>

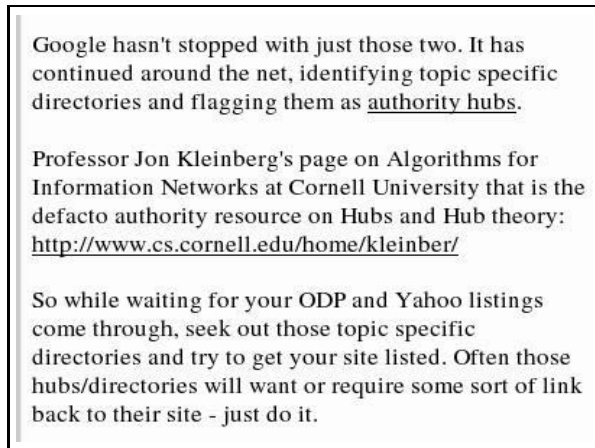


Fig. 3. A portion of a web page at Webmasterworld with a link to Jon Kleinberg's home page

The above examples demonstrate that:

- a link to a web page may in reality be referring to just one or more pagelets, and not the whole page itself. Resolving which pagelet is referred to by a link needs contextual information, and yet this may not be precise.
- a page may have to be artificially divided into pagelets or sections, to avoid the weight attributed by a link to one pagelet spilling over to other pagelets. As mentioned earlier, this is required for fair comparison during retrieval because, although this particular link is for a small portion of the page, the contents of the rest of the page benefit from it, thus enjoying a better status as compared to similar content elsewhere.

It may be noted that it is not claimed that one or the other is necessarily better, because some systems may assign more weight to more content, whereas, others may penalize it. All that is being argued for is that such disparities may lead to diverse results, and need to be addressed at an early stage of link and content analysis. We now formulate a basic methodology for fuzzy web surfer models.

3.2 Formulation

In what follows, we assume that the web pages have been preprocessed with the goal of increasing uniformity among them. By uniformity, we mean that the differences due to authoring styles, as explained earlier are reduced. There are two approaches for that. One approach is that big pages are split into pagelets [15,16], and each of them can be called a new page. The other option is to merge related pages, making each of them a section (or pagelet) of one large page [17,12]. The second approach is well suited for retrieval tasks by virtue of providing a larger coverage. However, for the purpose of link and content based analysis, we believe the first approach works better, because it generates a large number of small and coherent pages, thus avoiding topic drift.

As in existing surfer models, we label the available web pages (after preprocessing) from $\{1, 2, \dots, N\}$. We propose the methodology for fuzzy web surfer models by imitating that of existing surfer models. Similar to the concept of PageRank, we define the concept of FuzzRank, where the objective is to compute, for each given web page, a value which reflects the belief that a web surfer would be on that page. This value is proportional to the belief that the surfer would be on one of its backlinks. Similarly, associated with each link in a page, there is a fuzzy number that indicates the belief that this link would be followed, given that the surfer is on that page. These constitute the fuzzy transition matrix.

Formally, we are interested in computing $\mu(i)$ for each page i , which is the unconditional belief that a surfer would be on i . In other words, given a fuzzy transition matrix P , we want to obtain the eigen fuzzy set $\mu = (\mu(1), \dots, \mu(N))$ [7] of P which satisfies $\mu \circ P = \mu$. Here, the operation $\circ$ is the fuzzy max-min operation, as described in Eq. 6.

FuzzRank, the fuzzy counterpart of PageRank is now defined, as the greatest fuzzy eigen set of the fuzzy transition matrix, the existence of which has been proved in [18]. It is also known that this greatest fuzzy eigen set lies between $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$, where

$$x_k^{(0)} = \min_j \max_i P_{ij} \quad \forall k = 1, 2, \dots, N$$

and

$$x_k^{(1)} = \max_i P_{ik} \quad \forall k = 1, 2, \dots, N.$$

$\mathbf{x}^{(0)}$ is always an eigen fuzzy set, whereas, whenever, $\mathbf{x}^{(1)}$ is an eigen fuzzy set, it is the greatest. Now, it is also known [7] that the greatest eigen fuzzy set is of the form of $\mathbf{x}^{(1)} \circ P^k$, for some positive integer k .

Thus, computing FuzzRank makes use of the power method in max-min algebra, and is similar to computing PageRank, the difference being that one cannot start with an arbitrary vector. One may note that, $\mathbf{x}^{(1)}$ itself equals $\mathbf{1} \circ P$, and hence FuzzRank is of the form $\mathbf{1} \circ P^k$. Thus the initial vector for the power iterations for computing FuzzRank is always $\mathbf{1}$.

So, the task at hand is to obtain accurate values of the elements of the fuzzy transition matrix, because, once that is done, Eq. 6 is all that is required to compute the FuzzRanks of the web pages.

Whenever a page has a single link to another page, it is assumed that there is no fuzziness present there. This is usually the case when an original page say A has been split into pagelets which were originally its named sections and a link from a page, say B had specifically pointed to a named section, say $A\#C$. Then, after splitting, the page B points to just a single page representing $A\#C$. Had the link just pointed to A without referring to the intended section, the splitting would involve some fuzziness as to which section is being referred to. In that case, the membership values of the target of the link from B are non-zero for multiple pages representing the original sections of A . The membership values may be determined by considering similarity of the context around the anchor of the link and the potential target regions. Thus, the fuzzy transition matrix may be obtained.

3.3 Advantages and Limitations

We now discuss some features of the proposed class of fuzzy web surfer models. A list of advantages are listed first, following which we delve into the shortcomings of such a model.

We observe that theoretically, and intuitively, fuzzy web surfer models have the following merits:

1. Capture fuzziness in page contents: page boundaries may not be apparent all the time, especially when a single large page consists of several pagelets. Moreover, noise in web pages also affects the precise identification of the content of interest to the user.
2. Capture fuzziness in links: a page may contain several outlinks but not all of them may be intended for the same purpose. The reason for their presence may be ease of navigation, leading to advertisements, references, or pointing to authoritative resources. Similarly, a link to a particular page may in reality be actually for just one or two sections or pagelets of a page. These kinds of uncertainty may be better modeled by the proposed methodology.
3. Can take into account fuzzy contexts: context sensitive algorithms depend a lot on the modeling assumptions. For example, the context of a query may not be precisely clear, but the system may have a broad idea about it.
4. Robust computations: this is perhaps, the most emphatic reason for choosing fuzzy web surfer models. The computations in max-min algebra are more robust to perturbations as compared to usual addition and multiplication operations. There is an example in [7] that demonstrates the robustness of fuzzy Markov systems in comparison to regular Markov chains. When the entries of the transition matrix are perturbed by small quantities, the effects on the stationary distribution of the regular Markov chains are drastic, whereas, for fuzzy Markov chains, the changes are comparable to the perturbations.
5. Finite convergence: the stationary distribution of fuzzy Markov chains can be computed in finite number of steps, whereas, for regular Markov chains, only an approximation may be found as the convergence may not be achieved in finitely many steps. Existing web surfer models assume that, even though convergence is not attained, the order of the probabilities in the obtained distribution suffices.

We now study the possible limitations of the proposed methodology. It is well known that a Markov chain is ergodic if it is regular. However, in the case of fuzzy Markov chains, no such results are known. So, it is not clear when FuzzRank would actually exist, and even if it does, if it would be independent of the initial state of the process. There is an example in [7] where the rows of the limiting fuzzy transition matrix are distinct, thereby demonstrating the existence of non-ergodic fuzzy Markov chains.

This, however, need not be a limitation as all that it implies is that the final fuzzy distribution of the surfer being on a particular page may not be independent of his initial state. In practice, this may indeed be the case as a surfer starting from one set of pages may, in the long run, behave differently from another one who starts from a different set of web pages. Thus, that the fuzzy Markov chain of web pages being visited is not ergodic may be a blessing in disguise, which may be useful in computing topic sensitive page ranks or for detecting web communities.

4 Experimental Results

The objective of this section is to demonstrate the purpose and usefulness of fuzzy surfer models and to study the properties of FuzzRank. We begin this section with an example that serves as a preview of the experiments performed. The methodology is described as we present the details of the example.

The first step is to choose a (web) graph whose nodes are to be ranked. For this example, we choose a randomly generated directed graph. There are many generative graph models, as mentioned below:

- Erdos-Renyi model [19]: Given the number of vertices, edges are added randomly.
- Power-law models [20]: Here the in-degrees and out-degrees are assumed to arise from a power distribution of the form $y = x^a$. The R-MAT model [21] engulfs these and the Erdos-Renyi model.
- Lognormal models [22,20]: These models have been shown to do better than the power-law models for modeling the web graph and are faster and scalable.

We choose the exponential distribution, which is similar to the lognormal distribution, and is easier to simulate. In our case, to make sure that there are no orphans (that is, nodes with no in-links), for each node, we draw a random number from a (zero truncated) exponential distribution, and choose that many nodes at random from which in-links to the present node are created. Note that the out-degree for some nodes may be zero. A sample random graph with 10 nodes is presented in Fig. 4.

The in-degrees, in-links, out-degrees and out-links for this sample graph are listed in Table 1. In this case, none of the out-degrees are zero.

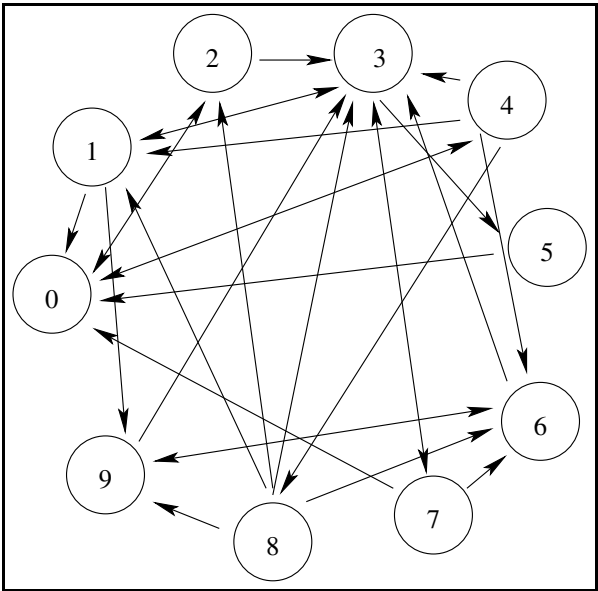


Fig. 4. A sample graph with 10 nodes and 28 links

Table 1. In-Links and Out-Links of the Sample Graph in Fig. 4

Node	Out-Degree	Out-Links	In-Degree	In-Links
0	2	2, 4	5	1, 2, 4, 5, 7
1	3	0, 3, 9	3	3, 4, 8
2	2	0, 3	2	0, 8
3	3	1, 5, 7	7	1, 2, 4, 6, 7, 8, 9
4	5	0, 1, 3, 6, 8	1	0
5	1	0	1	3
6	2	3, 9	4	4, 7, 8, 9
7	3	0, 3, 6	1	3
8	5	1, 2, 3, 6, 9	1	4
9	2	3, 6	3	1, 6, 8

Next, the PageRank vector is computed for this graph, and the vector at different iterations 1, 2, 3, and 31 (when it converged) are shown in Table 2. For the purpose of presenting consistently, all the vectors are max-min normalized, that is, they are linearly transformed so that their minimum is 0 and maximum is 1. The value of d (the probability for a random jump) is set to 0.15 during the computation of PageRank throughout our experiments.

Table 2. PageRank computations for the Sample Graph in Fig. 4

Node	Initial	Iter 1	Iter 2	Iter 3	...	Iter 31
0	1.000	0.915	0.618	0.935	...	0.879
1	1.000	0.225	0.430	0.326	...	0.365
2	1.000	0.211	0.527	0.225	...	0.390
3	1.000	1.000	1.000	1.000	...	1.000
4	1.000	0.126	0.494	0.201	...	0.359
5	1.000	0.056	0.339	0.200	...	0.242
6	1.000	0.436	0.370	0.374	...	0.367
7	1.000	0.056	0.339	0.200	...	0.242
8	1.000	0.000	0.000	0.000	...	0.000
9	1.000	0.352	0.412	0.279	...	0.327

Table 3. FuzzRank computations for the Sample Graph in Fig. 4

Node	Initial	Iter 1	Iter 2	Iter 3
0	1.000	1.000	1.000	1.000
1	1.000	0.166	0.444	0.444
2	1.000	0.375	1.000	1.000
3	1.000	0.375	1.000	1.000
4	1.000	0.375	1.000	1.000
5	1.000	0.166	0.444	0.444
6	1.000	0.375	1.000	1.000
7	1.000	0.166	0.444	0.444
8	1.000	0.000	0.000	0.000
9	1.000	0.375	1.000	1.000

We provide the FuzzRank vectors, too, as they evolve over iterations, in Table 3. Throughout our experiments, we set the fuzzy transition matrix to be the same as that used by the random surfer model. Again, we note that, although no normalization is performed during the actual computation of FuzzRank, they are max-min normalized while reporting them here.

We make the following observations by comparing Tables 2 and 3. We use Kendall discordance to measure the amount of disagreement between the two rank vectors. This discordance is defined as the proportion of discordant pairs among the total of $\frac{k(k-1)}{2}$ pairs. A pair (i, j) is called discordant with respect to two rank vectors, if i is ranked

ahead of j by one, and ranked behind j by the other. In the case of a tie, it is assumed that there is no discernible disagreement.

- The ordering of pages according to FuzzRank is achieved after the first iteration itself, and the actual convergence requires one more iteration. For the case of PageRank, the final ordering is achieved after the twelfth iteration.
- FuzzRank has clumped several nodes together. The number of distinct ranks are 3 for FuzzRank and 9 for PageRank. This indicates that FuzzRank is a more conservative way of ranking (compared to PageRank) where it concludes that the given information in the form of the structure of the graph is insufficient for strictly putting one node ahead of the other, and encourages the use of other factors, such as query relevance, to make this decision.
- The discordance between PageRank and FuzzRank is $\frac{2}{45}$. The disagreement is due to a single node (Node 1), which has a PageRank of 0.365. Had the PageRank value been 0.326 or less (with the PageRank values remaining the same for the remaining nodes), there would have been no discordance between the two rank vectors.

Often, one would be interested in finding the discordance between the top k ranked nodes. Generally, this is to indicate that a discordant pair among the top ranked pages of a ranked list is more significant as compared to the same at the bottom of the list. Comparing top k lists [23] involves obtaining the top k elements of both the lists and looking for discordant pairs among the union of those elements. If c is the number of elements in common to the two top k lists, the total number of elements in the union, n , is $k - c + k - c + c = 2k - c$. Unlike in [23], where it is assumed that the ranks of the nodes outside the top k lists are not known, we shall make use of the available information to compute the actual discordance, thereby avoiding the estimation of discordance suggested in [23]. To handle ties consistently, we shall keep all the tied elements together. So, for the FuzzRank vector, the sets top 1 to top 6 are all the same, consisting of the nodes 0, 2, 3, 4, 6, and 9, whereas the top 7 to top 9 lists have the nodes 1, 5, and 7, in addition to the aforementioned 6 nodes. As we vary k from 1 (we note that, for $k = 1$, the top elements of both lists may be the same, and hence a pair might not be possible at all, in which case the discordance would be defined to be zero) to 10, the number of discordant pairs between PageRank and FuzzRank is 0 for $1 \leq k \leq 5$, 1 for $k = 6$, and 2 for $7 \leq k \leq 10$. Fig. 5 plots the normalized discordance values, whereby, for each $k > 1$, the number of discordant pairs is divided by $\frac{n(n-1)}{2}$.

We now look at the effects of mutating the given graph on the rankings of the nodes. This is important because, often, a link to a page (or a section of the page), might not exist explicitly. The ranking algorithm would need to be robust to gracefully handle noise in links. It has been shown theoretically in [7] that the classical Markov chains may be severely impacted by small changes in the transition matrix, whereas, that is not the case for fuzzy Markov chains.

We perform a simple mutation on the sample graph, such as adding a node or removing a node. Each of the following three graphs is produced as a result of one such mutation.

M1: The link $4 \rightarrow 3$ is removed.

M2: The link $5 \rightarrow 0$ is removed. Note that, this means that there are no out-links from node 5, making it a *rank leak*.

M3: A new link $5 \rightarrow 2$ is added to the graph.

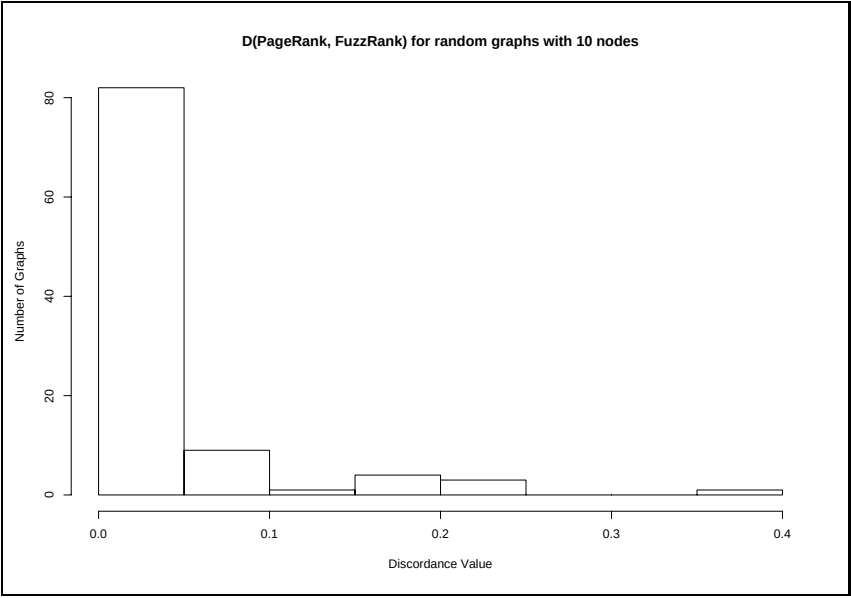


Fig. 6. Discordance values between PageRank and FuzzRank vectors for 100 randomly generated graphs with 10 nodes

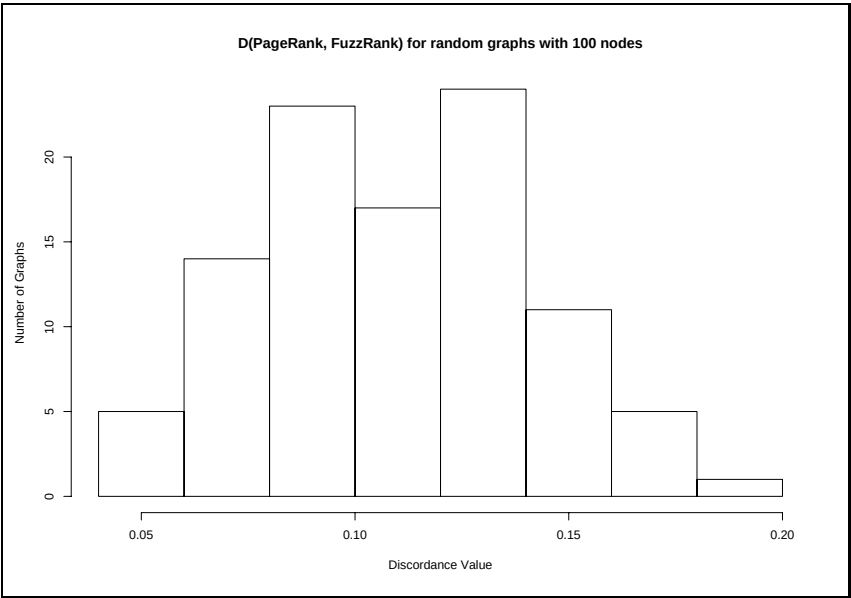


Fig. 7. Discordance values between PageRank and FuzzRank vectors for 100 randomly generated graphs with 100 nodes

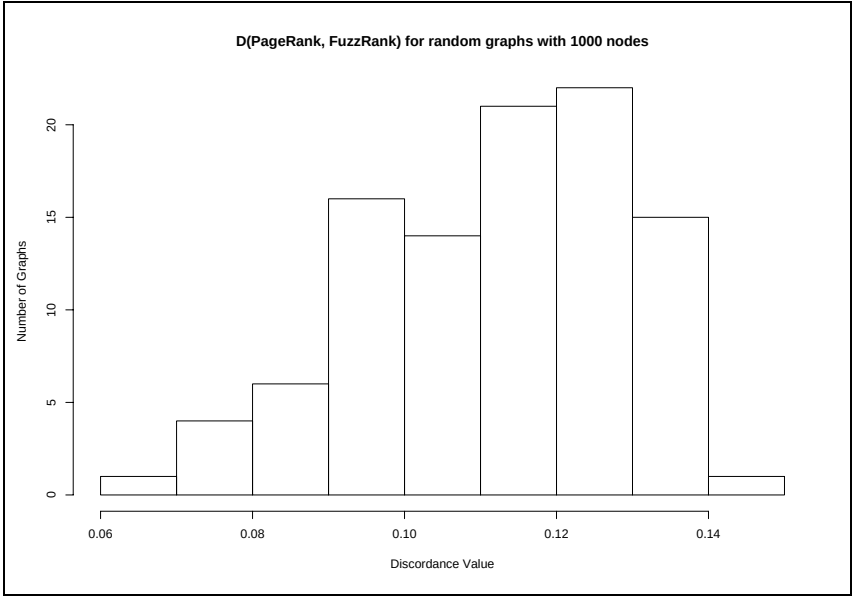


Fig. 8. Discordance values between PageRank and FuzzRank vectors for 100 randomly generated graphs with 1000 nodes

deviate from reality, the resultant ranking may be well away from the ideal one. The robustness of FuzzRank is related to the great number of ties in this case. By not committing itself to a strict ranking, it absorbs the effects of slight changes in the transition matrix.

We now describe the data sets that we have used in our experiments. We have generated 100 graphs each of sizes 10, 100, and 1000, randomly. We chose two real life data sets from Stanford’s WebBase [24] and named them WB1_7440 and WB4_7060, after the host and port numbers from which they are available. The former is a crawl of a part of the berkeley.edu domain, and there are about 140 thousand pages with over 1.6 million links to pages within the same data set. WB4_7060, which is a crawl of a part of the stanford.edu domain, consists of about 40 thousand pages and over 260 thousand links to pages within itself.

Having earlier detailed the methodology of the experiments on an example, the results are now quickly presented. Figs. 6, 7, and 8 present the discordance values between the PageRank and FuzzRank vectors for 100 randomly generated graphs with 10, 100, and 1000 nodes, respectively. As in the earlier example, the discordance values are generally low. Moreover, these values decrease as the number of nodes increases. The corresponding discordance values for the WB1_7440 and WB4_7060 data sets are 0.08 and 0.1, respectively, implying that PageRank and FuzzRank do not disagree much on even large real life data sets.

5 Conclusions and Future Research

The novel theoretical formulation of fuzzy web surfer models by integrating existing works on web surfer models and fuzzy Markov chains is very interesting. The definition of FuzzRank as the fuzzy surfer models counterpart of PageRank, which is based on the random surfer model is both simple and elegant. Experimental results confirm that FuzzRank has very similar ranking properties, and yet is more robust to noise. This robustness is a consequence of FuzzRank to avoid a strict ranking in the absence of strong evidence to that effect. While this may result in a large number of ties if this were the sole criterion for ranking web pages, given that several other factors, like query relevance, would be considered during the ranking process, the ability to consistently rank the pages in the presence of noise is an advantage. The walk-through with an example clearly shows how stable FuzzRank is over PageRank. Future directions for research on this topic involve stability analysis for various other kinds of noise, and obtaining an accurate fuzzy transition matrix based on both the links and the contextual information.

References

1. Zhong, N., Liu, J., Yao, Y.Y. (eds.): Web Intelligence. Springer, Heidelberg (2003)
2. Narayan, B.L., Pal, S.K.: A fuzzy web surfer model. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Compiègne, France, pp. 120–123 (2005)
3. Zhong, N., Liu, J., Yao, Y.: Envisioning intelligent information technologies (iIT) from the stand-point of web intelligence (WI). Communications of the ACM 50(3), 89–94 (2007)
4. Liu, J.: Web intelligence (WI): What makes wisdom web? In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 1596–1601 (2003)
5. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
6. Hoel, P.G., Port, S.C., Stone, C.J.: Introduction to Stochastic Processes. Waveland Press, Incorporated (1972)
7. Avrachenkov, K., Sanchez, E.: Fuzzy markov chains and decision-making. Fuzzy Optimization and Decision Making 1(2), 143–159 (2002)
8. Brin, S., Page, L.: The anatomy of a large-scale hypertextual search engine. Technical report, Stanford University (1998)
9. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
10. Richardson, M., Domingos, P.: The intelligent surfer: Probabilistic combination of link and content information in pagerank. In: Advances in Neural Information Processing Systems, vol. 14, MIT Press, Cambridge (2002)
11. Haveliwala, T.H.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge and Data Engineering 15(4), 784–796 (2003)
12. Pal, S.K., Narayan, B.L., Dutta, S.: A web surfer model incorporating topic continuity. IEEE Transactions on Knowledge and Data Engineering 17(5), 726–729 (2005)
13. Diligenti, M., Gori, M., Maggini, M.: A unified probabilistic framework for web page scoring systems. IEEE Transactions on Knowledge and Data Engineering 16(1), 4–16 (2004)
14. Narayan, B.L., Pal, S.K.: Detecting sequences and cycles of web pages. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Compiègne, France, pp. 80–86 (2005)

15. Ramaswamy, L., Iyengar, A., Liu, L., Douglass, F.: Automatic fragment detection in dynamic web pages and its impact on caching. *IEEE Transactions on Knowledge and Data Engineering* 17(5), 859–874 (2005)
16. Chakrabarti, S., Joshi, M.M., Tawde, V.B.: Enhanced topic distillation using text, markup tags and hyperlinks. In: *Research and Development in Information Retrieval* (2001)
17. Narayan, B.L., Murthy, C.A., Pal, S.K.: Topic continuity for web document categorization and ranking. In: *Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, Halifax, Canada pp. 310–315 (2003)
18. Sanchez, E.: Eigen fuzzy sets and fuzzy relations. *Journal of Mathematical Analysis and Applications* 81, 399–421 (1981)
19. Erdos, P., Renyi, A.: On random graphs i. *Publicationes Mathematicae*, 290–297 (1959)
20. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1(2), 226–251 (2004)
21. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-mat: A recursive model for graph mining. In: *Proceedings of SIAM Data Mining*, Florida, USA (2004)
22. Bi, Z., Faloutsos, C., Korn, F.: The DGX distribution for mining massive, skewed data. In: *Proceedings of KDD*. (2001)
23. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. *Siam Journal of Discrete Mathematics* 17(1), 134–160 (2003)
24. Hirai, J., Raghavan, S., Paepcke, A., Garcia-Molina, H.: Webbase: A repository of web pages. In: *9th International World Wide Web Conference*, Amsterdam (2000)
25. Buckley, J.J.: *Fuzzy probabilities and fuzzy sets for web planning*. Springer, Heidelberg (2004)
26. Zhong, N.: Impending Brain Informatics (BI) Research from Web Intelligence (WI) Perspective. *International Journal of Information Technology and Decision Making* 5(4), 713–727 (2006)
27. Zhong, N.J.L. (eds.): *Intelligent technologies for information analysis*. Springer, Heidelberg (2004)
28. Zhong, N., Liu, J., Yao, Y.Y.: In search of wisdom web. *IEEE Computer* 35(11), 27–31 (2002)

Intuitive Display for Search Engines Toward Fast Detection of Peculiar WWW Pages

Einoshin Suzuki¹, Shin Ando², Masayuki Hirose², and Masatoshi Jumi²

1. Department of Informatics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan
suzuki@i.kyushu-u.ac.jp
2. Electrical and Computer Engineering, Yokohama National University, Japan
{ando,mayusaki,jumi}@slab.dnj.ynu.ac.jp

Abstract. In this paper, we discuss utilization of data mining techniques in realizing intuitive display for search engines toward fast detection of peculiar WWW pages. A search engine can be regarded as a telescope for WWW because it serves as a means to find relevant information in the huge cyberspace. Most of the current display styles of search engines are, however, just text-based rankings thus are far from intuitive. They are also inadequate for certain activities such as browsing for unexpected Web pages. Detection of peculiar WWW pages is expected to lead to making profits and to stimulating our creativity. Our visualization method DPITT (Detecting Peculiar WWW pages from Image, Topic and Term) is based on several data mining techniques and outperforms that of Google in a problem setting which largely favors Google. In this paper, we mainly present our DPITT and we introduce our latest system GEMVIG.

1 Introduction

In the past few years, we have witnessed the rapid growth of WWW both in size and in importance. In Japan, WWW comes to have a considerable influence on people's daily activities such as shopping, dining, and traveling to name but a few and the situation seems to be similar in other advanced countries. A directory-type category such as the one provided at Yahoo! is considered to have limited usefulness thus search engines, despite of several deficiencies, have become popular. The fact that Google, the de facto standard search engine, is also used as a verb by many people supports this tendency.

We argue, however, that most of the display methods of search engines are weak. Showing a ranking list of URLs in text seems to be user-unfriendly. Our opinion is justified by a cliché in e-commerce: "user sees only the first page". A promising exception is Vivisimo (<http://vivisimo.com/>), a search engine which employs clustering to organize search results. Data mining, as it includes Web mining [10] and text mining, possesses various techniques for machine learning and visualization thus is expected to contribute to intuitive display of search results.

Based on the success of our multi-viewpoint and multi-granularity visualization in a medical domain [15], we have proposed various visualization methods based on data mining techniques in various domains. They include visual detection methods for malicious accesses [1,7] and a visualization method for multiple clustering results [3]. For the problem of displaying the results of a search engine, we have chosen the problem of detecting peculiar WWW pages and proposed DPITT [6] and GEMVIG [9]. It should be noted that for this problem we initially proposed a text-based method [5] but we have observed superior performance of visualization methods. In this paper, we mainly present DPITT and we also express our expectations for “Web Intelligence meets Brain Informatics” in our concluding remarks.

2 Detection of Peculiar WWW Pages

2.1 Motivations

Peculiarity or exception has attracted interests of researchers in data mining as a promising characteristics of usefulness [14]. For instance, in WWW, a peculiar WWW page can be a precious clue for making profits and for stimulating our creativity. For instance, early detection of a company which offers unique service on WWW is expected to lead to a profitable business relation or stock investment.

Yahoo! and Google represent two popular search engines and the former is also well-known for its predefined categories in the form of directories. In discovery of peculiar WWW pages, we believe that the interests of the user rarely fall into a few predefined categories thus we believe that the user is likely to utilize a search engine. Since most search engines display the title and a few lines which include words in the query for each WWW page, we think that the display style of a typical search engine poses a considerable workload to the user.

It is a consensus among data miners that Information Visualization [2,4] can be an effective measure for circumventing these difficulties. Discovered patterns should be inspected by the user as long as the user can afford the process. Effective visualization methods have been proposed in Web mining, e.g. Vivisimo (<http://vivisimo.com/>) and [12,16]. In discovery of peculiar WWW pages, however, we think it necessary to visualize a WWW page from various viewpoints at summary levels in terms of the user’s interest which is expressed in the form of a query. We believe that there exists no visualization method in Web mining that satisfactorily fulfills these requirements and is widely used.

2.2 Problem Description

We assume that a search engine accepts a set W_{key} of keywords (i.e. query) as input and outputs the number of WWW pages each of which contains words in W_{key} and a ranked list of links to the WWW pages¹. We represent the set of all

¹ We are aware that a typical search engine employs a sophisticated algorithm and may violate our assumption. However, we have experienced that such a violation can be neglected when the cardinality of W_{key} is small.

WWW pages considered by a search engine and its subset which consists of the top L WWW pages by $D(W_{\text{key}})$ and $D_L(W_{\text{key}})$, respectively.

The input to the problem is W_{key} and $D_L(W_{\text{key}})$ while its output is a set $D'(W_{\text{key}})$ of peculiar WWW pages. Here $D'(W_{\text{key}}) \subseteq D(W_{\text{key}})$ and whether a WWW page is peculiar is decided by the user. The evaluation indices are recall and precision of $D'(W_{\text{key}})$ in terms of $D_L(W_{\text{key}})$ as well as the computation time.

2.3 Probabilistic Model for Topics

We have proposed a visualization method called PrototypeLines, which has been proved to be successful with a medical application using mainly blood test data [15]. Its success taught us that dimension reduction with a probabilistic clustering method is highly effective. Likewise, an intuitive display which is based on an information criterion and which uses hue as a single medium is useful since it requires neither excessive training nor laborious parameter tuning. Since a typical WWW page contains images and documents unlike blood test data, we decided to develop a novel method.

PLSI [8], which infers semantic relevance, has been shown to be effective for document data with the bag-of-word model [13]. PLSI employs a set of latent classes $Z = \{z_1, z_2, \dots, z_K\}$ and assumes that a document d is generated probabilistically based on a generative model. For a word w ,

$$P(d, w) = P(d)P(w|d), \quad \text{where} \quad (1)$$

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \quad (2)$$

PLSI obtains maximum likelihood estimates for $P(z)$, $P(d|z)$, and $P(w|z)$. Various experiments including those in [8] report the effectiveness of PLSI for document data.

A latent class z in PLSI can be considered as representing a topic in the given set of document. Alternatively stated, we can detect K topics in the document set with PLSI. We believe that this kind of information is useful in detecting peculiar WWW pages.

2.4 Google Frequency for Detecting Peculiar Words

In [5], we have shown that peculiar WWW pages can be detected by a set of peculiar words contained in the page relatively accurately. We have experienced that the degree of generality cannot be inferred by the term frequency [13] estimated from a set $D_L(W_{\text{key}})$ of document obtained with a search engine and proposed an evaluation index Google Frequency (GF) which exploits statistics returned by a search engine [5].

We have defined a GF $gf_{\text{WWW}}(W_{\text{key}}, w_i)$ of a word w_i in terms a word set W_{key} of the query as follows

$$gf_{\text{WWW}}(W_{\text{key}}, w_i) = \frac{|D(W_{\text{key}} \cup \{w_i\})|}{\min(|D(W_{\text{key}})|, |D(\{w_i\})|)} \quad (3)$$

where “ $\min(a, b)$ ” returns the smaller term of a and b , and $|D|$ represents the cardinality of a set D . Intuitively, our GF represents generality of w_i in the set $D(W_{\text{key}})$ of documents returned by the search engine for W_{key} .

3 Multi-viewpoint Visualization System DPITT

3.1 Visualization Scheme

In this section, we propose our visualization method DPITT. DPITT employs PLSI and GF, which we explained in the previous section, for detecting topics and peculiar words.

As we explained previously, we believe that various information should be visualized to detect peculiar WWW pages effectively. DPITT is designed to visualize the page title, a representative image, the topic distribution, the generality distribution of words, a set of frequently used words, and a set of words specific to the WWW page.

The result of visualization is output as an html file which contains several pages, each of which contains visualization results of 25 WWW pages in the form of a 5×5 -matrix. Figure 1 shows a visualization result of a WWW page, where each of the title and the image can serve as a link to the original WWW page. The circled digits in the Figure will be explained in the next section.

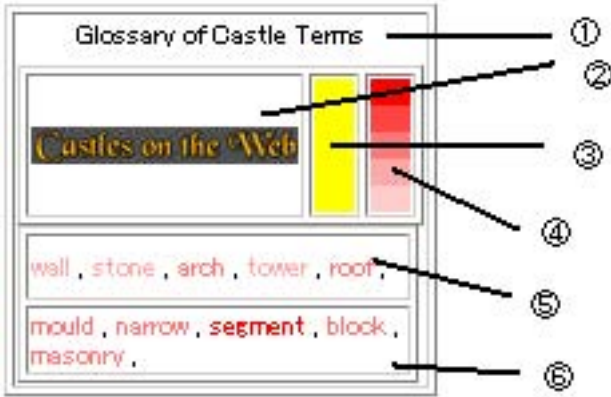


Fig. 1. Example of a visualization result of a WWW page

3.2 Details of DPITT

We use a window of size 150 pixel $\times$ 150 pixel for visualizing a WWW page d_j as shown in Figure 1. On the top row of the window is shown the title of the WWW page up to 40 letters (①).

In the second row, toward left, we show an image which is likely to be representative of the WWW page (②). The image is selected according to the following procedure and is transformed into 100 pixel $\times$ 100 pixel in size.

1. The largest image in the directory of the URL of the WWW page.
2. If 1. does not exist, the first-appearance image in the WWW page.
3. If there is no image in the WWW page, the capture image of the WWW page.

In the second row, in the middle, we show the topic distribution of the WWW page (③). In PLSI, we can obtain the conditional probability $P(z_c|d_j)$ of a topic z_c given a WWW page d_j . Since our objective lies in the detection of peculiar WWW pages, we settle $K = 2$ to detect a dominant topic z_1 and the remaining topic z_2 . The topic distribution which consists of $P(z_1|d_j)$ and $P(z_2|d_j)$ is represented by a bar colored in red and yellow. The lengths of the red and yellow parts are proportional to $P(z_1|d_j)$ and $P(z_2|d_j)$, respectively.

In the second row, toward right, we show the generality distribution of words (④) for the WWW page d_j . Let $N(d, w)$ be the occurrence of w in d . Typically only words that satisfy $N(w_i, d_j) \geq 2$ are used, though words that satisfy $N(w_i, d_j) = 1$ are used if $\forall i, N(w_i, d_j) \leq 1$. As the generality index of a word w_i , we use $gf_{WWW}(W_{key}, w_i)$. Figure 2 (left) shows our color assignment for GF. As we see in the Figure, the darker the color is the more peculiar the word is thus a WWW page with a large portion of dark colors is expected to be peculiar.

A frequently occurring word in a WWW page, i.e. a word with a high term frequency, is considered to represent an aspect of the WWW page. In the third row, we show the most frequently occurring words in the WWW page up to five (⑤). Each word w_i is colored according to its $gf_{WWW}(W_{key}, w_i)$ using the color assignment shown in Figure 2 (right).



Fig. 2. Color assignment to GF (left) and a word w_i (right) distributions

A word specific to a WWW page, i.e. a word which appears only in the WWW page is expected to represent an aspect of the unique content of the WWW page. In the fourth row, we show the most frequently occurring words that appear only in the WWW page d_j , i.e. $|D_L(W_{key} \cup \{w_i\})| = 1$ (⑥) up to five. The color of a word is decided as shown in Figure 2 (right).

4 Experimental Evaluation of DPITT

4.1 Experimental Conditions

We employed ten sets of WWW pages in the experiments. Each $D_L(W_{\text{key}})$ of the sets was collected by using a set W_{key} of keywords as a query to Google and storing the top 100 results ($L = 100$). Only WWW pages in the form of HTML were used in the experiments since other types of WWW pages are few in number and can be analyzed in the same way if we provide an appropriate interface. Table 1 shows the list of keyword sets, where a “+” represents a logical conjunction.

Table 1. List of keywords sets, where “Number” represents the number of WWW pages

Keywords	Number	Keywords	Number
castle	97	christmas	97
milk	100	mp3+player	95
rice	99	robot	99
tokyo	93	toshiba	93
world+cup	98	yokozuna	88

Figure 3 shows a display results of Google with “castle” as the query. In our environment, Google displays 7 and 12 WWW pages in average on a 21-inch screen and on a sheet of A4 paper, respectively. Figure 4 shows a display result of DPITT for this problem as the keyword set. As we see from the Figure and as we have explained, DPITT outputs 25 WWW pages on an A4 sheet thus displays more than twice as many WWW pages as Google.

We have performed a series of cognitive experiments with six subject users with the data set as follows.

1. A subject user reads explanation of the experiments.
2. The subject user writes down his/her criteria for classifying WWW pages into peculiar and others based on a list of frequent words in the WWW pages².
3. The subject user classifies WWW pages into peculiar and others from the output of DPITT.
4. The subject user classifies WWW pages into peculiar and others from the original WWW pages.
5. The subject user classifies WWW pages into peculiar and others from the output of Google.

² Because the subject user has not seen the search result, his/her criteria are not influenced by the result.

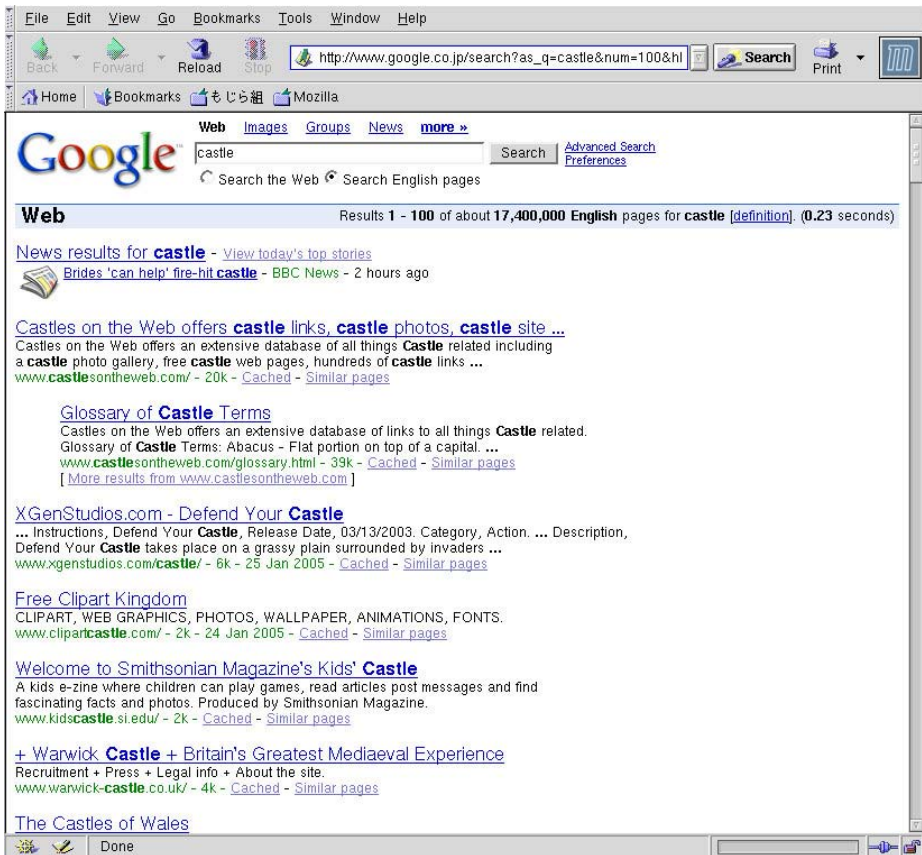


Fig. 3. Display result of searching “castle” with Google

The precision and recall are calculated by regarding the set of WWW pages judged peculiar in 4. as the target class. Consequently this procedure for the experiments largely favors Google because a subject user knows the correct answer in 5. Though we have designed DPITT for detecting peculiar WWW pages, DPITT was anticipated to be outperformed by Google since a subject user performs the classification with DPITT before 4. i.e. before knowing the correct answer.

4.2 Experimental Results

We show the experimental results in Figure 5. As we see from the Figure, DPITT outperforms or equals to Google in all data sets and in all indices. These results are surprising because the experimental procedure largely favors Google. Several subject users commented that the topic distributions were relatively accurate and so were the images. For instance, for the problem “castle”, the yellow



Fig. 4. Display result of DPITT with “castle” as the keyword set

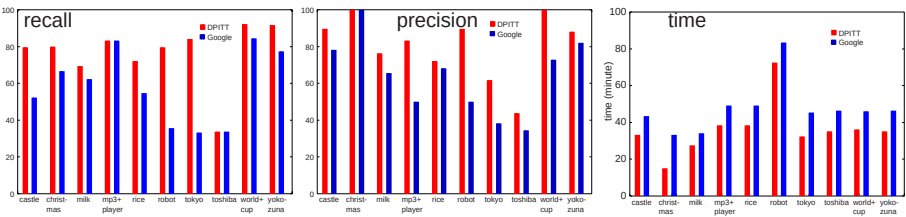


Fig. 5. Experimental results

and red topics often represent WWW pages related with a fortified building and other WWW pages, respectively. This characteristics resulted in fast detection of WWW pages not related with a fortified building. We consider that the topic distribution and the image information were especially effective in reducing time.

Since the subject users and the authors are academic members of a university, we have not succeeded in finding a WWW page which would certainly make profits in business or investment. This is also because we have employed the top 100 WWW pages of Google in our experiments only. However, several WWW pages stimulated our intellectual interests such as our results with “robots” as the query and we believe that they mean different things to business people in the robotics industry. We believe that the usefulness of DPITT largely depends on the problem setting and are planning to engage several kinds of domain experts in our future experiments.

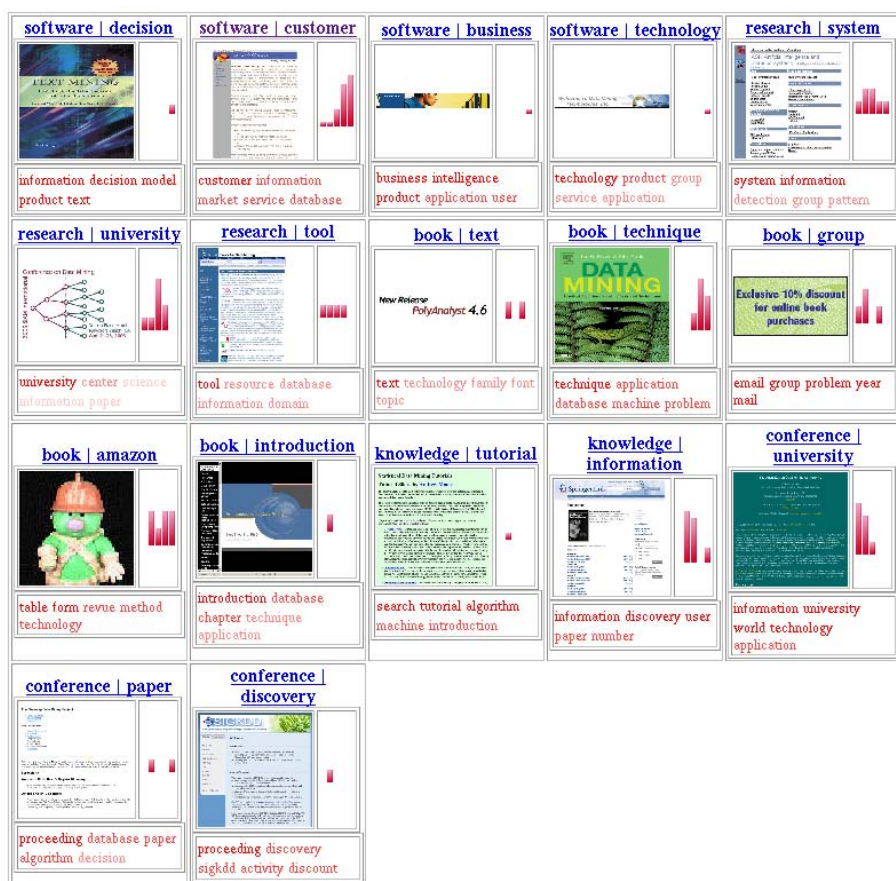


Fig. 6. Display result of WWW pages for query “data+mining”

5 GEMVIG

A drawback of DPITT is that the number of WWW pages that can be displayed on a screen is limited. To circumvent this problem, we have developed another display method GEMVIG which exploits clustering of WWW pages [9]. We will explain the details of GEMVIG in the future and just describe a comprehensible example in this paper.

Figure 6 shows a display result of WWW pages for a query “data+mining” where each of the 17 elements represents a cluster of WWW pages. Clusters are organized in a two-level tree and their names appear on the top of each element. For instance, the leftmost four elements in the first row all belong to a first-level cluster “software” and each of them belongs to its corresponding second-level cluster described after “software”.

The user can see the members of each cluster by clicking the corresponding element. Figure 7 shows the WWW pages which consist the cluster “research | system” in Figure 6.

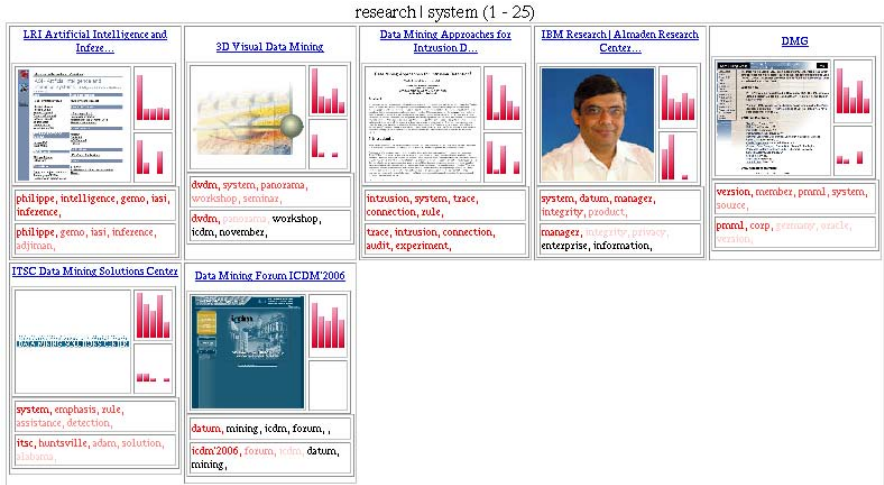


Fig. 7. Display result of WWW pages for cluster “research, system” in Figure 6

6 Concluding Remarks

In this paper, we have mainly presented our DPITT as a display method for a search engine for detecting peculiar WWW pages rapidly. Experimental results using the display method of Google for comparison justify our motivation and validate the effectiveness of DPITT. We have also explained our new display system GEMVIG, which allows many more WWW pages to be displayed on one screen with its use of WWW-page clustering.

Brain Informatics is a new interdisciplinary field to study human information processing mechanism systematically from both macro and micro points of view by cooperatively using experimental cognitive neuroscience and Web Intelligence (WI) centric advanced information technology (<http://wi-consortium.org/wimbi06/WImBI2006.htm>) [11,17,18,19,20,21,22,23,24]. We believe that research on WI including ours can benefit from brain informatics. It is well-known that current information visualization research is weak in evaluating various methods [4]. We hope research in Brain Informatics can provide answers to this problem as the research field seems to be most appropriate for exploring various challenging issues related with human cognition.

Acknowledgments

This work was partially supported by the grants-in-aid for scientific research on fundamental research (B) 16300042 and 18300047 from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. Ando, S., Suzuki, E.: Distributed Multi-objective GA for Generating Comprehensive Pareto Front in Deceptive Multi-Objective Problems. In: Proc. 2006 IEEE Congress on Evolutionary Computation (IEEE CEC), pp. 5718–5725 (2006)
2. Card, S.K., Makinlay, J.D., Shneiderman, B. (eds.): Readings in Information Visualization. Morgan Kaufmann, San Francisco (1999)
3. Durand, N., Cremilleux, B., Suzuki, E.: Visualizing Transactional Data with Multiple Clusterings for Knowledge Discovery. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 47–57. Springer, Heidelberg (2006)
4. Fayyad, U., Grinstein, G.G., Wierse, A. (eds.): Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, San Francisco (2002)
5. Hirose, M., Suzuki, E.: Using WWW-Distribution of Words in Detecting Peculiar Web Pages. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 355–362. Springer, Heidelberg (2004)
6. Hirose, M., Suzuki, E.: DPITT: Multi-viewpoint Visualization System for Detecting Unexpected WWW Pages Rapidly. In: 2006 IEEE International Conference on Granular Computing (IEEE-GrC 2006), pp. 538–541 (2006)
7. Hirose, N., Suzuki, E.: Engineering Web Log for Detecting Malicious Sessions to a Web Site by Visual Inspection. WSEAS Transactions on Computers 10(4), 1249–1258 (2005)
8. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR), pp. 50–57 (1999)
9. Jumi, M.: Research on Multi-viewpoint and Multi-granularity Visualization of a Set of Searched Web Pages Based on Hierarchical Clustering. Master of Engineering Dissertation, Department of Electrical and Computer Engineering, Division of Advanced Physics, Electrical and Computer Engineering, Graduate School of Engineering, Yokohama National University, Japan (in Japanese) (2006)

10. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. *ACM SIGKDD Exploration* 2, 1–15 (2000)
11. Liu, J.: Web Intelligence (WI): What Makes Wisdom Web? In: *Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1596–1601 (2003)
12. Renteria, J.C., Lodha, S.K.: WebVis: a Hierarchical Web Homepage Visualizer. In: *Proc. SPIE*, vol. 3960, pp. 50–61 (2000)
13. Salton, G., McGill, M.J. (eds.): *Introduction to Modern Information Retrieval*. McGraw-Hill, New York (1983)
14. Suzuki, E.: Undirected Discovery of Interesting Exception Rules. *International Journal of Pattern Recognition and Artificial Intelligence* 16(8), 1065–1086 (2002)
15. Suzuki, E., Watanabe, T., Yokoi, H., Takabayashi, K.: Detecting Interesting Exceptions from Medical Test Data with Visual Summarization. In: *Proc. Third IEEE International Conference on Data Mining (ICDM)*, pp. 315–322 (2003)
16. Toyoda, M., Kitsuregawa, M.: Extracting Evolution of Web Communities from a Series of Web Archives. In: *Proc. Fourteenth ACM Conference on Hypertext and Hypermedia (Hypertext)*, pp. 28–37 (2003)
17. Yao, Y.Y., Zhong, N., Liu, J., Ohsuga, S.: Web Intelligence (WI): Research Challenges and Trends in the New Information Age. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) *WI 2001. LNCS (LNAI)*, vol. 2198, pp. 1–17. Springer, Heidelberg (2001)
18. Zhong, N.: Impending Brain Informatics (BI) Research from Web Intelligence (WI) Perspective. *International Journal of Information Technology and Decision Making* 5(4), 713–727 (2006)
19. Zhong, N., Liu, J. (eds.): *Intelligent Technologies for Information Analysis*. Springer, Heidelberg (2004)
20. Zhong, N., Liu, J., Yao, Y.Y.: In Search of the Wisdom Web. *IEEE Computer* 35(11), 27–31 (2002)
21. Zhong, N., Liu, J., Yao, Y.Y. (eds.): *Web Intelligence*. Springer, Heidelberg (2003)
22. Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies (iT) from the Stand-Point of Web Intelligence (WI). *Communications of the ACM* 50(3), 89–94 (2007)
23. Zhong, N., Liu, J., Yao, Y.Y., Ohsuga, S.: Web Intelligence (WI). In: *Proc. the 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC)*, pp. 469–470 (2000)
24. Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.): *WI 2001. LNCS (LNAI)*, vol. 2198. Springer, Heidelberg (2001)

GridMiner: An Advanced Grid-Based Support for Brain Informatics Data Mining Tasks

Peter Brezany¹, Ivan Janciak¹, Jarmila Brezanyova², and A Min Tjoa²

¹ Institute of Scientific Computing, University of Vienna, Nordbergstrasse 15/C/3
A-1090 Vienna, Austria

brezany@par.univie.ac.at, janciak@par.univie.ac.at

² Institute of Software Technology and Interactive Systems, Vienna University of Technology
Favoritenstrasse 9-11/E188, A-1040 Vienna, Austria
tjoa@ifs.tuwien.ac.at

Abstract. Brain Informatics is a new interdisciplinary field of science, which studies the mechanisms of human information processing and, in some cases, also mechanisms causing the development of cognitive and other mental disturbances. Like other e-Science disciplines, it is accompanied with collection of large distributed datasets, which have to be efficiently managed, processed and analyzed. The Grid is an appropriate platform allowing solution of such ambitious tasks. This paper presents a Grid-based infrastructure called GridMiner providing the kernel functionality, like workflow management, visualization, data preprocessing, filtering, transformation and integration, data mining and data warehousing services, etc., supporting analytical tasks required by the Brain Informatics research and applications.

1 Introduction

In the past ten years, the Internet and World Wide Web significantly affected scientific research processes. However, a much more powerful infrastructure is needed for that many new areas of science that are becoming increasingly reliant on new ways of collaborative, multidisciplinary working. They typically require very large scale computing resources and access to very large data collections, often well in the multi-terabyte range, generated, e.g., by scientific simulations and experiments performed on a new generation of high-resolution scientific instruments, and by medical techniques like fMRI (functional magnetic resonance imaging) and EEG (electroencephalogram). Gaining insights and extracting latent knowledge from that large volumes of data, which can be heterogeneous and geographically distributed, requires I/O hardware and software that can move data rapidly across networks and to and from arrays of storage devices, intelligent data reduction, preprocessing and visualization, high performance data mining methods, and software mechanisms for efficient coordination of data analysis processes. The communities of users that need to access and analyze this data are often large a spread across different research sites. These requirements resulted in complex and stringent performance demands that, until recently, were not satisfied by any existing computational and data management infrastructure.

The novel infrastructure to enable this science revolution is generally referred to as the *Grid*; it allows flexible, secure, and coordinated resource sharing among dynamic

collections of individuals and institutions. Since the 1990's (presentation and discussion of the very first Grid vision), Grid technology has traversed different phases or generations – from *Computational Grids* (concerning, e.g., job scheduling, system information services, life cycle management) to *Data Grids* (addressing issues such as distributed data access, efficient file transfer, metadata management, data replication, data integration), and recently to *Semantic and Knowledge-Oriented Grids* – these terms denote several development directions, including Knowledge Discovery (very often used as a synonymum to Data Mining), Semantic, and Knowledge Grids.

The Grid is a driving force of *e-Science analytics*, a dynamic research field, that includes rigorous and sophisticated scientific methods of data preprocessing, integration, analysis, and visualization. There has been already several Grid research activities addressing some of these methods, e.g. data mining, data integration, etc. The *Grid-Miner* project (<http://www.gridminer.org>) at the University of Vienna and the Vienna University of Technology aims, as the first Grid research effort, to cover all aspects of the knowledge discovery process and integrate them into an advanced service-oriented Grid application. The innovative architecture provides (1) a robust and reliable high performance data mining and On-Line Analytical Processing (OLAP) environment, (2) seamless access to intermediate data and results of the discovery process for further reuse in a standardized way, and (3) a persistent workspace for continuous and evolving data mining tasks supported via a flexible GUI supporting both desktop and mobile devices. It hides the complexity of the Grid but is still offering possibilities to interfere during the execution phase, control the task execution and visualize results. The interactive cooperation of a set of optimized services – for data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation and finally its storage – within the GridMiner architecture is the key to productive knowledge discovery in large datasets. The aim is to provide a generic framework – software toolkit, application development methodology, and a set of appropriate use-cases – which can be easily extended for domain specific data exploration applications.

This paper shows how the GridMiner can support data analysis tasks and decision processes in the *Brain Informatics* domain, a new interdisciplinary field of science, and its rest is organized as follows. Section 2 characterizes Brain Informatics and its tasks. Section 3 presents the KDD process and its phases from the point of view of the GridMiner system. Basic building blocks of the system and their roles in the architecture are discussed in Section 4. Here, we also introduce an illustrative example, how Brain Informatics data collection and analysis workflows can be constructed and executed in the GridMiner environment. Section 5 discusses how to build a Wisdom Grid, a significant contribution to the Web Intelligence research, on top of the GridMiner system. We briefly describe related work in Section 6 and conclude in Section 7.

2 Brain Informatics

Due to N. Zhong's specification [21], Brain Informatics (BI) studies the mechanisms of human information processing from both the macro and micro viewpoint by combining experimental cognitive neuroscience with advanced information technology [25]. It allows new understanding and discovery of human intelligence models. Some of its

branches also aim to find and analyze complex etiological and neurobiological mechanisms in the central neural system that regulates psychological functions and the development of cognitive and other mental disturbances [11]. New instrumentation, like fMRI and EEG and advanced information technology significantly support the BI research.

In the Brain Informatics Lab of the School of Biomedical Engineering of the Beijing University of Technology, in the current research phase, the following four-step data collection and analysis workflow is applied [23]:

1. Multiple brain data such as fMRI images and EEG brain waves extracted from human multi-perception mechanism involved in a particular task, e.g., visual and auditory calculation experiments are collected into databases.
2. To be able to apply data mining methods, the data sets created in Step 1 are filtered and transformed into a structured form.
3. Data mining methods, e.g. peculiarity oriented mining for multi-aspect analysis [22], are applied to the selected data.
4. The mined results are evaluated and interpreted. If necessary, the workflow returns to Step 2.

The Swedish Human Brain Informatics (HUBIN) project reports that their neurological databases have been analyzed by conventional and new new statistical and data mining methods; the technical report [11] provides an overview of the methods applied and discusses the results achieved.

The above discussed tasks were implemented by specialized programs developed from scratch. Next sections discuss how they can be supported by our GridMiner framework.

3 Knowledge Discovery Process in GridMiner

Knowledge discovery in databases (KDD) can be defined as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [7]. KDD is a highly interactive process and to achieve appealing results the user must permanently have the possibility to influence this process by applying different algorithms or adjusting their parameters. Traditionally it consists of several phases which are depicted in Fig. 1 and discussed in this section. The phases are composed by the graphical user interface to an appropriate graphical workflow representation, which is then processed and interactively executed by a specific workflow management subsystem.

Data Cleaning and Integration Phase. According to [18], data preprocessing is usually the most challenging and time consuming step in the whole KDD process. The aim of this task is to improve data quality which has significant impact on the final model quality and, therefore, on the success of whole KDD process. Data in operational databases are typically not clean. This means that those databases contain errors, due to wrong inputs from users or application failures. Besides, this data may be incomplete, and essential information may not be available for some attributes. Hence, for data mining

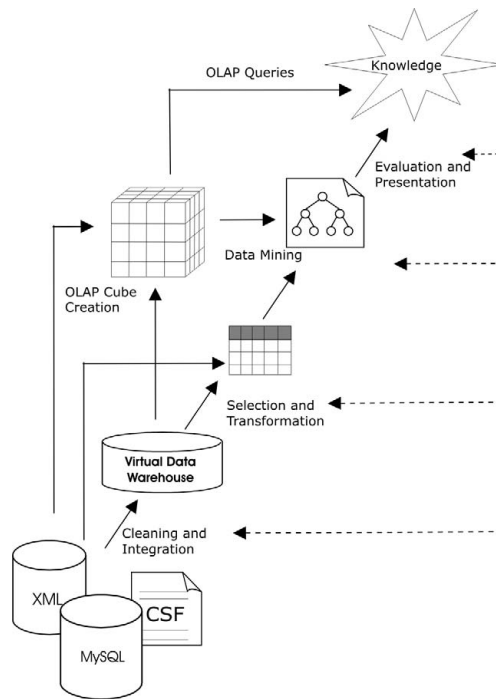


Fig. 1. Knowledge discovery in GridMiner

and other analysis tasks, it is necessary to clean data and to integrate the data into one common dataset. Because of large size of typical Grid data sources, building traditional data warehouses is impractical. Therefore, we work with a “Virtual Data Warehouse” – we clean and integrate the data, currently, from relational and XML databases and Comma Separated Value Files (CSF), dynamically on-the-fly at the time the query is issued [19]. The integrated data is provided to the applications in the WebRowSet¹ format (XML files or streams).

Data Selection and Transformation Phase. The size of the input data for modeling can enhance accuracy of the predicted model as well as it has significant impact on the time of model building. Data selection step allows to choose an appropriate subset of the whole dataset which can be used as a training set to build a model as well as a test set for the model evaluation.

Data transformation is a step which brings data in a different form, which is more appropriate for data mining algorithms. If for example the quality of the results of the mining step is very low, it is possible, that the data are not in a suitable form. Transformation is a simple method, which can increase the quality of data mining tremendously.

Data Mining Phase. This phase deals with the selecting, applying and tuning a modeling technique on the prepared data set. It involves the application and parameterization

¹ <http://java.sun.com>

of a concrete data mining algorithm, to search for structures and patterns within the dataset, like classification, clustering, association, characterization, or comparison algorithms.

OLAP Phase. OLAP is used for interactive analysis of data stored in datasets. It is based on the Data Cube concept. Data Cube is a data structure, which is used as a multidimensional view of data repositories. In GridMiner, we apply specific data mining techniques to OLAP cubes; this technology is called On-Line Analytical Mining [12].

Model Evaluation and Presentation Phase. The results of the data mining phase have to be evaluated by application of distinct measures. Data mining algorithms deliver a set of data patterns but not all of them are interesting to the user or represent knowledge, so this phase discriminates relevant from irrelevant patterns. The main task of the presentation step is to give to a user all discovered information in an appropriate form. There are techniques for representing data mining models in, for example, tables, charts or more sophisticated visualization mechanisms.

4 GridMiner Architecture - Components and Their Roles in KDD

In this section, we discuss the current development status of the GridMiner system, its components and implementation. The GridMiner's component architecture is depicted in Fig. 3. Following parts provide its description.

4.1 Knowledge Base (KB)

KB is used to store and share all documents needed by the other components in KDD process. KB is implemented as an XML database (XMLDB) that manages the life cycle of a document from its creation till reusing. KB is used as a central registry of all information and knowledge resources in the user's projects.

Information Resources

The KDD process includes the information flows between its phases – their inputs and outputs have to be specified in a unified way. XML based documents have already become a standard for storing and sharing information in Web environments because they are very well understandable, can be easily maintained, distributed and accessed by their URLs². In our approach, five types of XML documents are used to store such information, namely XSLT, PMML, Workflow (DSCL) Document, Mapping Schema, and Perform Document. The set of these documents can be treated as a distributed knowledge base and, as depicted in Fig. 2, each of them is used by different Web applications or Grid services for exchanging information.

Role of XSLT. XSLT³ is a transformation document which, when applied to the WebRowSet, allows to select or transform only a subset of the data without changing the

² URL - Uniform Resource Locator.

³ <http://www.w3.org/TR/xslt>

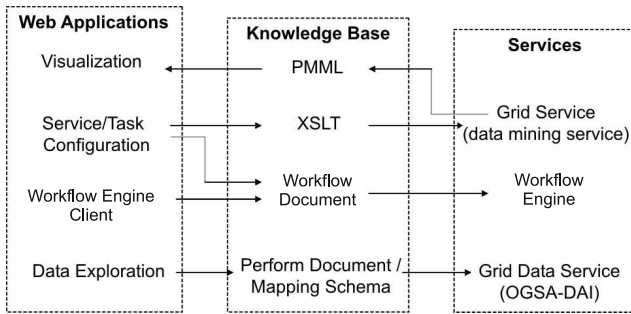


Fig. 2. Inter-component information flow

original data source or without creating a new one. It allows vertical selection of the columns and horizontal selection of the rows and modify their metadata or attribute values. The XSLT document is first created in the Data Preprocessing phase, due to the user specification, and also later used during the Data Selection phase. For example, the user can specify that the input data source must include only 'N' rows in which attribute 'A' has values 'x1' or 'x2', and all values of attribute 'B' are lower then 'R'.

Role of PMML. The Predictive Model Markup Language is an XML-based language for statistical and data mining models developed by the Data Mining Group⁴. It describes the inputs to data mining models, the transformations used prior to prepare data for data mining and the parameters, which define the models themselves. A PMML document is in our case an information resource which follows whole data mining process and is incrementally extended from its creation in 'Data Survey' phase, where input parameters are defined for the model; in the 'Model Building' step a new model is inserted into the document and is later used in the 'Model Evaluation' step and also in the final visualization.

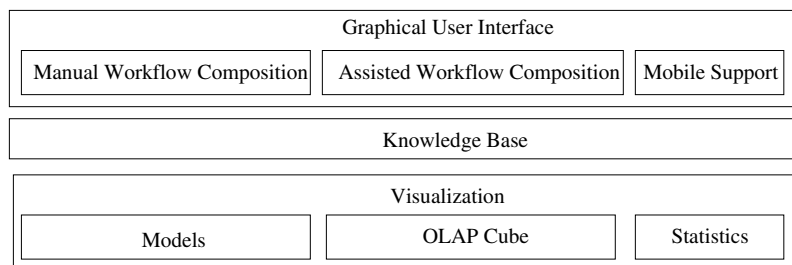
Role of Workflow (DSCL) Document. The Dynamic Service Control Language (DSCL) [14] is an XML-based language, which allows to specify data exploration workflows consisting of various Grid services and sequences of activities and their execution. This language was constructed for interactive control of Grid services as well as for defining their input and output parameters.

Role of Mapping Schema. Mapping Schema describes one virtual data source of data distributed in separated databases and files. It includes description of metadata (types and names of attributes) and operations (UNION or JOIN) for different types of data sources which come into the process of integration.

Role of Perform Document. This is a document specifying data resource access, transformation and delivery requests for the Grid Data Service (GDS). It contains a set of activities which are analyzed, validated and executed by the GDS engine to prepare the data set for processing by data mining services.

⁴ <http://www.dmg.org>

Non Grid Components



Grid Components

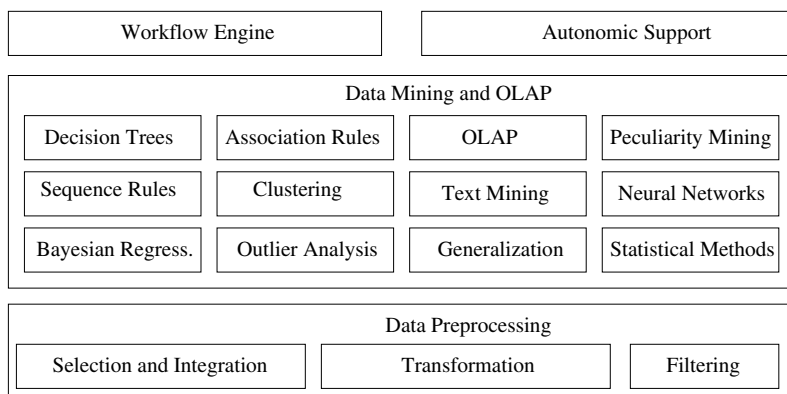


Fig. 3. Components Building the GridMiner Architecture

Knowledge Resources

Knowledge obtained as a result of KDD is managed by knowledge management infrastructure and is represented by ontologies, facts and rules as follows.

Role of Ontologies. Ontologies describe particular domains and offer semantic models of the resources and detailed description of their functionality and usability. We have created our own ontologies, which describe three basic domains: data sources, activities and data mining. All ontologies are implemented in the KB and specified in the Ontology Web Language (OWL)⁵.

Data source ontology. Data source ontology is one of the basic building blocks of the KB. It describes all frequently used data types which could be involved as inputs in KDD.

Data Mining Ontology. Data mining ontology is a special ontology about the data mining domain and areas related to it. This ontology is helpful in the process of knowledge discovery when an appropriate method and algorithm must be selected to reach the

⁵ <http://www.w3.org/2001/sw/WebOnt>

demanding results. The ontology is based on the categorization of data mining tasks, namely classification, association, sequences, clustering and OLAP, and some statistical functions.

Activity ontology. This ontology describes executive parts of the system responsible for knowledge discovery. Our system is service oriented and therefore, this ontology deals with a detailed description of Grid services able to perform demanded tasks.

Role of Facts. Facts are basic nuggets of knowledge and in our case, facts are instances of ontologies. They are used to state things that are unconditionally true in the domain of interest. From the combination of facts and rules, it is also possible to deduce new facts (inferred knowledge), that are not explicitly stored in the KB, but they are there implicitly present.

Role of Rules. Some data mining tasks produce a set of rules within the models they create. Rules can be applied on domains ontologies in order to retrieve some knowledge about those domains.

4.2 Visualization

Several advanced visualization methods for different data mining models, OLAP and statistical results have been introduced in the GridMiner project. The Web application responsible for visualization needs just a PMML document including the model specification. The document is transformed into its graphical representation including model description and its textual representation like, for example, the decision tree image and a list of decision rules.

4.3 Grid Services

Each service in the GridMiner architecture is implemented as a stand-alone Grid service specified by Open Grid Services Architecture (OGSA) [8].

Data Mining Services. The data mining process within the GridMiner is supported by several Grid services able to perform data mining tasks and OLAP. The suite of data mining services includes sequential, parallel and distributed implementations of data mining algorithms which are able to deal with data in the WebRowSet format provided by OGSA-DAI [1]. They all can deliver their results (data mining models) in the form of PMML documents, what makes them compatible with third party visualization applications.

Workflow Engine (WE) Service. The WE service [15] is an implementation of a workflow engine, which executes and controls Grid services according to the activities specified in a DSCL document. It is implemented as a stateful transient OGSA Grid service that can be instantiated by the WE client. The Autonomic Support component supports the workflow execution adaptivity to monitored changes of the execution environment.

Grid Data Service (GDS). The data mining services implemented in GridMiner expect to get input data in WebRowSet format. It is a collection of rows from database tables in XML, also called disconnected rowset, because it represents SQL rowset. Nowadays it is becoming a standard format for data exchange for Web based applications and is also supported by the latest Java development kit release. It has three basic elements a) properties, b) metadata, and c) data. The *properties* element describes basic dataset and its properties. The *metadata* gives information about the columns in the result set and the actual data is contained within the *data* element.

OGSA-DAI (Data Access and Integration) [1] is standard middleware implementation of GDS for supporting access and integration of data from separate data sources within the Grid. The latest distribution allows to define activities which can transform a result set into required form and deliver it as an XML file or as a data stream in the WebRowSet format. Therefore, it was also included in the architecture as a separate service supporting GridMiner services with the access to the data sources on the Grid.

Data Integration. The data integration in the GridMiner is based on the wrapper/mediator approach supported by the Grid Data Mediation Service (GDMS) [19], which allows integrating heterogeneous relational databases, XML databases and comma separated value files into one logically single homogeneous virtual data source. It is implemented as an activity in the OGSA-DAI and the mediation is specified in details by 'Mapping Schema' created during the data integration task.

4.4 Graphical User Interface

The aim of the GUI is to give the user a tool which hides the whole complexity of the architecture and offers an easy-to-use front end for the system administrator and data mining expert. It is designed as a standalone application able to be remotely started by Java Web Start (JWS)⁶ in any operating system supporting Java. It closely interacts with the knowledge base⁷ from which it gets information about all the resources to manage. It reads the KB at the startup phase and retrieves all information included in the user's data exploration project, e.g. used data sources, services and configurations, workflow description, etc. There are two main tasks of the GUI: (1) the interaction with Web applications and (2) workflow construction and steering its execution.

Web Applications Interaction. The GUI was proposed as a thin client, which interacts over a browser window with the "User Service Tier" of a supporting Web application and delivers just an HTML front end to the user. It invokes the Web application with required parameters and delivers output results back to the GUI, where they can be used as inputs to the other Web applications or written to the knowledge base for later reusing.

Workflow Construction and Execution. The other contribution of the GUI is to interactively construct workflows in a graphical representation at a high abstraction level. This

⁶ <http://java.sun.com/products/javawebstart>

⁷ Location of XMLDB is the only parameter needed to start the GUI.

representation is then automatically converted into the DSCL form. In this process, the user can be helped by an advanced Workflow Construction Assistant [17].

During the workflow run, the user can stop, resume, or cancel its execution.

Mobility Extension to the GridMiner. To allow interaction with the workflow engine from mobile devices, a new component, Mobile Support, was developed. The mobile GUI interacts directly with the workflow engine and gives to a user the ability of monitoring and controlling workflow execution. Moreover, it is able to interact with the OLAP service and data mining services - submit queries and display their results.

4.5 GridMiner Application in Brain Informatics - An Illustrative Example

In this subsection, we further elaborate the four-step data collection and analysis workflow outlined in Section 2 and briefly describe how it is constructed and executed in the GridMiner environment.

The workflow is constructed by the GUI means on so called workspace. There are areas of the GUI containing specific icons. For example, there are icons denoting data mining services placed in the data mining services GUI area. The icons can be drawn on the workspace, where they create nodes of the workflow graph.

The workflow construction begins with placing icons denoting fMRI, EEG and traditional psychometrics databases on the workspace. After clicking on an icon, it is possible to specify the logical name of the data source (This is called as configuration of the workflow node.), or, to accept the predefined one. Then icons denoting services for data filtration, transformation, mining, visualization are put on the workspace and appropriately configured. Then the icons (nodes) are connected by arrows denoting the appropriate flows of workflow. This results in the graphical representation of the *abstract workflow*. Clicking on the “Run” button, this workflow starts to be interpreted in three phases:

1. *Converting the graphical form into an abstract workflow document.* The internal representation of the graphical form is automatically converted into a DSCL representation of the abstract workflow, which is stored in a workflow document. This document is passed to the workflow engine (service).
2. *Resource discovery and allocation.* The workflow engine searches in registries, discovers physical resources (databases and services) and their replicas, and, based, on the built-in cost models selects the most appropriate resources. In this way, the abstract workflow document is converted into the *concrete workflow* document, also represented in DSCL.
3. *Workflow execution.* Now the workflow engine orchestrates the execution of concrete services. The user can stop the execution, investigate intermediate results, and resume or cancel the execution. The GridMiner also allows to observe and control the execution from mobile devices.

When the workflow execution has finished, the user can investigate the data mining results, estimate their interestingness and, if necessary, modify some workflow parameters and repeat the workflow processing.

5 From the GridMiner to the Wisdom Grid

The Wisdom Grid project aims, as the first research effort, to cover all aspects of the knowledge life cycle on the Grid - from discovery in Grid data repositories, to processing, sharing and finally reusing of knowledge as input for a new discovery [16] and implement them as a novel infrastructure based on Web and Grid services. The Wisdom Grid extends the GridMiner's functionality presented in the previous sections.

We think that the best way of grounding the subsequent discussion on the Wisdom Grid architecture, its design goals, functionality, and the techniques used is in terms of a scenario. To this end, we will use the following scenario, which is derived from discussions with healthcare professionals involved in treatment of patients with traumatic brain injuries (TBIs) [3]: A physician has a patient with a specific diagnosis and wants to know what should be done next in the treatment of this patient, what are the possible risks, what kind of drugs should be used, what outcome for this and similar patients can be assumed, etc. The physician (user) has several options to find out answers to these questions. Traditionally, he can discuss these issues with his colleagues, with a specialist by an on-line chat, he can look into a patient database to see what development can be observed from other patients with the same or similar symptoms and diagnosis, he could search the Internet for appropriate medical articles and reports on this problem, etc. However, in this way, it is not possible to obtain urgent information, which may often be a very critical issue in many medical applications, like TBI, or in crisis management environments in general. Analogous scenarios can be developed for Brain Informatics and other application domains.

To solve the above problems, an advanced information technology support is needed which fulfils a set of requirements, which include: (1) ability to access and analyze a huge amount of information, which is typically heterogeneous and geographically distributed; (2) intelligent behavior - ability to maintain, discover, extend, present, and communicate knowledge; (3) high performance (real-time or nearly real-time) query processing; and (4) high security guarantee.

Fig. 4 pictures a layered architecture of the Wisdom Grid infrastructure and depicts which actor is associated with each layer.

The *Knowledge Consumer* is an actor asking for a knowledge. It could be an agent, a service of a graphical user interface able to construct questions (e.g., in the FIPA ACL/RDF⁸ message format) in a way the *Intelligent Interface* understands. This actor initializes the whole process of knowledge discovery and receives the final results. The knowledge search process is organized by the *Knowledge Management Infrastructure* - either the knowledge is available in the Knowledge Base and can be immediately retrieved and passed the Intelligent Interface, or it is not available and has to be searched for by the *Data Mining and OLAP Infrastructure* in the databases attached to the Grid; the functionality of this infrastructure is based on the *Generic Grid Services* layer. The *Domain Application Expert* is responsible for building and management of appropriate ontologies, the *Data Mining and OLAP Expert* for data preparation for data mining and OLAP, selection of appropriate exploration methods and their parameters, and the *Service Provider Administrator* configures Generic Grid Services (e.g. the Globus

⁸ <http://www.fipa.org>

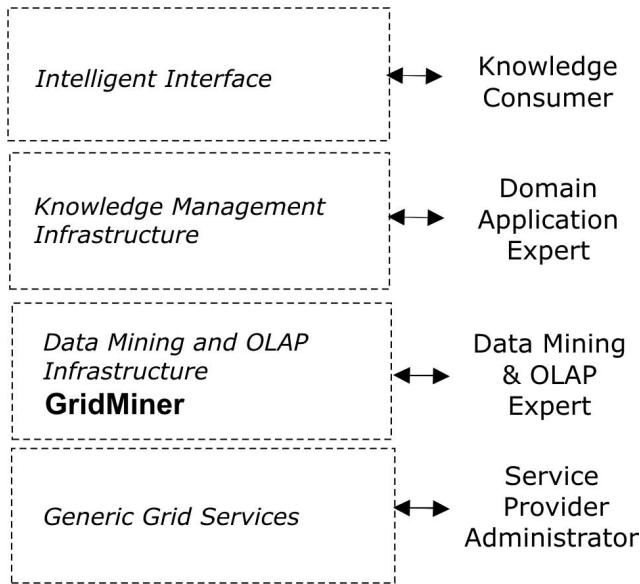


Fig. 4. Wisdom Grid Architecture

toolkit) and optimizes mapping of high-level Wisdom Grid services onto available Grid resources.

The overall Wisdom Grid architecture and participation of its components in the knowledge discovery workflow were described in our paper [4].

6 Related Work

In the past five years, data mining has become a subject of the Grid research, which has been reflected in several projects. The recently finished IST project DataMiningGrid⁹ developed advanced tools and services for deploying and using data-mining applications on the Grid. It is based on state of the art, Web Services Resource Framework-compliant technology (<http://www.globus.org>). The myGrid project [10] developed high-level middleware for data and legacy application resource integration such as resource discovery, workflow enactment and distributed query processing to support in silico experiments in biology. Their developed workflow system Taverna [20] provides semantic support for the composition and enactment of bioinformatics workflows for the life sciences community. The Science Environment for Ecological Knowledge (SEEK) project [13] aims to create a distributed data integration and analysis network for environmental, ecological and systematics data. Cannataro and Talia [5] present design of a Knowledge Grid architecture based on the non-service-based version of the Globus Toolkit. Discovery Net [6] provides a service-oriented computing model for knowledge discovery, allowing users to connect to and use data analysis software as

⁹ <http://www.datamininggrid.org/>

well as data sources that are made available online by third parties. The ideas of the Semantic Web [2] led Grid scientists to the notion of *Semantic Grid*, where they plan to apply Semantic Web technologies in Grid computing developments [9]. Web Intelligence is a new direction for scientific research and development that explores the next generation of the Web [24]. The Wisdom Grid concept discussed in this paper can be considered as a basis for development of Grid Intelligence and an important support for Web Intelligence.

7 Conclusions

We have presented the framework and initial implementation of our GridMiner system, which is the first information infrastructure covering all facets of knowledge discovery in data sources integrated to Grid environments. GridMiner can be coupled to Brain Informatics and other knowledge discovery and management applications, as we shown by an illustrative example. Its open and flexible architectures allows the integration of other data analysis and visualization services and their composition into advanced data exploration workflows. The ongoing research work in the GridMiner project includes development of novel interactive workflow management concepts based on the WS-BPEL language standard. Further development of the resulting WEEP workflow engine framework (<http://weep.gridminer.org>) has been accepted as a project within the Globus Alliance initiative.

References

1. Atkinson, M., Baxter, R., Hong N.C.: Grid data access and integration in OGSA (2002), <http://www.cs.man.ac.uk/grid-db/papers/OGSA-DAI-spec-1.2.pdf>
2. Berners-Lee, Y., Hendler, J., Lassila, O.: The semantic web. Scientific American (May 2001)
3. Brezany, P., Tjoa, A.M., Rusnak, M., Janciak, I.: Knowledge grid support for treatment of traumatic brain injury victims. In: Kumar, V., Gavrilova, M., Tan, C.J.K., L'Ecuyer, P. (eds.) ICCSA 2003. LNCS, vol. 2667, Springer, Heidelberg (2003)
4. Brezany, P., Goscinski, A., Janciak, I., Tjoa, A.M.: The development of a Wisdom Autonomic Grid. Knowledge Grid and Grid Intelligence 2004, Beijing (2004)
5. Cannataro, M., Talia, D.: Parallel and distributed knowledge discovery on the Grid: A reference architecture. In: Fourth International Conference on Algorithms and Architectures for Parallel Processing ICA³PP, Hong Kong, December 11-13, 2000, pp. 662-673. World Scientific (2000)
6. Curcin, V., Ghanem, M., Guo, Y., Kohler, M., Rowe, A., Syed, J., Wendel, P.: Discovery net: Towards a Grid of knowledge discovery. In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002)
7. Fayyad, U.M., Uthurusamy, R.: From data mining to knowledge discovery in databases. In: Fayyad, U.M., et al. (eds.) Advances in Knowledge Discovery and Data Mining, AAAI, pp. 37-54 (1996)
8. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The physiology of the Grid: An Open Grid Services Architecture for distributed systems integration (July 2002), <http://www.globus.org/research/papers/ogsa.pdf>
9. Goble, C., De Roure, D.: The Semantic Grid: A future e-science infrastructure, <http://www.semanticgrid.org>

10. Goble, C., Wroe, C., Stevens, R.: The myGrid project: services, architecture and demonstrator. Technical report, EPSRC e-Science Pilot Project myGrid (2003)
11. Hall, H.: Human brain informatics (HUBIN). Annual scientific report 2003. Karolinska Institute, Stockholm, March 5 (2004)
12. Han, J.: Data Mining. Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
13. Jones, M.B., Ludaescher, B., Pennington, D., Pereira, R., Rajasekar, A., Michener, W., Beach, J.: H., Schildhauer, M.: A knowledge environment for the biodiversity and ecological sciences. *Journal of Intelligent Information Systems* (2006)
14. Kicking, G., Brezany, P.: The Grid knowledge discovery process and corresponding control structures (March 2004), <http://www.gridminer.org/publications/gridminer2004-02.pdf>
15. Kicking, G., Hofer, J., Tjoa, A.M., Brezany, P.: Workflow management in GridMiner. In: 3rd Cracow Grid Workshop (2003)
16. McElroy, M.W.: The new knowledge management. *Journal of the KMCI* (2000)
17. Brezany, P., Janciak, I., Tjoa, A.M.: Ontology-based construction of data mining workflows for Grid environments. A book chapter to be published in, by Idea Group Inc. (2007)
18. Pyle, D.: Data preparation for data mining. Morgan Kaufmann, San Francisco (1999)
19. Woehrer, A., Brezany, P.: Mediators in the Architecture of Grid Information Systems (February 2004), <http://www.gridminer.org/publications/gridminer2004-01.pdf>
20. Wolstencroft, K., Oinn, T., Goble, C., Ferris, J., Wroe, C., Lord, P., Glover, K.: Panoply of Utilities in Taverna (2005)
21. Zhong, N.: How to make Web Intelligence (WI) meets Brain Informatics (BI) successfully. In: *COMPSAC* (1), pp. 24–25 (2006)
22. Zhong, N., et al.: Peculiarity oriented fMRI brain data analysis for studying human multi-perception mechanism. *Cognitive Systems Research* 5, 241–256 (2004)
23. Zhong, N., Hu, J., Motomura, S., Wu, J.: L., Liu, C.: Building a data-mining grid for multiple human brain data analysis. *Computational Intelligence* 21(2) (2005)
24. Zhong, N., Liu, J., Yao, Y. (eds.): *Web Intelligence*. Springer, Heidelberg (2003)
25. Zhong, N., Liu, J., Yao, Y.: Envisioning Intelligent Information Technologies (iIT) from the Stand-Point of Web Intelligence (WI). *Communications of the ACM* 50(3), 89–94 (2007)
26. Zhong, N.: Impending Brain Informatics (BI) Research from Web Intelligence (WI) Perspective. *International Journal of Information Technology and Decision Making* 5(4), 713–727 (2006)

A Semantically Enabled Service Oriented Architecture

Darko Anicic¹, Michael Brodie², Jos de Bruijn¹, Dieter Fensel¹,
Thomas Haselwanter¹, Martin Hepp¹, Stijn Heymans¹, Jörg Hoffmann¹,
Mick Kerrigan¹, Jacek Kopecky¹, Reto Krummenacher¹, Holger Lausen¹,
Adrian Mocan¹, James Scicluna¹, Ioan Toma¹, and Michal Zaremba¹

¹ Digital Enterprise Research Institute (DERI)

University of Innsbruck, Austria

firstname.lastname@deri.org

² Verizon Communications

USA

michael.brodie@verizon.com

Abstract. The researchers in DERI Innsbruck have been building an execution infrastructure for the Semantic Web Services (SWS) based on the Services Oriented Architecture (SOA) paradigm of loosely coupled components. While SOA is widely acknowledged for its potential to revolutionize the world of computing, that success depends on resolving several fundamental challenges, and especially in the case of open SOA environment the existing specifications do not address several issues. We aim in DERI Innsbruck to define a skeleton of the SWS system and implement the overall infrastructure with the aim of automating service discovery, negotiation, adaptation, composition, invocation, and monitoring as well as service interaction requiring data, protocol, and process mediation. We call this infrastructure a Semantically Enabled Service oriented Architecture (SESA). While there are already several specifications in the space for Web Services there are still elements missing, for example there is no specification describing how the particular components/services of the SWS infrastructure would work together. That work is carried out by DERI researchers in standardization bodies such as OASIS and W3C. In the near future a service-oriented world will consist of an uncountable number of services. Computation will involve services searching for services based on functional and non-functional requirements and an interoperating with those that they select. Services will not be able to interact automatically and SOAs will not scale without signification mechanization of a fixed set of components/services. Hence, machine processable semantics are critical for the next generation of computing, services and SOAs, to reach their full potential. The contribution of DERI Innsbruck is to define and implement the fixed set of services of an infrastructure that must be provided to enable a dynamic discovery, selection, mediation, invocation and inter-operation of the Semantic Web Services to facilitate the SOA revolution towards open environments. We recognize in DERI Innsbruck that SOA outside of tightly controlled environment cannot succeed until/unless the semantics issues are addressed. Only with semantics can critical subtasks can be automated leaving humans to focus on higher level problems.

1 Introduction

The most important issue in today's design of software architectures is to satisfy increasing software complexity as well as new IT needs, such as the need to respond quickly to new requirements of businesses, the need to continually reduce the cost of IT or the ability to integrate legacy and new emerging business information systems. In the current IT enterprise settings, introducing a new product or service and integrating multiple services and systems present unpredicted costs, delays and difficulty. Existing IT systems consist of a patchwork of legacy products, monolithic off-shelf applications and proprietary integration. It is even today's reality that in many cases users on the spinning chairs manually re-enter data from one system to another within the same organization. The past and existing efforts in Enterprise Application Integration (EAI) don't represent successful and flexible solutions. Several studies showed that the EAI projects are lengthy and the majority of these efforts are late and over budget. It is mainly costs, proprietary solutions and tightly-coupled interfaces that make EAI expensive and inflexible.

Service Oriented Architecture (SOA) solutions are the next evolutionary step in software architectures. SOA is an IT architecture in which functions are defined as independent services with well-defined, invocable interfaces. SOA will enable cost-effective integration as well as bring flexibility to business processes. In line with SOA principles, several standards have been developed and are currently emerging in IT environments. In particular, Web Services technology provides means to publish services in a UDDI registry, describing their interfaces using the Web Service Description Language (WSDL) and exchanging requests and messages over a network using SOAP protocol. The Business Process Execution Language (BPEL) allows composition of services into complex processes as well as their execution. Although Web services technologies around UDDI, SOAP and WSDL have added a new value to the current IT environments in regards to the integration of distributed software components using web standards, they cover mainly characteristics of syntactic interoperability. With respect to a large number of services that will exist in IT environments in the inter and intra enterprise integration settings based on SOA, the problems of service discovery or selection of the best services conforming users needs, as well as resolving heterogeneity in services capabilities and interfaces will again be a lengthy and costly process. For this reason, machine processable semantics should be used for describing services in order to allow total or partial automation of tasks such as discovery, selection, composition, mediation, invocation and monitoring of services.

In [3], the vision of serviceware as the next natural step beyond hardware and software is introduced: After four decades of rapid advances in computing, we are embarking on the greatest leap forward in computing that includes revolutionary changes at all levels of computing from the hardware through the middleware and infrastructure to applications and more importantly in intelligence. There we refine this towards a comprehensive framework that constitutes the bases for the technologies developed by DERI researchers, which integrates two complimentary and revolutionary technical advances, namely Service-Oriented Architectures (SOA) and Semantic Web, into a single computing architecture, that we call SESA. While SOA is widely acknowledged for its potential to revolutionize the world of computing, that success is dependent on resolving two fundamental challenges that SOA does not address, namely integration, and

search or mediation. In a service-oriented world, millions of services must be discovered and selected based on requirements, then orchestrated and adapted or integrated. SOA depends on but does not address either search or integration.

This report is structured as follows. Section 2 is providing a general summary of research goals pursued by DERI Innsbruck. We present DERI technologies in the following Section 3. Section 5 summarizes this report.

2 Objectives

Within DERI Innsbruck each of our researchers is working on a research goal. Such a research goal usually combines a research area, i.e., a major research challenges in SWS and SESA together with an implementation effort related to it. A research goal typically has a corresponding architectural component, and vice versa. A tight coupling between research goals and architectural components is desirable. The WSMX platform [5] will provide a SESA environment, which facilitates prototype development.

We distinguish 4 different types of elements of an overall SESA where each element type is composed by some sub-functionalities:

- The problem-solving layer which consists of (1) Ontologies, (2) Applications (e.g., e-tourism, e-government) and (3) Developer tools (GUI tools such as those for engineering ontology/web service descriptions; generic developer tools such as language APIs, parsers/serializers, converters, etc.).
- The broker layer which consists of (4) Discovery, (5) Adaptation (including selection and negotiation), (6) Composition (web service composition techniques such as planning), (7) Choreography, (8) Mediation ((a) Ontology mediation: techniques for combining Ontologies and for overcoming differences between Ontologies; (b) Process mediation: overcoming differences in message ordering, etc.), (9) Grounding, (10) Fault Handling (Transactionality, Compensation, etc.), and (11) Monitoring.
- The base layer that is providing the exchange formalism used by the architecture, i.e., (12) Formal languages (static ontology and behavioral, i.e., capability, choreography, orchestration languages, connection between higher-level descriptions, e.g., WSMML), (13) Reasoning (techniques for reasoning over formal descriptions; LP, DL, FOL, behavioral languages, etc.) and (14) Storage and Communication.
- Finally, vertical services such as (15) Execution management and (16) Security (authentication/authorization, encryption, trust/certification).

Figure 1 presents the current status of the WSMX and SESA infrastructure.

Currently, DERI Innsbruck focuses on the following essential components to bootstrap the overall approach: (1) Ontologies, (2) Applications, (3) Developer tools, (4) Discovery, (5) Adaptation, (6) Composition, (7) Choreography, (8) Mediation, (9) Grounding, (12) Formal languages, (13) Reasoning, (14) Storage and Communication, and (15) Execution management. There are no concrete plans yet for (10) Fault Handling, (11) Monitoring, and (16) Security. Some of this work may be provided by external DERI cooperation partners.

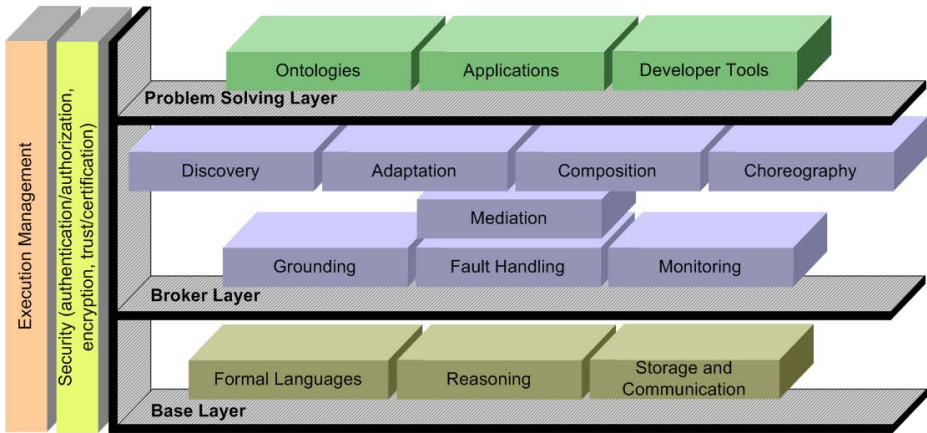


Fig. 1. SESA Infrastructure

While some of these functionalities are provided as services, the others remain the entities required to let the overall system to function, but they are not services in terms of a Service oriented Architectures. These results in SESA being called an infrastructure, not just an architecture. While the Base layers (without Formal Languages) builds SESA architecture in terms of the services, the Problem Solving layer adds the set of tools and entities, which causes that SESA becomes a complete Semantic Web Services oriented infrastructure.

SOAs typically consist of a set of services, and a coordinator that combines the services and puts them to use. Talking about SOA in the context of SESA can sometimes be misleading since SESA is a SOA and at the same time it is the coordinator of another, larger SOA. The SESA differentiates two types of services: platform services (such as Discovery, Choreography, Data and Process mediation etc.) and user services (e.g. back-end applications services). Platform services are necessary to enable the infrastructure to deliver its functionality as defined by its execution semantics. User services are exposed by information system external to the SESA infrastructure, but they are still coordinated using SESA platform services. The SESA recommendation defines the scope of particular platform services in terms of their functionality, while it remains silent about the scope and the functionality of user services.

The SESA infrastructure consists of several decoupled services. This enables independent refinement of these services, so each of them can have its own architecture, without hindering the overall SESA infrastructure. Following the SOA design principles, the SESA infrastructure separates the concerns of the individual components thereby separating the service descriptions and their interfaces from the implementation. This adds flexibility and scalability for upgrading or replacing the implementation of the services provided by the components that adhere to the required interfaces.

The SESA recognizes **vertical** and **horizontal** services. Vertical services remain invisible to horizontal services, and during its execution, the horizontal services remain

unaware that vertical services are executed. This type of vertical service is provided through the inversion of control.

3 Technologies/SESA Services

As presented in section 2, we distinguish four different types of elements of an overall SESA where each element type is composed by some sub functionalities:

- The problem-solving layer,
- The broker layer,
- The base layer,
- The vertical layer

This section describes in more detail functional components that play a role in the SESA architecture.

3.1 The Problem-Solving Layer

The problem-solving layer consists of Ontologies, Applications (e.g., e-tourism, e-government) and (3) Developer tools (GUI tools such as those for engineering ontology/web service descriptions; generic developer tools such as language APIs, parsers/serializers, converters, etc.).

Ontologies (chair: Martin Hepp). In this research topic, the working group advances the state of the art in the use of ontologies for advancement of the automation of business processes [7,13,12]. Ontologies in our understanding are community contracts about a representation of a domain of discourse. Representation in here includes (1) formal parts that can be used for machine reasoning, and (2) informal parts like natural language descriptions and multimedia elements that help humans establish, maintain, and renew consensus about the meaning of concepts.

Our contributions address the following two main dimensions of using ontologies for real application:

1. *Maturing Semantic Web foundations*, so that they become compatible with the real world complexity and scale. This includes three main branches of research:
 - *Ontology Engineering Methodologies* for and prototypes of industry-strength business ontologies, e.g. the gen/tax methodology for deriving ontologies from existing hierarchical standards and taxonomies (UNSPSC, eCI@ss, ...) and eClassOWL, the first serious attempt of building an ontology for e-business applications; and in general advancing the state of the art in e-business data and knowledge engineering, including metrics for content.
 - *Community-driven Ontology Building* For quite a while, we have been trying to hand back control over the evolution of ontologies to the user community, including semi-automated approaches and OntoWiki, a Wiki-centric ontology building environment. In this segment also fall quantitative comparisons of community-centric and engineering-based ontology building.

– *Economic Aspects of Ontology Building and Usage*

Building ontologies consumes resources, and in an economic setting, these resources are justified and will be spent (by rational economic actors, at least) only if the effort needed to establish and keep alive a consensual representation of a domain of discourse is outweighed by the business gain, either in terms of cost, added value, or strategic dimensions, e.g. process agility. This research branch is rather young and underdeveloped, but an important piece of understanding and fueling the use of ontologies in business applications.

2. *Building actual ontologies for core challenges of Information Systems* in order to realize and evaluate the business benefit, and to identify the open research challenges. We currently focus on five specific application domains:

- Semantics-supported Business Process Management, i.e. the idea to mechanize Business Process Management by using Semantic Web techniques and especially Semantic Web Services. There is a first vision paper and a Working Group being founded.
- Semantic Web services, especially WSMO/WSML/WSMX, i.e. the use of ontologies and related technology for the automation of Web services discovery, composition, execution, and monitoring.
- Electronic Markets and Electronic Procurement, including a reference framework for ontology-supported electronic procurement and an analysis of the true complexity of business matchmaking.
- eTourism, e.g. the automation of the discovery and booking of tourism offerings.
- Financial reporting, e.g. the automated mediation between financial data (e.g. XBRL data) so that balance sheets and other documents from multiple sources can be integrated on the fly.

Applications (chair: Michal Zaremba). The mission of the Application working group is to develop a common understanding of various technologies intended to facilitate the use of other services of SESA. This working group develops (1) use case scenarios that help validate the real-world fitness of SESA components and (2) domain-specific implementations which will be used for testing of SESA services.

Semantic Web Services challenge¹ has been the first attempt to provide a test-bed for the future Application working group [25]. The goal of the SWS Challenge is to develop a common understanding of various technologies intended to facilitate the automation of mediation, choreography and discovery for Web Services using semantic annotations. The challenge explores the trade-offs among existing approaches. Additionally we would like to figure out which parts of the problem space may not yet be covered. The challenge aims to provide a forum for discussion based on a common application. This Challenge seeks participation from industry and academic researchers developing software components and/or intelligent agents that have the ability to automate mediation, choreography and discovery processes between Web services.

¹ <http://www.sws-challenge.org>

Developer Tools (chair: Mick Kerrigan). The mission of the Developer Tools working group is to produce high quality tools related to Semantic Web Services that can be used by users of all competency levels. To this end DERI Innsbruck provide a large number of tools that can be used by users with different skill sets. Members of the working group are working on tools for managing WSMO ontologies, web services, goals and mediators, for creating mappings between WSMO ontologies for runtime mediation, for executing WSDL web services and managing WSMO execution environments.

The developer tools implemented with DERI are broken down into a number of plug-ins for Eclipse. The plug-ins are bundled together as two different products, namely the Web Services Modeling Toolkit (WSMT) and the DERI Ontology Management Environment (DOME). The WSMT [18,17] is aimed at covering all the functionality of WSMO, WSML and WSMX. With the WSMT the user is able to create and manage WSMO ontologies, web services, goals and mediators through the WSML human readable syntax, create mappings between two WSML ontologies for the purposes of instance transformation and thirdly manage and interact with the WSMX environment. The primary focus of DOME [11] is the use of WSMO as an ontology language and thus focuses only on the ontology and mediator parts of the WSMO specification. Users of DOME can create and manage their WSMO ontologies and mediators through the WSML human readable, XML and RDF syntaxes. They can also create mappings between two ontologies that can be used later by an execution environment.

3.2 The Broker Layer

The broker layer consists of Discovery, Adaptation (including selection and negotiation), Composition (web service composition techniques such as planning), Choreography, Mediation ((a) Ontology mediation: techniques for combining Ontologies and for overcoming differences between Ontologies; (b) Process mediation: overcoming differences in message ordering, etc.), Grounding, Fault Handling (Transactionality, Compensation, etc.), and Monitoring. There are no concrete plans yet for Fault Handling and Monitoring.

Discovery (chair: Holger Lausen). The Discovery working group develops different discovery implementations that are compatible with WSMO, WSML and specifically WSMX. The scope of the working group is to develop solutions based on existing descriptions (WSDL, UDDI, and text) as well as on more advanced descriptions based on semantic annotations using WSML [16,8]. Some tasks in this component are the development of a discovery engine based on keywords and existing annotations (WSDL), extended beyond for example WSDL description to related documentation, interpretation of, the semantic descriptions.

The working group has a parallel approach with respect to implementation of the component. On the one hand we implement some basic infrastructure including key word search on existing service documentation and on the other hand we will evaluate the different existing implementation to assess how they can be extended and integrated in a more common framework.

The working group develops a discovery engine based on keywords and existing annotations (WSDL, HTML docs, etc). The data set we will operate on will come from

publicly available Web service descriptions. Initially this has been limited to the information that can be obtained from the WSDL files. A search request can be expressed using keywords or advanced template search that allows to query for specific operation names or similar. WSDL documents can also be retrieved by URL. This phase will also include basic monitoring functionality for determining if the service specified in the given WSDL document is available.

Besides the pure WSDL description the working group extends the search to related documentation. This related information can be extracted from UDDI (taxonomies, specific tmodels, etc) or relevant web pages. Based on the initial work of the WSRD group in DERI Galway some information retrieval techniques will be applied to the keywords found in the WSDL file corpus.

The working group has also started interpreting the semantic descriptions, where the search expressivity increases with the expressivity of the underlying ontology language. The working group will be open to various approaches, at the beginning those approaches might not be compatible (i.e. annotations using language/model A will not necessarily work with annotations using language/model B). However the long-term goal is to make them interoperable or merge them.

Adaptation (chair: Ioan Toma). After discovering a set of potentially useful services, a SESA needs to check whether the services can actually fulfill the users concrete goal and under what conditions. Those that cannot fulfill the goal are removed from the list of discovered services. This step is required as it is not feasible for a service to provide an exhaustive semantic description. Giving the Amazon bookstore service as an example, it is not feasible for Amazon to update the semantic description of their Web service every time a new book is available or an existing book is changed, therefore we must whether that Amazon actually currently has a copy of the book requested by the user, and at an acceptable price. The process of checking whether and under what conditions a service can fulfill a concrete Goal is called negotiation in SESA, and it also encompasses by so-called filtering. By filtering we understand the process of narrowing the set of discovered services which provide the functionality requested, by considering only the services that have the appropriate non-functional properties requested by the user. Furthermore building a ranking/order relation based on non-functional properties criteria like price, availability, etc. is also part of the filtering process [29].

Once a list of Web services that can fulfil the user's concrete goal is prepared, a SESA must then choose one of the services to invoke. It is important that this selection is tailored to the user's needs, as for example while one user may require high quality another may prefer low price. This process is called selection. Negotiation, filtering and selection are tasks of the Adaptation working group.

Composition (chair: Jörg Hoffmann). The Composition working group develops methods to do Web Service composition (WSC), starting from web service descriptions at various levels of abstraction, specifically, the functional level and process level components of WSMO. Such methods are implemented as tools in the relevant contexts, in particular WSMX. Potential applications of WSC technology are researched, and modelled using WSMO/WSML; case studies are run with the developed tools, ultimately resulting in technology export.

In more detail, the working group revolves around the following topics:

- Language subset/capability extensions. The working group deals with as large as possible subsets of WSMO/WSML. Naturally, the developed technology has started with restricted language subsets, and incrementally moves on to richer subsets. If new features/scenarios become relevant on the side of WSMO/WSML, these will become new targets for WSC.
- Applications, case studies, benchmarking, technology export [15]. A vital ingredient to WSC research is to stay as close as possible to the envisioned fields of commercial/industrial application. The working group uses and strengthens DERI's contacts in this respect. Possible areas of application shall be identified, and increasingly realistic scenarios shall be modelled. These models play a crucial role in evaluating the developed WSC techniques, and thus guiding the research into which kinds of methods will work and which will not. The case studies may eventually lead to fostered collaborations and, ultimately, to technology export.
- Addressing efficiency problems [14]. Since WSC is a notoriously hard problem it is PSPACE-complete even in extremely simple formalisms it is essential to develop heuristic techniques that have the potential to scale satisfyingly in practical instances of the WSC problem.
- The working group expects that, eventually, notions of optimality will become relevant for WSC: What is the best service satisfying the composition task, and how can we compose that service? We intend to contribute to both the development of such notions and to their algorithmic treatment.

Choreography (chair: James Scicluna). The Choreography part of SESA is meant to provide a process language which should allow for formal specifications of interactions and processes between the service providers and clients, define reasoning tasks that should be performed using this language, and implement an engine to support the execution of interactions, as well as to support reasoning in this language.

The model for WSMO Choreographies is currently stable [27]. It is inspired by the Abstract State Machines (ASM) methodology and inherits the core principles such as the state, transition rules and flexibility to model any kind of behavior. The syntax of the choreography language has been defined as a result of the model. It is similar to the ASM language with some obvious constructs that have been introduced in order for it to fit with the WSML language. The semantics are defined using a set-based approach and describe the operational behavior of choreographies on the same lines as for ASMs.

The work of the Choreography API has been divided in different parts, namely, the API (i.e. the interfaces), the implementation, the parser and the serializer. The API defines the interfaces and methods (with no implementation) for the objects within the language constructs. The implementation part implements the interfaces so that a user can easily create and manage the language constructs. The parser loads up an object model representation in the memory from a choreography description in a WSML file. The serializer, performs the reverse operation, that is, it saves the memory representation of the language to the equivalent syntax representation in a WSML file. All of these modules have been completed.

The main steps involved in the implementation of the choreography engine are the design with particular emphasis on the interaction with other WSMX components and the actual programming. Both of these aspects are in a stable condition but eventually they evolve as WSMX gets better and as requirements change. Particularly, we will work further towards a more expressive and intuitive language to deal with problems related to choreography reasoning with special emphasis on web service compatibility [23]. This language will be as a layer on top of the ASM methodology and hence compliant with such a formalism. Finally, this language will be able to express the existing interaction and workflow patterns which capture the major possible use cases in business processes and service interactions.

Mediation (chair: Adrian Mocan). Mediation in SESA aims at providing flexible mediation service at both data and process level. Data Mediation provides automatic data transformation from the format used by the source party to the format required by the target party involved in conversation, while Process Mediation is concerned with the heterogeneity of the public processes of the participants in a conversation.

- Data Mediation provides automatic data transformation from the ontology used by the source party to the ontology required by the target party involved in conversation [24]. As WSMX is a semantic enabled service execution environment, we assume that the data to be mediated is semantically described, i.e. it consists of ontology instances. As a consequence the WSMX Data Mediation Service has to support instance transformation from terms of one ontology to the terms of another ontology, based on the set of already created mappings between the two given ontologies.
- The Process Mediator service has the task of solving the communication (behavioral) mismatches that may occur during the communication between a requestor and a provider of a service [4]. As in WSMO, the requestor is a WSMO Goal, while the provider is a Semantic Web Service, the Process Mediators task is be to accommodate the mismatches between the goals requestedChoreography and the SWSs choreography.

Grounding (chair: Jacek Kopecky). Apart from discovering web services and composing them, a SESA also needs to communicate with the Web services send the necessary request messages and receive the responses. Because internal communication within the SESA uses semantic data and practically all currently deployed Web services use their specific XML formats, the External Communication component needs to translate between the involved data forms. This translation is also known as data grounding [20]. Above that, this SESA also needs to support concrete network protocols (HTTP, SOAP, other bindings) to be able to exchange messages with the Web service.

As grounding has to be based on the Web Services Description Language, the work on this component also contains W3C efforts towards Semantic Web Services. In particular, this means the WSDL RDF mapping [19] from Web Service Description WG, and the Semantic Annotations for WSDL [22] in the SA-WSDL WG at W3C.

3.3 The Base Layer

The base layer provides the exchange formalism used by the architecture, i.e., Formal languages, Reasoning (techniques for reasoning over formal descriptions; LP, DL, FOL, behavioral languages, etc.) and Storage and Communication.

Formal Languages (chair: Jos de Bruijn). Descriptions in a Semantically-Enabled Service Oriented Architecture (SESA) need different formal languages for the specification of different aspects of knowledge and services [16,26,6]. The descriptions in a SESA can be decomposed into four dimensions:

- Static knowledge (ontologies)
- Functional description (capabilities)
- Behavioural description (choreography and orchestration)
- Non-functional Properties

Tasks for this working group include the integration of FOL-based and nonmonotonic LP-based languages, the explicitization of context for use with scoped negation, and the development of rules for the, Semantic Web (through the W3C RIF working group). Furthermore, requirements on the functional descriptions of services and as well as a semantics for web service functionality need to be devised. Requirements need to be gathered on the description of a choreography and an orchestration and a semantics needs to be devised. Finally, purpose and usage of non-functional requirements will be investigated. Future work of this working group will focus on:

- Static knowledge integrating knowledge based on classical first-order logic and nonmonotonic logic programming; important issues are the representational adequacy of the integration, as well as decidable subsets and a proof theory, so that reasoning becomes possible; scoped default negation; rules for the Semantic Web RIF working group; connection between Semantic Web languages RDF, OWL.
- Functional description requirements need to be gathered on the functional specification of services and a semantics needs to be devised which can be combined with the language for the description of ontologies, in order to enable the use of ontologies for the description of web service functionality. An important use case for the functional description of services is discovery. Therefore, it is expected that many requirements on the functional description of services will come from the discovery research goal.
- Behavioural description there exist several formal languages which are suitable for behavioural description. Examples are transaction logic, situation calculus, and action languages. Requirements need to be gathered on the description of choreography and an orchestration and semantics needs to be devised. A key challenge is the combination of this language with ontology languages in order to enable the reuse of ontology vocabulary in the choreography and orchestration descriptions. Finally, this language needs to be connected to the language for capability description in order to prove certain correspondences between the functional and behavioural description of services.

- **Non-functional Properties** Non-functional properties can at least be divided into two categories: (1) meta-data, e.g., author, description, etc., of the WSML statements in a description and (2) actual non-functional properties, i.e., actual properties of services (e.g. pricing, QoS, transactions). NFPs require a deeper investigation into their purpose and their usage.

The work on Formal Languages establishes the theoretical foundations for the WSML family of languages, used for the description of Web Services.

Reasoning (chairs: Darko Anicic and Stijn Heymans). The Reasoning working group develops an efficient and extensible reasoning engine for expressive rule-based languages (WSML Core/Flight/Rule), as well as description logic based languages (WSML-DL) [1,2]. The reasoner is based on state-of-the-art reasoning algorithms (for query answering, logical entailment, etc.). The SESA needs the reasoning component for service discovery as well as both process and data mediation. Mission critical features of the Reasoning component are: hybrid reasoning based on DLs and logic programming, reasoning with very large instance bases, reasoning with heterogeneous and conflicting information, and reasoning in distributed environments. Also one of the major objectives of this working group is the implementation of a Rule Interchange Format (RIF)². RIF aims to specify a common format for rules in order to allow rule interchange between diverse rule systems. This format (or language) will function as an interlingua into which rule languages can be mapped, allowing rules written in different languages to be executed in the same reasoner engine. The RIF layer of our reasoner engine will be capable of handling rules from diverse rule systems and will make WSML rule sets interchangeable with rule sets written in other languages that are also supported by RIF.

Storage and Communication (chair: Reto Krummenacher). The storage components, plural on purpose, shall provide repositories to store objects needed to ensure successful processing of user request to SESA. There might be a need for different storages tailored to the particular needs: web service descriptions, goals, mediation rules, workflows, and execution semantics. It is already known that the Execution Management component requires repositories for ontologies and data instances (service descriptions in particular). The idea is to use a Triple Space infrastructure to do so. The mission of the Storage Component team is thus to find out which means of storage are required and in what way these requirements can be fulfilled in the easiest and simplest way to provide optimal service to the application layer components and the vertical services [28,21,9].

3.4 Vertical Services

Vertical services consist of Execution management and Security (authentication and authorization, encryption, trust/certification). There are no concrete plans yet for Security.

Execution Management (chair: Thomas Haselwanter). The Execution Management working group is responsible for the management of WSMX as a platform and for

² <http://www.w3.org/2005/rules/>

the coordination of the individual components [30,10]. As the kernel of the system it enables and realizes the overall operational semantics of WSMX that let the system achieve the promised functional semantics of its client-side interface. It takes the functionality offered by the individual components of the framework and orchestrates these atomic pieces into a coherent whole in an orderly, and consistent fashion. These properties are guaranteed by the execution semantics, which are executed over the set of services that are available to the execution management component

4 Technical Task Force

A comprehensive framework as described in the previous section can only be effective if the different components are aware of the global vision as well as communicate among each other to accomplish this joint vision. Chaired by Mick Kerrigan, the mission of the Technical Task Force is to oversee the implementation efforts within each of the DERI objectives to ensure that the different prototypes are interoperable. The Technical Task Force aims to improve communication between the different objectives to ensure transparency and understanding of current development efforts. This will be achieved by identifying dependancies between working groups, sharing the requirements one working group has on another and aiding a working group in prioritizing certain implementation efforts based upon the needs of other groups. The Technical Task Force meets on a monthly basis to discuss the current status of the implementation efforts across the objectives and devise plans to bring these prototypes into a coherent architecture.

5 Summary

This paper outlined a comprehensive framework that integrates two complimentary and revolutionary technical advances, Service-Oriented Architectures (SOA) and Semantic Web, into a single computing architecture, that we call Semantically Enabled Service oriented Architecture (SESA) and details of how these technologies are developed within DERI Innsbruck have been provided. While SOA is widely acknowledged for its potential to revolutionize the world of computing, this success is dependent on resolving two fundamental challenges that SOA does not address, namely integration, and search or mediation. In a service-oriented world, millions of services must be discovered and selected based on requirements, then orchestrated and adapted or integrated. SOA depends on but does not address either search or integration. The contribution of DERI Innsbruck is to provide the semantics-based solutions to search and integration that will enable the SOA revolution. The paper provides a vision of the future, enabled by SESA, that places computing and programming at the services layer and places the real goal of computing, problem solving, in the hands of end users.

Acknowledgements

The work is funded by the European Commission under the projects ASG, DIP, enI-RaF, InfraWebs, Knowledge Web, Musing, Salero, SEKT, SEEMP, SemanticGOV, Super, SWING and TripCom; by Science Foundation Ireland under the DERI-Lion Grant

No.SFI/02/CE1/I13 ; by the FFG (Österreichische Forschungsförderungsgesellschaft mbH) under the projects Grisino, RW², SemNetMan, SEnSE, TSC and OnTourism.

References

1. Abiteboul, S., Hull, R., Vianu, V.: *Foundations of Databases*. Addison-Wesley, Reading (1995)
2. Beerli, C., Ramakrishnan, R.: On the power of magic. In: *PODS 1987: Proceedings of the sixth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 269–284. ACM Press, New York (1987)
3. Brodie, M., Bussler, C., de Bruijn, J., Fahringer, T., Fensel, D., Hepp, M., Lausen, H., Roman, D., Strang, T., Werthner, H., Zaremba, M.: *Semantically enabled service oriented architectures: A manifesto and a paradigm shift in computer science*. Technical report, DERI Innsbruck (2006)
4. Cimpian, E., Mocan, A., Stollberg, M.: Mediation enabled semantic web services usage. In: Mizoguchi, R., Shi, Z., Giunchiglia, F. (eds.) *ASWC 2006*. LNCS, vol. 4185, Springer, Heidelberg (2006)
5. Cimpian, E., Moran, M., Oren, E., Vitvar, T., Zaremba, M.: D13.0 overview and scope of wsmx
6. de Bruijn, J., Lausen, H., Polleres, A., Fensel, D.: The web service modeling language: An overview. In: Sure, Y., Domingue, J. (eds.) *ESWC 2006*. LNCS, vol. 4011, pp. 590–604. Springer, Heidelberg (2006)
7. Fensel, D.: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, Heidelberg (2004)
8. Fensel, D., Keller, U., Lausen, H., Polleres, A., Toma, I.: What is wrong with web services discovery. In: *W3C Workshop on Frameworks for Semantics in Web Services* (2005)
9. Harth, A., Decker, S.: Optimized index structures for querying rdf from the web. In: *Proc. 3rd Latin American Web Congress (LA-WEB 2005)* (2005)
10. Haselwanter, T., Zaremba, M., Zaremba, M.: Enabling components management and dynamic execution semantic in wsmx. In: *Proceedings of the WIW 2005 Workshop on WSMO Implementations* (2005)
11. Henke, J.: The table metaphor: A representation of a class and its instances. In: Gómez-Pérez, A., Euzenat, J. (eds.) *ESWC 2005*. LNCS, vol. 3532, Springer, Heidelberg (2005)
12. Hepp, M.: Representing the hierarchy of industrial taxonomies in owl: The gen/tax approach. In: *ISWC Workshop Semantic Web Case Studies and Best Practices for eBusiness (SWCASE 2005)* (2005)
13. Hepp, M.: Products and services ontologies: A methodology for deriving owl ontologies from industrial categorization standards. *Int'l Journal on Semantic Web and Information Systems (IJSWIS)*, pp. 72–99 (2006)
14. Hoffmann, J., Brafman, R.: Conformant planning via heuristic forward search: A new approach. *ai* 170(6–7), 507–541 (2006)
15. Hoffmann, J., Edelkamp, S., Thiebaux, S., Englert, R., Liporace, F., Trüg, S.: Engineering benchmarks for planning: the domains used in the deterministic part of ipc-4. 26, 453–541 (2006)
16. Keller, U., Lausen, H., Stollberg, M.: On the semantics of functional descriptions of web services. In: Sure, Y., Domingue, J. (eds.) *ESWC 2006*. LNCS, vol. 4011, Springer, Heidelberg (2006)
17. Kerrigan, M.: The wsmx editor plug-in to the web services modeling toolkit. In: *Proceedings of the 2nd WSMO Implementation Workshop (WIW)* (2005)

18. Kerrigan, M.: WSMOViz: An Ontology Visualization Approach for WSMO. In: Proceedings of the 10th International Conference on Information Visualization (July 2006)
19. Kopecky, J., Parsia, B. (eds.): Web services description language (wsdl) version 2.0: Rdf mapping. Technical report
20. Kopecky, J., Roman, D., Moran, M., Fensel, D.: Semantic web services grounding. In: Proc. of the Int'l Conference on Internet and Web Applications and Services (ICIW 2006) (2006)
21. Krummenacher, R., Hepp, M., Polleres, A., Bussler, C., Fensel, D.: Www or what is wrong with web services. In: Proc. of the 2005 IEEE European Conf. on Web Services (ECOWS 2005) (2005)
22. Lausen, H., Farrell, J. (eds.): Semantic annotations for wsdl. Technical report
23. Martens, A.: On Compatibility of Web Services. Petri Net Newsletter 65, 12–20 (2003)
24. Mocan, A., Cimpian, E., Kerrigan, M.: Formal Model for Ontology Mapping Creation. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L. (eds.) ISWC 2006. LNCS, vol. 4273, Springer, Heidelberg (2006)
25. Petrie, C.: It's the programming, stupid. IEEE Internet Computing, "Peer to Peer" (2006)
26. Polleres, A., Feier, C., Harth, A.: Rules with contextually scoped negation. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, Springer, Heidelberg (2006)
27. Roman, D., Scicluna, J.: D14 ontology based choreography of wsmo services. Technical report (2006)
28. Shafiq, O., Krummenacher, R., Martin-Recuerda, F., Ding, Y., Fensel, D.: Triple space computing middleware for semantic web services. In: Proc. 2006 Middleware for Web Services (MWS 2006) Workshop at the 10th Int'l IEEE Enterprise Computing Conference (EDOC 2006) (2006)
29. Toma, I., Foxvog, D.: D28.4 non-functional properties in web services. Technical report (2006)
30. Zaremba, M., Moran, M., Haselwanter, T.: Applying semantic web services to virtual travel agency case study. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, Springer, Heidelberg (2006)

Spam Filtering and Email-Mediated Applications

Wenbin Li^{1,5}, Ning Zhong^{1,2}, Y.Y. Yao^{1,3}, Jiming Liu^{1,4}, and Chunnian Liu¹

¹ The International WIC Institute, Beijing University of Technology, China

² Dept. of Life Science and Informatics, Maebashi Institute of Technology, Japan

³ Dept. of Computer Science, University of Regina, Canada

⁴ Dept. of Computer Science, Hong Kong Baptist University, Hong Kong

⁵ Shijiazhuang University of Economics, China

Abstract. This chapter reviews and examines two important research topics related to intelligent email processing, namely, email filtering and email-mediated applications. We present a framework to show a full process of email filtering. Within the framework, we suggest a new method of combining multiple filters and propose a novel filtering model based on ensemble learning. For email-mediated applications, we introduce the concept of operable email (OE). It is argued that operable email will play a fundamental role in future email systems, in order to meet the need of the World Wide Wisdom Web (W4). We demonstrate the use of OE in implementing an email assistant and other intelligent applications on the World Social Email Network (WSEN).

1 Introduction

Email is one of the most useful communication tools over the Internet. In the last few decades, the functionality of email has been in a constant evolution, from simple message exchange to multimedia/hypermedia contents communication, and to push technologies for direct marketing. Email becomes indispensable to our academic research and daily life. In the meantime, when we feel that *we cannot live without email*, we often strongly feel that *we cannot live with it*.

Email has brought us many problems, such as fast-spreading viruses, security threats, and especially massive spam [11]. In recent years, many efforts have been made to solve these issues. For example, machine learning based methods were used to filter spam [1,9,15,18,19,20,24,27]. Agent-based techniques were used to develop an email processing assistant [3,8,12,21]. Measures of Social Network Analysis (SNA) were used for the task of anti-virus [5]. Although these studies have reported some encouraging results, more research efforts and better email filtering techniques are in urgent demand.

Traditional email typically does not have semantic features. Without machine-understandable semantics of email, it is very difficult to automate and support email-mediated intelligent applications. Thus, the existing email systems can cause many inconveniences and sometimes serious consequences. For example, users may frequently forget an important appointment. They may be tired of

those repeated tasks which could be automated. They may not be able to remember where an attachment has been stored. Email users need new systems that can support them and help them to solve their problems effectively.

For the next generation Web, the concepts of traditional email and associated systems are no longer sufficient. To some extent, the recently proposed model of semantic email may remedy such a difficulty [23]. It is crucial to explore other novel ideas and email models to meet the challenges of the new generation Web. We need new theories, technologies, tools and applications for relieving users from the email malaise.

This chapter mainly addresses and demonstrates two levels of WI technologies in the context of email: 1) *infrastructure-level spam filtering* and 2) *application-level intelligent assistance*. For the former, we provide a full process of email filtering, in which a novel method is suggested by combining multiple filters based on ensemble learning. For the latter, we propose an email system called operable email (OE) to meet the need of the World Wide Wisdom Web (W4). We discuss how to use OE to implement an email assistant and several intelligent applications on the World Social Email Network (WSEN).

2 Email Filtering

2.1 Formal Description

Email messages can be modeled as semi-structured documents that consist of a set of classes and a number of variable length free-text. Thus, many text mining techniques, especially Automated Text Categorization (ATC), can be used to develop an email filtering and management system. Under the framework of ATC, email filtering is viewed as a 2-class categorization task. Two kinds of errors will occur when a filter labels new emails, i.e., the false positive error and the false negative error. The former is mislabeling a legitimate email as a spam and the latter is mislabeling a spam as a legitimate email. The costs of the two types of errors are different. Following the definition of ATC in [26], we give a formal definition of the Automated Cost-Sensitive Email Filtering problem.

Suppose $D = \{d_1, d_2, \dots, d_{|D|}\}$ is the training set and $C = \{c_0 = \text{"spam"}, c_1 = \text{"legitimate"}\}$ is the set of their classes or categories, where $|D|$ denotes the cardinality of D . Each email $d_i \in D$ belongs only to one of the email categories. Formally, this can be expressed as a function $\phi : D \times C \longrightarrow \{\text{true}, \text{false}\}$, where $D \times C$ is the Cartesian product of D and C . The function ϕ assigns *true* to (d_i, c_j) if c_j is the real category of d_i and *false* otherwise. The key task of email filtering is to learn the classification function ϕ . In general, without any constraints on the form and properties of the classification function ϕ , this is almost an impossible task. In what follows, we use a learning algorithm ℓ to obtain an approximation $\hat{h} : D \times C \longrightarrow \{\text{true}, \text{false}\}$ of the unknown target function ϕ . The function $\hat{h}$ is called a *filter* and should be as close to ϕ as possible.

Unlike the cost-insensitive classifiers which minimize zero-one loss or an error rate, cost-sensitive email filters choose the class that minimizes the expected cost of a prediction as given by [10]:

$$c(c_i|\vec{x}) = \sum_{j=0}^1 P(c_j|\vec{x})c(i,j) \quad (1)$$

where $\vec{x}$ is the vector representation of an email, $c(i,j)$ ($i, j \in \{0,1\}$) denotes the cost of classifying an email of c_i into c_j , $P(c_j|\vec{x})$ is the conditional probability that $\vec{x}$ belongs to c_j , and $c(c_i|\vec{x})$ is the expected cost predicting $\vec{x}$ to c_i . A filter $\hat{h}$ is obtained by using Bayes optimal prediction, which guarantees that it achieves the lowest possible overall cost.

Cost-sensitive email filtering can be defined as a task that learns an approximate classification function $\hat{h} : D \times C \rightarrow \{true, false\}$ from full or partial data in D with a cost-sensitive algorithm ℓ_c , and it minimizes the expected cost of prediction. The main issues for building an email filter include training set preparation, email representation, feature selection, the filtering model learning with ℓ_c and filter evaluation.

Training Datasets. In general, a user is unwilling to release legitimate emails because of receivers' or senders' privacy. Collecting benchmark email datasets is very difficult. There are only a few collections that are publicly available for the research communities: PU1, Ling-Spam, SpamAssassin, Spambase and TREC05.

PU1¹ consists of 481 real legitimate emails and 618 spam. Header fields except the subject and HTML tags of messages in PU1 are removed. In order to bypass privacy issues, each token was mapped into a unique integer. PUA, PU2, PU3, together with PU1 are called PU collection. Comparing with PU1, the other three adopt different pre-processing methods.

Ling-Spam² consists of 481 spam messages received by the provider, and 2412 legitimate messages retrieved from the archives of a mail list. According to its providers, legitimate messages in Ling-Spam are more topic-specific than the legitimate messages most users receive. The performance of a learning-based anti-spam filter on Ling-Spam may be an over-optimistic estimate of the performance that can be achieved on the incoming messages of a real user, where topic-specific messages may be less dominant among legitimate messages. Ling-Spam is currently the best (although not perfect) candidate for the evaluation of spam filtering [2]. Like PU collection, Ling-Spam also has other three versions.

SpamAssassin³ contains 6047 messages, 4150 of which are marked as legitimate and 1897 as spam. Legitimate emails in SpamAssassin are messages collected from BBS or real emails donated by personal users. Androutsopoulos and colleagues claimed that the performance of a learning-based filter on the SpamAssassin may be an under-estimate of the performance that a personal filter can achieve [2].

Spambase⁴ only distributes information about each message rather than the messages themselves to avoid privacy issues. With 57 pre-selected features, each real email was represented by a vector. This corpus contains 4601 vectors about

¹ <http://www.iit.demokritos.gr/~ionandr/publications/>

² <http://www.iit.demokritos.gr/~ionandr/publications/>

³ <http://spamassassin.org/publiccorpus/>

⁴ <http://www.ics.uci.edu/~mllearn/databases/spambase/>

emails. Among them, 2788 vectors are about legitimate messages and 1813 are about spam. Spambase is much more restrictive than above three corpus. Its messages are not available in raw form. It is impossible to experiment with features other than those chosen by its creators.

The 2005 TREC⁵ was created for the TREC Spam Evaluation Track. TREC Spam Corpus (TREC'05) contains 92,189 email messages. 52,790 messages are labeled as spam while 39,399 are labeled as ham.

Email Representation and Feature Selection. As a prerequisite for building a filter, one must represent each message so that it can be accepted by a learning algorithm ℓ_c . The commonly used representation method is the term/feature weight vector in the Vector Space Model (VSM) in information retrieval [25]. Suppose $V = \{f_1, f_2, \dots, f_{|V|}\}$ is the set of vocabulary which consists of features (i.e., words or phrases) appeared in D , where $|V|$ is the size of V . A vector representation of an email is defined as a real-value vector $\vec{x} \in \mathbb{R}^{|V|}$, where each component x_j (also called weight) is statistically related to the occurrence of the j th vocabulary entry in the email. The value of x_j can be computed based on two types of frequencies: the absolute feature frequency and the relative feature frequency. The absolute frequency is simply the count of f_j appearing in the email. Based on absolute frequency, one can easily obtain a binary weight scheme. That is, $x_j \in \{0, 1\}$ simply indicates the absence or the presence of feature f_j in the email. A very popular weighting scheme is the so-called *tf* $\times$ *idf* weighting defined by:

$$\chi_j = tf_j \cdot \log_2\left(\frac{N}{df}\right) \quad (2)$$

where tf_j is the number that f_j occurs in the email, N is the number of training emails, and df is the number of training emails in which f_j occurs.

For a moderate-sized email test collection, the size of V will reach tens or hundreds of thousands. The size of V is prohibitively high for many learning algorithms, such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT) and so on. A method that automatically reduces the dimensions without sacrificing filtering accuracy is highly desirable. This process is called feature selection (FS for short). In general, FS mainly includes two steps. The first step is to calculate feature weight for each entry in V with an FS function φ , and the second step is to rank all features and extract the top M (in general $M \ll |V|$) features from V . Formally, FS can be defined as a mapping process from V to V' by using φ , i.e., $V \xrightarrow{\varphi} V'$, where $V' \subset V$, V' is a substitute of V .

Many functions are available for completing the above mapping process, such as Information Gain (IG), χ^2 -test (CHI), Document Frequency (DF), Mutual Information (MI), Term Strength (TS), Odds Ratio (OdR) and so on [28]. By comparing the effectiveness of the five functions in the context of general ATC tasks, the reported experimental results show that IG, CHI and DF are more effective than MI and TS.

⁵ <http://plg.uwaterloo.ca/~gvcormac/trecrcorpus/>

Filter Evaluation. In ATC tasks, performance is often measured in terms of *precision*, *recall*, *F1*, broken point, ROC analysis and so on. All these criteria can be used to evaluate email filters. A disadvantage of these criteria is that they assign equal weights to the false positive errors and the false negative errors. They tell us little about a cost-sensitive filter's performance.

Some cost-sensitive evaluating methods have been proposed, such as the cost-ROC [6], *weighted accuracy* ($WAcc$), *weighted error rate* ($WErr$) and so on [1]. $WAcc$ and $WErr$ are computed based on an assumption that a false positive costs λ times as a false negative:

$$WAcc = \frac{\lambda \cdot TP + TN}{\lambda \cdot N_l + N_s}, \quad WErr = \frac{\lambda \cdot FP + FN}{\lambda \cdot N_l + N_s} \quad (3)$$

where FP and FN are the numbers of false positive and false negative errors, TP and TN denote the numbers of legitimate and spam messages correctly treated by filters, N_l is the size of legitimate training dataset, and N_s is the total number of training spams.

The *Total Cost Ratio* (TCR) defined by:

$$TCR = \frac{N_s}{\lambda \cdot FP + FN}, \quad (4)$$

is another cost-sensitive measure. A filter with better performance will have a greater TCR value.

Although $WAcc$, $WErr$ and TCR are effective for comparing the performance of multiple filters, they cannot properly reflect the performance of a single filter. They treat each legitimate message as λ messages. Consequently, when we wonder how many legitimate emails were mislabeled as spam and how many opposite errors occur, we cannot get answers from the above criteria directly. Hence, we proposed three new evaluation metrics [18], *error rejecting rate* (ERR), *error accepting rate* (EAR), and *total error cost* (TEC):

$$ERR = \frac{FP}{TP + FP}, \quad EAR = \frac{FN}{TN + FN} \quad (5)$$

$$TEC = c(0, 1) \cdot P(c_0) \cdot ERR + c(1, 0) \cdot P(c_1) \cdot EAR, \quad (6)$$

where $P(c_0)$ and $P(c_1)$ are the probabilities of an arbitrary email being spam and legitimate, respectively.

A good filter should have very low ERR value (zero is the best) and has a little high but acceptable EAR value (lower is better). When values of ERR and EAR of two filters are close, it is still difficult to determine which one is better. In order to address this problem and carry out cost-sensitive comparison, we also introduce TEC , a cost-sensitive criterion, which allows us to know the "total cost" caused by two kinds of errors. In general, since $c(0, 1)$ is larger than $c(1, 0)$, the effect on the TEC of the ERR is more than the EAR . To reflect the fact that more errors will occur in a large dataset, we respectively multiply

$P(c_0)$ and $P(c_1)$ by two components in Eq. (6). In conclusion, an excellent cost-sensitive filter should have the following features:

- 1) *ERR* value is very low (zero is the best),
- 2) *EAR* value is a little high, but it should be in an acceptable range, and
- 3) *TEC* value is as low as possible.

To compare the robustness of cost-sensitive email filters, a criterion suggested in [34] is given by:

$$r_{\ell_c} = \frac{cost_{\ell_c}}{\max_i cost_i} \quad (7)$$

where $cost_{\ell_c}$ is the average cost of the algorithm ℓ_c and $\max_i cost_i$ denotes the largest average cost among all the compared methods. The smaller the value of r_{ℓ_c} is, the better the performance of the algorithm is. Furthermore, when there are multiple test corpuses, the smaller the value of sum over all datasets is, the better the robustness of the method is.

2.2 A Novel Filtering Framework Based on Ensemble Learning

As mentioned earlier, there are two main steps for constructing a filter: feature subset selection and training the filter. Only when both of the two processes reach the best effect, the filter can obtain the best performance. It may be difficult for these two steps to achieve the best performance at the same time. Recently, combining multiple filters or classifiers is a popular way to produce effective and efficient filters [18,20]. In what follows, we propose a novel filtering framework called Two-Phase Cost-Sensitive Ensemble Learning (TPCSL).

The TPCSL is an instance of Two-Phase Ensemble Learning (TPEL). Algorithm 1 shows an overview of TPEL. It includes two types of learning. The first one is called *direct learning* (see Step 1 in Algorithm 1), while the other one is *indirect learning* (see Steps 3 and 4 in Algorithm 1). Direct learning denotes the process that *directly* learns multiple homogeneous or heterogeneous filters from a training dataset. After that, to some extent, each of these filters holds the **knowledge** for detecting spam and legitimate emails. When these filters inform their decisions to a committee, the committee can distinguish spam and legitimate. That is, the committee's decision is determined by the voting filters. We call such a process the *indirect learning*, as shown in Step 4 in Algorithm 1.

When one or more cost-sensitive techniques are used in the phases of *direct learning* or/and *indirect learning* of TPEL, we get the TPCSL. In the *direct learning* phase, we can use cost-sensitive techniques, such as oversampling/undersampling in each training subset D_i according to a cost matrix, or the cost-sensitive algorithm to learn each ensemble filter. In the *indirect learning* phase, K and K_j can be represented in a cost-sensitive way. In addition, K_j can be estimated with oversampling or undersampling techniques. A cost-sensitive learning method can be used to learn a committee. Thus, the TPCSL is a very flexible framework.

The difference between *direct learning* and *indirect learning* is that the former directly learns knowledge (i.e., T models $\tilde{h}_1, \dots, \tilde{h}_T$) from the data, while the

Data: T is the count of ensembled filters, M is the times of *indirect learning*, Q_i is the count of ensembled committees learned in the i th *indirect learning*, x is an instance, $E_{ij}(x)$ ($j=1, 2, \dots, Q_i$) is the j th ensemble function constructed in the i th round

Result: $E(x)$: the final ensemble function

```

1. for  $i = 1$  to  $T$  do
    | 1.1 Sampling from  $D$  with replacement to get  $D_i$ ;
    | 1.2. Learning  $\tilde{h}_i$  from  $D_i$ ;
end
2. Constructing knowledge  $K$  of  $\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_T$ ;
3. if  $M == 1$  then
    | 3.1 Learning an ensemble  $E(x)$  function from  $K$ ;
    | 3.2 Return  $E(x)$ ;
end
4. for  $i = 1$  to  $M$  do
    | 4.1 for  $j = 1$  to  $Q_i$  do
    |   | Sampling from  $K$  to get sub-knowledge  $K_j$ ;
    |   | Learning an Ensemble  $E_{ij}(x)$  from  $K_j$ ;
    | end
    | 4.2  $K =$  Constructing the knowledge of  $E_{i1}(x), \dots, E_{iQ_i}(x)$ ;
end
5. Learning an ensemble  $E(x)$  function from  $K$ ;
6. Return  $E(x)$ ;

```

Algorithm 1. The TPEL Algorithm

latter learns knowledge (i.e., an ensemble function $E(x)$) from $\tilde{h}s = \{\tilde{h}_1, \dots, \tilde{h}_T\}$. The key problem is how to represent the knowledge obtained from each model. A method is described as follows. With the knowledge of $\tilde{h}_i$ ($i = 1, \dots, T$), one may tell the label or class probability of each training example. For a hard filter $\tilde{h}_i$, we use a vector $\vec{k}_i \equiv \langle c_{i1}, c_{i2}, \dots, c_{i|D|} \rangle$ to denote its own knowledge, where c_{ij} ($j=1, 2, \dots, |D|$) is the class obtained from $\tilde{h}_i$ for the j th training instance, $|D|$ is the number of training examples. If $\tilde{h}_i$ is a soft filter, there are two possible methods for representing knowledge. First, we use a vector $\vec{k}_i \equiv \langle P_{i1}, P_{i2}, \dots, P_{i|D|} \rangle$ to denote the knowledge obtained from the model $\tilde{h}_i$, where P_{ij} ($j=1, 2, \dots, |D|$) is the output from $\tilde{h}_i$ and is the class probability (or function value) of the true class of the j th training instance. Second, the soft filter can be converted to a hard one. For example, suppose that x is the j th training example, its true class is c_1 , and the output of $\tilde{h}_i$ is the class probability $P_{ij}(c_1|x)$ for x . If $P_{ij}(c_1|x) > (P_{ij}(c_0|x) + \alpha)$, then $\tilde{h}_i$ labels x as c_1 , else it is classified into c_0 . Here, α is a threshold assigned by a user. If we want to adopt a cost-sensitive method to represent the knowledge of $\tilde{h}_i$, then set $\alpha > 0$, else let α be 0. Thus, $K \equiv [\vec{k}_1^t, \dots, \vec{k}_T^t]$ can be viewed as the knowledge obtained from all the models, where $\vec{k}_i^t$ is the transpose of $\vec{k}_i$ ($i=1, \dots, T$). In order to apply a supervised learning method to learn the $E(x)$ from K , we redefine the matrix K as $[\vec{k}_1^t, \dots, \vec{k}_T^t, \vec{k}_{(T+1)}^t]$, where $\vec{k}_{(T+1)}^t$ stores the real and corresponding labels of training examples.

From the knowledge, the algorithm used in the *indirect learning* process can learn a committee. The next subsection provides a case study of TPCSL.

2.3 A Case of TPCSL: Combining Multiple Filters Based on GMM

This subsection gives an instance of TPCSL named GMM-filter [18]. The GMM-Filter consists of two main phases: training and filtering. Training in the GMM-Filter includes four main steps: learning multiple filters, constructing the knowledge with respect to filters, using correspondence analysis to reduce the space complexity and noise related to the knowledge, and learning a committee from the knowledge. The training phase is divided into the following steps: a) dividing a training dataset into T partitions, b) training the k th NB filter on the k th subset ($k=1, 2, \dots, T$), c) constructing a **training matrix** (i.e., **knowledge**), d) using correspondence analysis on the training matrix to generate two distributions: legitimate distribution and spam distribution, and e) learning GMM-filter from the distribution. The a)~d) steps refer to the direct learning process, while e) is the indirect learning process. *The distributions are another kind of knowledge of combined filters.* Thus, learning a committee can be denoted by a Gaussian Mixed Model (GMM) from the above two distributions. When a new email arrives, the GMM is used to compute the probability of legitimate and spam, respectively.

The GMM-Filter adopts the following way for knowledge representation. Suppose a matrix $TM_{(N \times 2(T+1))}$ is the knowledge with respect to legitimate or spam of the combined filter, N is the total count of training emails. The i th line is used to represent the i th training email, which mainly reflects the performance of each filter on the i th training email. Let the i th row vector in TM be $\bar{v}_i$, and $\bar{v}_i = \langle P_{i,1}, P_{i,2}, \dots, P_{i,2T-1}, P_{i,2T}, P_{i,2T+1}, P_{i,2T+2} \rangle$, where $P_{i,2k-1}$ ($k = 1, \dots, T$) is the posterior probability of the i th training email belonging to spam computed by the k th filter, $P_{i,2k} = 1 - P_{i,2k-1}$; if the i th training email is a spam, then $P_{i,2T+1} = 1$, else $P_{i,2T+1} = 0$; if the i th training email is legitimate, then $P_{i,2T+1} = 0$, else $P_{i,2T+1} = 1$.

After the knowledge with respect to T filters is constructed, the GMM-filter uses correspondence analysis to reduce the dimensions and eliminate the noise in the new training dataset.

Correspondence Analysis of Training Matrix. Correspondence analysis [13] provides a useful tool for analyzing associations between rows and columns in a contingency table. A contingency table is a two-entry frequency table where the joint frequencies of two variables are reported. For instance, a (2×2) table could be formed by observing from a sample of n individuals with two variables: the individual's gender and whether the individual smokes, and their joint frequencies are reported. The main idea of correspondence analysis is to develop simple indices that show the relationship between the row and column categories. These indices tell us simultaneously which column category has a bigger weight in a row category and vice-versa. Correspondence analysis is also related to the issue of reducing the dimension of a data table.

For a training matrix TM , the calculation of correspondence analysis may be divided into three stages (see Algorithm 2). The first stage consists of some preprocessing calculations performed on $TM_{I \times J}$ ($I = N, J = 2(T + 1)$) which leads to the standardized residual matrix, S . In the second stage, a singular value decomposition (SVD) is performed on S to redefine it in terms of three matrices: $U_{I \times L}$, $\sum_{L \times L}$, and $V_{L \times J}$, where $L = \min(I - 1, J - 1)$. In the third stage, U , $\sum$, V are used to determine $Y_{I \times L}$ and $Z_{J \times L}$, the coordinates of the rows and columns of TM , respectively, in the new space, where $Y_{I \times L}$ is principal coordinates of rows, and $Z_{J \times L}$ is principal coordinates of columns. Note that not all L dimensions are necessary. Hence, we can reduce the dimension to $\tilde{L}$, with some loss of information. Definition 1 is used to reflect the degree of information loss.

Data: TM

Result: $U, \sum, V, Y$;

1. $\text{sum} = \sum_{i=1}^I \sum_{j=1}^J tm_{i,j}$;
2. $P = (1/\text{sum})TM$;
3. $r = \langle r_1, r_2, \dots, r_I \rangle$, $r_i = P_{i \cdot}$ ($i = 1, 2, \dots, I$);
4. $c = \langle c_1, c_2, \dots, c_J \rangle$, $c_i = P_{\cdot i}$ ($i = 1, 2, \dots, J$);
5. $D_r = \text{diag}(r_1, r_2, \dots, r_I)$, $D_c = \text{diag}(c_1, c_2, \dots, c_J)$;
6. $\tilde{P} = D_r^{-1/2}(P - rc^t)D_c^{-1/2}$;
7. $\tilde{P} = U \sum V^t$;
8. $Y = D_r^{-1/2}U \sum$;
9. $Z = D_c^{-1/2}V \sum$;
10. RETURN $U, \sum, V, Y$;

Algorithm 2. The CA Algorithm in GMM-filter

Definition 1. *Information Loss (IL) is defined as follows:*

$$IL = 1 - \sum_{i=1}^{\tilde{L}} \lambda_i / \sum_{i=1}^L \lambda_i \quad (8)$$

where, λ_i is the diagonal element in $\sum$.

According to Definition 1, given a value of IL , the $\tilde{L}$ can be computed. Then, we use $Y_{I \times \tilde{L}}$ denoted by Y^* to replace $Y_{I \times L}$ below. In the new geometric representation, rows $\overline{y}_i$ and $\overline{y}_j^*(i, j = 1, 2, \dots, T)$ in Y^* , corresponding to rows i and j in TM . All data representing training mails of spam in Y^* form a distribution in the $\tilde{L}$ dimensions space, and data about legitimate training emails form another distribution. We can learn GMM from the two distributions by the method to be described in the next subsection.

Learning GMM with EM Algorithm. The use of GMM is motivated by the capability of Gaussian mixtures to model arbitrary densities. A Gaussian mixture density is a weighted sum of M component densities, as given by the equation:

$$P(c_i|\vec{x}; \lambda_i) = \sum_{j=1}^M w_{ij} P_{ij}(\vec{x}) \quad (9)$$

where c_i ($i=0, 1$) is the categories of “spam” and “legitimate”, respectively, $\vec{x}$ is a vector of an instance, $P_{ij}(\vec{x})$ ($j=1, \dots, M$) is the component density, w_{ij} is the mixture weight, and $\sum_{j=1}^M w_{ij} = 1$.

Each component density is a L -variate Gaussian function, i.e., $P_{ij}(\vec{x}) \sim N_L(\mu_{ij}, \nu_{ij})$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation $\lambda_i = \{w_{ij}, \mu_{ij}, \nu_{ij}\}$ ($i=0, 1; j=1, \dots, M$). Given the training data of c_i , the goal of model training is to estimate the parameters of the GMM, λ_i , which in some sense best matches the distribution of the training feature vectors. By far the most popular and well-established method is the maximum likelihood (ML) estimation. ML parameter estimates can be obtained by iteratively using a special case of the expectation-maximization (EM) algorithm. The basic idea of the EM is, beginning with an initial model λ_i , to estimate a new model λ'_i . The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached. There are two main steps in EM [7], described as follows.

E step, computing

$$h_j^{(k)}(\vec{x}_n) = \{w_{ij}^{(k)} P_{ij}^{(k)}(\vec{x}_n)\} / \{\sum_{j=1}^M w_{ij}^{(k)} P_{ij}^{(k)}\}. \quad (10)$$

M step, updating parameters according to following three formulas

$$w_{ij}^{(k+1)} = \frac{1}{N} \sum_{n=1}^N h_j^{(k)}(\vec{x}_n), \quad (11)$$

$$\mu_{ij}^{(k+1)} = \{\sum_{n=1}^N h_j^{(k)}(\vec{x}_n) \vec{x}_n\} / \sum_{n=1}^N h_j^{(k)}(\vec{x}_n), \quad (12)$$

$$\nu_{ij}^{(k+1)} = \sum_{n=1}^N h_j^{(k)}(\vec{x}_n) [\vec{x}_n - \mu_{ij}^{(k+1)}][\vec{x}_n - \mu_{ij}^{(k+1)}]^T / \{\sum_{n=1}^N h_j^{(k)}(\vec{x}_n) \vec{x}_n\}, \quad (13)$$

where $N=|c_0|$ when we estimate a spam model λ_0 , and if we estimate a model λ_1 for legitimate, then $N=|c_1|$.

Filtering Method in the GMM-Filter. For simplification, the following discussion only considers combining multiple Naive Bayes (NB) filters. When a new email e arrives, $\vec{v}_0$ and $\vec{v}_1$ are constructed based on the computation on T filters, where $\vec{v}_0 = \langle P_1, 1-P_1, \dots, P_T, 1-P_T, 1, 0 \rangle$, $\vec{v}_1 = \langle P_1, 1-P_1, \dots, P_T, 1-P_T, 0, 1 \rangle$, and P_i is the posteriori probability of e for spam computed by the

i th filter. Furthermore, we normalize $\vec{v}_i$ ($i=0, 1$): $\vec{v}_i = \vec{v}_i^* / (T + 1)$, and then transform $\vec{v}_i^*$ into the $\tilde{L}$ -dimensional space by using the following equation.

$$\vec{x}_i = \vec{v}_i^* Z \Sigma^{-1}. \quad (14)$$

Next, we can compute $P(c_0|x_0, \lambda_0)$ and $P(c_1|x_1, \lambda_1)$, as well as normalizing them. The former value reflects the degree that e belongs to legitimate, and the latter is the posterior probability of spam. Then, for e , we classify it into spam if and only if:

$$\frac{P(c_1/e, \lambda_1)}{P(c_0/e, \lambda_0)} \geq \frac{c(0, 1)P_0}{c(1, 0)P_1}. \quad (15)$$

In the above equation, the right hand side is a constant, which is called α .

Experimental Results of the GMM-Filter. Experiments were carried out on the Ling-Spam and PU1 corpus. Table 1 shows the distributions of testing and training of spam and legitimate emails of two corpus. The feature subset selection method used in our experiments is Information Gain (IG) [28], the ratio of IG is 1%, $T=8$. We set $K=120$ in the KNN method, the model count in GMM-Filter is 6, $IL=0.1$. And, $c(0, 1)=0.5$, $c(1, 0)=4$. The single filter in voting is NB too, the total count of filters in voting is 8.

Table 1. Distributions of testing and training emails of two corpus

	training emails count		testing emails count	
	legitimate	spam	legitimate	spam
PU1	488	384	122	96
Ling-Spam	1929	384	483	97

Figure 1(a) and 1(b) show the comparative results of five filtering algorithms on PU1 and Ling-Spam, respectively. On PU1, the GMM-Filter is the best because its values of three criteria are all the lowest. The GMM-Filter also shows ideal performance on Ling-Spam. From Figure 1(b), we can see that EJR of the GMM-Filter is 0, TEC and ERR of the GMM-Filter are very low. Although NB has the similar result to the GMM-Filter on Ling-Spam, the performance of NB is worse than the GMM-Filter on PU1.

The motivation of combining multiple filters is to reduce the effect of the factors existing in the training process. The size of a feature subset is one of the most important factors affecting the performance of filter. The experimental results investigating the effect of feature subset on the filter performance are shown in Tables 2 and 3. On all datasets, TEC of the GMM-Filter has little change when the size of feature subset ranges from 0.2% to 10%.

The experimental results show that the GMM-Filter has better performance, and it is insensitive to the ratio of feature subset selection. The GMM-Filter shows that TPCSL is an effective framework of ensemble learning.

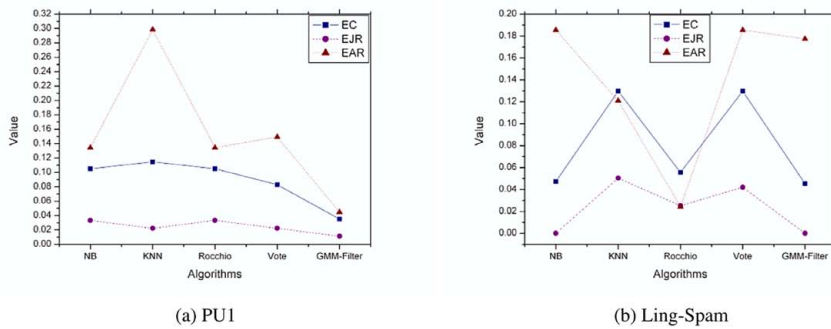


Fig. 1. Comparative results of five filtering algorithms on two corpus

Table 2. Comparison of the GMM-Filter to other filters on PU1 when the ratio of feature subset selection is changed

	NB	Rocchio	Voting	KNN	GMM-Filter
0.2%	0.157643	0.187898	0.109873	2.292994	0.030255
0.5%	0.124204	0.133758	0.078025	2.292994	0.035032
1%	0.105096	0.105096	0.082803	0.11465	0.035032
2%	0.068471	0.068471	0.085987	0.143312	0.046178
10%	0.074841	0.367834	0.08121	0.205414	0.041401

Table 3. Comparison of the GMM-Filter to other filters on Ling-Spam when the ratio of feature subset selection is changed

	NB	Rocchio	Voting	KNN	GMM-Filter
0.2%	0.046178	0.038217	0.186306	1.515924	0.033439
0.5%	0.023885	0.05414	0.184713	1.515924	0.030255
1 %	0.047325	0.055556	0.12963	0.12963	0.045267
2 %	0.035032	0.022293	0.046178	1.515924	0.041401
10 %	0.044586	0.036624	0.08758	0.207006	0.047352

This work mainly addresses and demonstrates two levels of WI technologies in the context of email. In the above sections, we discuss the *infrastructure-level spam filtering*. Below, we will depict *application-level intelligent assistance*.

3 Operable Email and Its Applications

3.1 Motivations of Operable Email

Web Intelligence (WI) [29,30,31,32,33] has been recognized as a new direction for scientific research and development to explore the fundamental roles as well as practical impacts of Artificial Intelligence (AI) and advanced Information Technology (IT) on the next generation of Web-empowered products, systems,

services, and activities. It is one of the most important IT research fields in the era of Web and agent intelligence. A long-standing goal of WI is to develop the World Wide Wisdom Web (W4). According to Liu et al. [22]: “The next paradigm shift in the WWW will lie in the keyword of wisdom. The new generation of the WWW will enable users to gain new wisdom of living, working, playing, and learning, in addition to information search and knowledge queries.” The Wisdom Web attempts to layout the ultimate dream of the intelligent Web. It focuses more on the knowledge level and intelligent Web systems for real world, complex problem solving. The Wisdom Web covers a wide spectrum of issues, such as the intelligent Web, autonomic Web support, social intelligence, intelligent-agent technology, and so on.

Among those research topics, recasting email for intelligent services is a special goal of W4. Needless to say, semantic dimension is the basis of realizing intelligent services based on email. To implement automatic applications via email, Luke and colleagues add semantic features to current email to design semantic email [23]. A Semantic Email Process (SEP) should be constructed for a semantic email task. A SEP contains three primary components: *originator*, *manager* and *participants*. A SEP is initiated by the originator. The manager may be a shared server or a program run directly by the originator. A new SEP invoked by an originator is sent to the manager. Then, the manager sends email messages to the participants. After that, the manager handles responses, and requests changes as necessary to meet the originator’s goals. The participants respond to messages received about the process. The semantic email has the greatest promise for implementing automated functions. Some applications based on such a kind of email have been developed. However, such semantic emails have several disadvantages. First, an automated task needs to be defined as a Semantic Email Process (SEP) by a trained user. Second, it needs an additional server (namely Manager) to support the functions of a SEP. Third, many bread-and-butter tasks, in fact, cannot be defined as a SEP.

Using email to implement ubiquitous social computing is another special goal of W4. As an indicator of collaboration and knowledge exchange, e-mail provides a rich source for extracting informal social network across the organization. As a result, it is a highly relevant area for research on social networks. Like the WWW that consists of websites and hyperlinks which *explicitly connect* sites, the World Social Email Network (WSEN) consists of e-mail address and communication relationship which *implicatively connect* users. With the development of other branches of computer science, the Web has gained great advances in recent years. An important point is that applications such as search engine, question-answer, e-business and so on evolve with the Web. However, on the WSEN, only a few researchers focus on developing technical foundations, physical infrastructures, software systems, applications.

Figure 2 shows a multilevel description of the Web, the Semantic Web, the WSEN and email. The goal of the Semantic Web is to create an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users [4], by establishing machine-understandable Web resources.

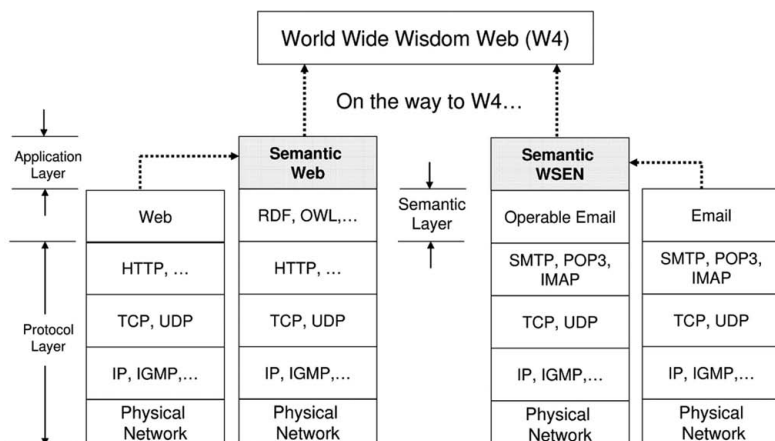


Fig. 2. A Multilevel description of the Web, the Semantic Web, the WSEN and email

To do that, researchers plan to create ontology and logic mechanisms and replace HTML with XML, RDF, OWL, and other markup languages [14]. Being independent of the Web or the Semantic Web, the WSEN aims to establish a novel application platform which promotes sharing, cooperating, asynchronously communicating, and even controlling distributed computing devices, such as PC, PDA, mobile, home electronic appliances etc. As shown in Figure 2, we plan to accomplish this by creating necessary semantic mechanisms and replacing traditional email with operable email.

In contrast to the traditional human-readable email, the operable email provides a language and an infrastructure where agents on the WSEN can automatically cooperate and deal with tasks by expressing information in an email with a machine-processable form. We summarize the basic ideas of the operable email as follows. It is a technology for enabling machines to make more sense of some specifically emails sent from an agent to another, with the result of implementing automated functions or services on the WSEN.

3.2 Issues with Operable Email Research

The operable email and intelligent applications on the WSEN present excellent opportunities and challenges for the research and development of novel and intelligent agent-based applications in the next-generation Web, as well as for exploiting business intelligence and social intelligence [16,17]. With the rapid growth of the techniques related to WI, research and development on W4 have received much attention. We expect that more attention will be focused on operable email and the WSEN in the coming years. In order to study operable email systematically, and develop advanced operable email- and agent-based intelligent applications on the WSEN, we list several major research topics pertinent to operable email below.

- **Designing PSML (Problem Solver Markup Language).** The issue deals with theories, methods, and languages for representing problems, queries, autonomous entities, knowledge etc. All these studies are essential for developing PSML with which users share files, create local knowledge database, publish information, search resource, question, even control remote devices and so on. The PSML provides a single gateway to assist people or agents to represent multifarious semantic information in a machine-understandable form.
- **Coding, decoding and executing operable email.** From sender to receivers, the operable email appears in two ways. In the first method, the operable email does not enclose any machine-understandable information. This kind of operable email serves for the human-human communications. In the other method, the operable email encloses messages represented by the PSML. In this case, PSML messages are coded at the side of the sender, and are decoded at the side of receivers. Under permissions, the decoded PSML messages are executed on the corresponding plug-ins of operable email's agent.
- **Routing mechanism for operable email.** Over the WSEN, knowledge and resource are stored with a totally distributed method. A crucial problem is to determine a route through which an operable email enclosed with PSML messages is sent to providers of sharing files and knowledge. Suppose that A denotes an agent for searcher and G is the agent for provider. From A to G , there are many routes to penetrate the operable email initiated from A , such as " $A \rightarrow B \rightarrow E \rightarrow G$ ", " $A \rightarrow C \rightarrow E \rightarrow G$ ". Thus, the first issue of this topic is how to choose an optimized route on the dynamic WSEN. The so-called dynamic WSEN refers to two aspects: the nodes on the WSEN are changing momentarily, and the trust relationship between nodes are also changing. The second issue is how to avoid to automatically forward an operable email in a circle. The third issue is to study the hop of forwarding operable emails.
- **Distrust and trust promulgation mechanism.** From cognitive and mathematical points of view, trust can be broadly classified into two types. The first type views trust as some underlying beliefs, and defines trust as a function of the value of these beliefs. The mathematical view ignores the role of underlying beliefs and uses a trust metric, based on variables like perceived competence, perceived risk, utility of a situation for the agent involved, importance of a situation etc. These models incorporate some aspects of game theory and the evolution of cooperation models. Both views see trust as a variable with a threshold for action. The agent grants access rights based on whether the trusted value of the requesting entity is above a threshold. On the WSEN, the following issues about trust should be considered: how to define the trust or distrust between two nodes; how to promulgate distrust or trust on the WSEN that is constantly changing in both size and topology; and how to implement the cooperation or delegation mechanism based on trust.

- **From operable email-based agent to intelligent applications.** As mentioned above, each node on the WSEN is designed as an agent. Each agent has the functions of traditional email clients. Moreover, it is a computational entity that is capable of making decisions on behalf of their users and self-improving their performance in dynamically changing and unpredictable task environments. Thus, this issue mainly includes the following topics: push or pull, matchmaking, collaborative work support, decision and delegation and so on.

3.3 Implementing an Assisatant (ECIPA) Using Operable Email

Brief Introduction to ECIPA. We discuss an Email-Centric Intelligent Personal Assistant (ECIPA), which is implemented based on the operable email. The main objective of an ECIPA is to design customizable and extensible agents that work together to process incoming, outgoing, and existing emails. Figure 3 shows the architecture of ECIPA. It adopts a three-tier software architecture that makes use of an effective distributed client/server design to provide increased performance, flexibility, maintainability, reusability, and scalability. It also hides the complexity of distributed processing from the user. These characteristics make the three-tier architecture an ideal choice for Internet applications and net-centric information systems. The data tier provides database management related functionalities and is dedicated to data and file services. The middle tier (i.e., the application tier) comprises multiple agents and their interactions to provide the process management, such as sending/receiving email, querying, parsing/executing the operable email. The top tier (i.e., the presentation tier) provides user services, such as session, text input, dialog, and display management.

Moreover, as shown in Figure 3, the middle tier in the ECIPA consists of a family of collaborative agents. Some of those agents are transparent to the users. In other words, they work as background agents, such as M/SA, FA, ETA, PA, EA, LA, ExA and Q/AA. On the other hand, the foreground agents are CA and IA, respectively. The functions of the agents are summarized as follows:

- Monitoring/Sending Agent (M/SA):
monitors the user's inbox and sending email;
- Filtering Agent (FA):
labels spam automatically;
- Extracting Agent (EtA):
distills information from diverse sources;
- Parsing Agent (PA):
parses and decodes an operable email;
- Running Agent (EA):
executes commands enclosed in an operable email;
- Querying/Alerting Agent (Q/AA):
answers/alerts the user;
- External Agent (ExA):
communicates with the archiving system;

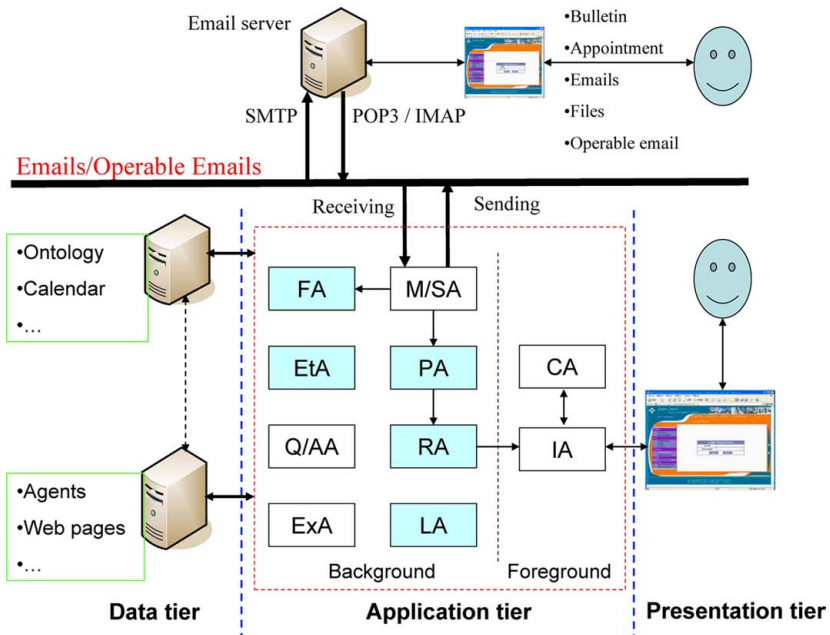


Fig. 3. The architecture of the ECIPA. FA: Filtering Agent, EtA: Extracting Agent, CA: Configuring Agent, EA: Executing Agent, Q/AA: Querying/Altering Agent, IA: Interacting Agent, ExA: External Agent, PA: Parsing Agent, M/SA: Monitoring/Sending Agent, LA: Learning Agent. Some of agents exchange information directly, denoted by the line between the two corresponding agents, and other agents communicate indirectly. The highlighted agents are described in detail in this work.

- Learning Agent (LA):
learns dynamic user behaviors;
- Configuring Agent (CA):
sets the running parameters;
- Interface Agent (IA):
provides user services.

Some of the agents communicate directly. That is, an agent *A* directly sends messages to another agent *B*. Some of them communicate indirectly. That is, an agent *B* accesses the information stored in server by an agent *A*. The agents in the ECIPA do not need a special platform to support them. The notion of “agent” is used to characterize certain autonomous, loosely coupled components that discover and communicate with each other by exchanging highly structured messages [3].

The main functions of ECIPA are summarized below:

- Intelligent cooperation based on the operable email (supported by PA and EA): For email-mediated user tasks, many of them can be carried out automatically [23]. The use of the operable email allows an ECIPA to aid its user to automatically carry out the formalized tasks enclosed in the operable email sent by another user or the user's ECIPA.
- Ontology-mediated classification, query, and archiving (supported by EtA, ExA, Q/AA): An ontology is designed to store the user's background knowledge, the information of local or global resources, emails and attachments within their context. ECIPA provides flexible functions for classifying emails into virtual folders based on the operations on concepts in ontology, as well as those for queries based on concepts.
- Sorting/responding based on dynamic user behavior learning (supported by LA): ECIPA learns dynamic user behaviors based on the time-window techniques. It prioritizes incoming messages according to the habit processing emails of a user. In addition, by finding and analyzing the user behaviors, ECIPA can identify whether the user is on vacation. ECIPA responds for the user automatically in the case of vacation.
- Automated and cost-sensitive spam filtering (supported by FA): ECIPA combines multiple Naive Bayesian filters to block unwanted emails. The filtering method is sensitive to the cost of the false positive errors and the false negative errors.

The ECIPA demonstrates that useful tasks can be accomplished by means of agent interaction. The key techniques adopted by ECIPA are very useful for developing other intelligent applications on the Web. For example, the dynamic behavior learning approach based on time-window can be adopted in other personalizing recommendation system on the Web. Although the architecture described is developed as a three tiers structure, it can be easily implemented in other ways.

To appreciate the usefulness of ECIPA, one may image the following scenario. The intelligence of ECIPA is reflected by the process of scanning or tracking emails, analyzing user behaviors, filtering, finishing tasks automatically, etc. Some or all of those tasks are performed manually before. Upon registration, a user uses a corresponding email address and proper password to log in the ECIPA. The personalized interface is shown in the IA. On the user's homepage, the sorted incoming emails, the urgent items, the summary and digest of the tasks automatically completed by the EA etc. are given. When the user is operating on the client side of ECIPA, the IA captures and records some of its user's operations. The LA learns dynamic behaviors from those operated records passed by IA. The M/SA is running on the server side, it can monitor the new incoming messages for all the users in the ECIPA. When a new traditional email is found by the M/SA, it is passed to the FA. After the FA labels this email, the EtA extracts information from this message, and then store them into an ontology. If the new email is an operable email, the M/SA informs the PA to

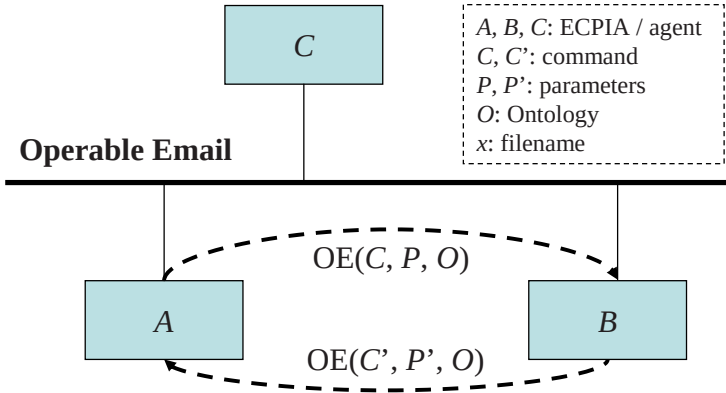


Fig. 4. The model of the operable email

parse it. The EA gets it from the PA. Furthermore, EA executes the commands enclosed in the operable email under the permission of the user.

Usage of Operable Email in the ECIPA. The ECIPA supports two kinds of email-mediated tasks: automated tasks and non-automated tasks. The implementation of the first kind is supported by the operable email. Figure 4 presents the working model of the operable email. As shown in this figure, if *A* wants to “operate” *B*, *A* should send an operable email that encloses a command with corresponding parameter *P*. When *B* receives the message, it parses the command and automatically replies *A* by another operable email. This model shows that the *B*’s user is released from some of the manual operations, such as reading email, finding file, attaching file, and replying email.

The operable email opens the door for implementing a wide range of email-mediated applications with automated functions of response that are infeasible today. These automated processes brought by the operable email offer tangible productivity. Below, we describe how an ECIPA can explore the operable email for automatic processing of very common email tasks, and illustrate the key ideas through some examples:

- Publishing and managing bulletin. Suppose that you send a message through the operable email. With suitable semantics attached to the email, the operable email can result in automatically extracting and posting the announcement to a Web calendar, and sending reminders a day before the event.
- Making appointment automatically. Imagine that you are making an appointment with a user. Currently, you must check your calendar and reply manually. While, if the user makes the appointment through an operable email, your ECIPA does the above work for you automatically. The ECIPA informs you the appointment in advance.
- Negotiating the schedule of a meeting. Suppose that you are organizing a meeting and you want to hold it when most of the members are free. In

tradition, you must ask the members through emails one by one, then compile the replies manually. With the operable email, your ECIPA can automatically negotiate with the members. At last, it reports the round-table result to you. Furthermore, under your permission, your assistant automatically informs each member the final time of the meeting.

- Sharing files. Imagine that you are a team leader and tired of the frequent requests for sending the project documents through emails. The ECIPA provides two ways for sharing files between the groups. The first one is “push”. Namely, after you set a new sharing for the members in a given contact list, the ECIPA sends the table of sharing files to the members in that list. Thus, a member can use the “download” command to request your ECIPA to automatically send a file in the sharing table. Another one is “pull”. Before a file is downloaded from your ECIPA, a user should use the “listfile” command to get the file table which shows files the user can access. Then, the user sends an operable email which encloses the “download” command to your ECIPA. After that, the EA in your ECIPA responds the user automatically.

Using a special field in the header of an operable email, the M/SA in the ECIPA can distinguish an operable email from a traditional email. The content in an operable email is generated by the ECIPA (or written by users) according to the syntax shown in Table 4.

Table 4. The command syntax of the operable email in BNF

$\langle \text{message} \rangle :: = \langle \text{command} \rangle (\langle \text{blank} \rangle \langle \text{para-value} \rangle)^*$
$\langle \text{command} \rangle :: = \langle \text{identifier} \rangle$
$\langle \text{para-value} \rangle :: = \langle \text{paraName} \rangle \langle \text{blank} \rangle \langle \text{value} \rangle$
$\langle \text{paraName} \rangle :: = \langle \text{identifier} \rangle$
$\langle \text{value} \rangle :: = (\langle \text{ascii} \rangle)^*$
$\langle \text{identifier} \rangle :: = \langle \text{alphabetic} \rangle (\langle \text{alphabetic} \rangle \langle \text{numeric} \rangle)^*$
$\langle \text{blank} \rangle :: = \langle \text{whitespace} \rangle (\langle \text{TAB} \rangle \langle \text{ENTER} \rangle)^+$
“*” indicates any number of occurrences.
“+” means that the number of occurrences should be greater than 1.

To support the automated tasks in the ECIPA, we define some commands as shown in Table 5. Due to the limited space of this chapter, the parameters of those commands are omitted. To simplify the parsing process, we represent each script of the commands as an XML document according to a scheme of the body of operable email before sending [16]. In other words, the ECIPA encodes message-as-XML documents into operable emails, and decodes the message-as-XML documents back into messages that represent the commands as shown in Table 5.

Hence, we can see that the operable email provides an email-based communication means, facilitated by an assistant of the users. This means that the operable email supports an email-based agent infrastructure where agents can automatically deal with some tasks that are impracticable currently. There are

several reasons why using the operable email as a communication media for “bear the weight” of the communication script. Email clients are lightweight and available on most computational devices. Emails are peer-to-peer and symmetric communication protocols. An email-based agent communication channel does not need a router component. Again, firewalls are not an issue for email-based communication. Agents on either side of a firewall can communicate more easily with an email-based infrastructure than with a TCP/IP infrastructure. Finally, the ISP mailbox acts as a message queuing facility, obviating the need for a specialized message queuing component. Consequently, the operable email or its transmutation developed in future is a very useful tool for implementing WI applications on the next generation Web.

Table 5. The main commands used in the ECIPA

Command	Meaning
download	Download a file from another assistant
listfile	Ask the receiver to list the sharing files the sender can access
sendfile	Ask the receiver to send a file
appointment	Make an appointment with the receiver at a given time
meeting	Hold a meeting with the receivers at a given time
bulletin	Publish a bulletin to receivers
subscribe	Subscribe a piece of information from the receiver
ask	Ask the receiver about something

In practice, operable emails can be generated by either a user or a program. The user-generated case is tedious, time-consuming, and error-prone. Thus, in the ECIPA, we adopt the latter to form operable emails for the user according to the input in the IA of the assistant.

4 Conclusions

Email filtering and email-mediated intelligent applications have been discussed in this chapter. A two-phase cost-sensitive ensemble learning (TPCSL) framework for email filtering is given. The TPCSL consists of two types of learning: *direct learning* which refers to learn one or multiple filters from training dataset directly; *indirect learning* refers to the process of constructing a committee from the knowledge of multiple filters. An instance of TPCSL named GMM-Filter is studied. The GMM-Filter includes four main steps: learning multiple filters, constructing the knowledge of those filters, using correspondence analysis to reduce the space complexity and noise related to the knowledge, and learning a committee from the knowledge. The experimental results show that the GMM-Filter achieves a better performance, and it is insensitive to the ratio of feature subset selection. A very promising research topic, *email intelligence*, is given. In our context, email intelligence refers to the automated applications or functions provided by email. The traditional email without semantic cannot support such

a need. To bridge this gap, we propose the concept of operable email. Although the design of operable email is not depicted in detail, an illustrative case using operable email is provided.

Acknowledgments

This work is partially supported by the NSFC major research program: “Basic Theory and Core Techniques of Non-Canonical Knowledge” (NO. 60496322), NSFC research program (NO. 60673015), the Open Foundation of Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, the Project (NO. 07213507D and 06213558) of Dept. of Science and Technology of Hebei Province, and the Project (NO. Y200606) of Shijiazhuang University of Economics.

References

1. Androutsopoulos, I., Koutsias, J., Chandrinou, K.V., Spyropoulos, C.D.: An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with encrypted personal e-mail messages. In: Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp. 160–167 (2000)
2. Androutsopoulos, I., Georgios, P., Michelakis, E.: Learning to filter unsolicited commercial e-mail. Technical Report 2004/2, NCSR Demokritos00 (2004)
3. Bergman, R., Griss, M., Staelin, C.: A personal email assistant. Technical Report HPL-2002-236, HP Labs Palo Alto (2002), <http://citeseer.ist.psu.edu/bergman02personal.html>
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* 284(5), 34–43 (2001)
5. Boykin, P.O., Roychowdhury, V.: Personal email networks: an effective anti-spam tool. *IEEE Computer* 38(4), 61–68 (2005)
6. Chris, D., Robert, C.H.: Cost curves: an improved method for visualizing classifier performance. *Machine Learning* 65(1), 95–130 (2006)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society* 39(1), 1–38 (1977)
8. Deng, Y.H., Tsai, T.H., Hsu, J.: P@rty: a personal email agent. In: Proc. of Agent Technology Workshop, pp. 61–64 (1999)
9. Drucker, H., Wu, D., Vapnik, V.N.: Support vector machines for spam categorization. *IEEE Transactions on Neural Networks* 20(5), 1048–1054 (1999)
10. Duda, R.O., Hart, P.E.: *Pattern Classification and Scene Analysis* (1973)
11. Fawcett, T.: In vivo spam filtering: a challenge problem for data mining. *KDD Explorations* 5(2), 140–148 (2003)
12. Ho, V., Wobcke, W., Compton, P.: EMMA: an email management assistant. In: Proc. of 2003 IEEE/WIC International Conference on Intelligent Agent Technology (IAT 2003), pp. 67–74 (2003)
13. Hardle, W., Simar, L.: *Applied Multivariate Statistical Analysis*, 341–357 (2003)

14. Hendler, J.: Agents and the Semantic Web. *IEEE Intelligent Systems* 16(2), 30–37 (2001)
15. Jason, D.M., Rennie, J.: ifile: an application of machine learning to e-mail filtering. In: *Proc. of the KDD-2000 Text Mining Workshop*, pp. 95–98 (2000)
16. Li, W.B., Zhong, N., Liu, J.M., Yao, Y.Y., Liu, C.N.: A perspective on global email networks. In: *Proc. of 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006)*, pp. 117–120 (2006)
17. Li, W.B., Zhong, N., Yao, Y.Y., Liu, J.M., Liu, C.N.: Developing intelligent applications in social e-mail networks. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) *RSCTC 2006. LNCS (LNAI)*, vol. 4259, pp. 776–785. Springer, Heidelberg (2006)
18. Li, W.B., Liu, C.N., Chen, Y.Y.: Combining multiple email filters of naive Bayes based on GMM. *Acta Electronica Sinica* 34(2), 247–251 (2006)
19. Li, W.B., Zhong, N., Liu, C.N.: Design and implementation of an email classifier. In: *Proc. of International Conference on Active Media Technology (AMT 2003)*, pp. 423–430 (2003)
20. Li, W.B., Zhong, N., Liu, C.N.: Combining multiple email filters based on multivariate statistical analysis. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) *ISMIS 2006. LNCS (LNAI)*, vol. 4203, pp. 729–738. Springer, Heidelberg (2006)
21. Li, W.B., Zhong, N., Liu, C.N.: ECPIA: An email-centric personal intelligent assistant. In: Wang, G.-Y., Peters, J.F., Skowron, A., Yao, Y. (eds.) *RSKT 2006. LNCS (LNAI)*, vol. 4062, pp. 502–509. Springer, Heidelberg (2006)
22. Liu, J.M.: Web intelligence (WI): What makes Wisdom Web? In: *Proc. of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, pp. 1596–1601 (2003)
23. McDowell, L., Etzioni, O., Halevy, A., Henry, L.: Semantic email. In: *Proc. of the Thirteenth Int. WWW Conference (WWW 2004)* (2004)
24. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In: *Proc. of the AAAI-98 Workshop on Learning for Text Categorization*, pp. 55–62 (1998)
25. Salton, G.: *Automatic text processing: the transformation, analysis, and retrieval of information by computer* (1989)
26. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
27. Sun, D., Tran, Q.A., Duan, H., Zhang, G.: A novel method for Chinese spam detection based on one-class support vector machine. *Journal of Information and Computational Science* 2(1), 109–114 (2005)
28. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: *Proc. of 14th International Conference on Machine Learning (ICML 1997)*, pp. 412–420 (1997)
29. Zhong, N.: Developing intelligent portals by using WI technologies. In: *Proc. of Wavelet Analysis and Its Applications, and Active Media Technology (AMT 2004)*, pp. 555–567 (2004)
30. Zhong, N., Liu, J.M.: The alchemy of intelligent IT (iIT): blueprint for future of information technology. In: *Intelligent Technologies for Information Analysis*, Springer Monograph, pp. 1–16 (2004)
31. Zhong, N., Ohara, H., Iwasaki, T., Yao, Y.Y.: Using WI technology to develop intelligent enterprise portals. In: *Proc. of International Workshop on Applications, Products and Services of Web-based Support Systems*, pp. 83–90 (2003)

32. Zhong, N., Liu, J.M., Yao, Y.Y.: In search of the Wisdom Web. In: IEEE Computer, pp. 27–31 (2002)
33. Zhong, N., Liu, J.M., Yao, Y.Y.: Envisioning intelligent Information Technologies (iT) from the stand-point of Web Intelligence (WI). Communications of the ACM 50(3), 89–94 (2007)
34. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transaction on Knowledge and Data Engineering 18(1), 63–77 (2005)

Ontology Based Web Mining for Information Gathering

Yuefeng Li¹ and Ning Zhong²

¹ School of Software Engineering and Data Communications
Queensland University of Technology, Brisbane QLD 4001, Australia
y2.li@qut.edu.au

² Department of Life Science and Informatics
Maebashi Institute of Technology, Maebashi 371-0816, Japan
zhong@maebashi-it.ac.jp

Abstract. There exists a gap between Web mining and the effectiveness of using Web data. The main reason is that we cannot simply utilize and maintain the discovered knowledge using the traditional knowledge-based techniques due to the huge amount of discovered patterns, many noise in discovered patterns and even some useful patterns with uncertainties. In this chapter we discuss ontology-based problem solving approaches for building a bridge between Web mining and the effectiveness of using Web data, which tend to automatically construct and maintain ontologies for representations, application and updating of discovered knowledge. We mainly discuss two models: the pattern taxonomy model and the ontology mining model. The former uses the up-to-date techniques of association mining and latter uses granule mining that directly discovers granules rather than patterns.

1 Introduction

We have witnessed an explosive growth of the available information on the Web over the last decade. However, there are two fundamental issues regarding the effectiveness of Web information gathering: mismatch and overload. The mismatch means some useful and interesting data has been missed out, whereas, the overload means some gathered data is not what users want.

Although information retrieval (IR) based techniques have touched fundamental issues [3,15], IR-based systems neither explicitly describe how the systems act like users nor discover interesting and useful knowledge from very large datasets to answer what users really want. This issue has challenged the artificial intelligence (AI) community to address “what has information gathering to do with AI” [20,25]. For a short while, many intelligent information agents have been presented for this challenge. Unfortunately, information agents can only show us the architectures of Web information gathering [22,23,24,27]. They have not provided more significant contributions for finding interesting and useful knowledge from Web data.

Web Intelligence (WI) is an alternative way that can provide a new thought for solving this problem [67,68,66,44,72,74]. Currently, there are three main directions for the effectiveness of using Web data in WI: Web mining, adaptive Web systems and information foraging agents [71].

The application of data mining techniques to Web data, called Web mining, is used to discover knowledge (patterns) from Web data. Currently, a Web mining system can be viewed as the use of data mining techniques to automatically retrieve, extract, generalize, and analyze information on the Web [7] [53]. Web mining can be classified into four categories: Web usage mining, Web structure mining, Web user profile mining, and Web content mining [14,47,28,59].

An adaptive Web system [56] is able to identify the interrelationships among distributed electronic information on the Web based on the discovery of Web mining [5]. The adaptive Web systems can help to facilitate or personalize Web surfing operations. It is another version of Web usage mining. An information foraging agent is able to discover user navigation strategies during Web surfing operations based on empirical regularities observed from Web log data [45]. It takes one more step further to adaptive Web systems.

The research on adaptive Web and information foraging agents provides basic problem solving approaches for a particular area, Web surfing operation. In this chapter, we discuss a more general Web-based problem solving approach “ontology-based Web mining” for Web information gathering.

It is indubitable that the existing Web mining models can return numerous discovered patterns (knowledge) from Web data. However, there is a gap between the effectiveness of using Web data and Web mining because the traditional knowledge-based techniques cannot be simply transplanted into this area. The first reason is too many discovered patterns available. The second reason is that there exists many noise patterns in the set of discovered patterns [62]. Another reason is that some useful patterns might include uncertainties when they are automatically extracted from the Web data [30,35]. These problems prevent the effectively application of discovered knowledge for Web information gathering.

In this chapter, we discuss two methods that based on using of ontology-based approaches for Web information gathering: the pattern taxonomy model (PTM) and ontology mining technique. PTM [62] is a pattern taxonomy, a tree-like structure, that illustrates the “*is_a*” relation between discovered patterns. It also provides a suitable structure for pruning some noise patterns, for example, the non-closed patterns. The usage of discovered patterns in PTM is based on the traditional measures of patterns.

For the ontology mining technique [35], syntactically an ontology consists of two parts: the top backbone and the base backbone. The former illustrates the linkage between discovered patterns. The latter illustrates the linkage between terms and patterns. Different to PTM, a pattern in the ontology mining model is a representation of a group of objects rather than a set of terms. The initial ontology can be automatically built according to the set of discovered patterns. A mathematical model, called the association set, was set up to represent the correlation between patterns. The ontology can also be updated based on the user feedback.

The remainder of the chapter is structured as follows. We begin by introducing the problem about Web information gathering in Section 2. An outline of the architecture of ontology based Web data mining is presented in Section 3. The basic methods of text mining, association mining and decision tables is introduced in Section 4. Section 5 provides the structures of representation of discovered knowledge in both PTM and the

ontology mining technique, and Section 6 describes data mining models based on the structures of knowledge representation. Furthermore, we discuss knowledge evolution in Section 7 and related work in Section 8, respectively. Finally, the concluding remarks is given in Section 9.

2 The Problem

A major challenge in WI is to build effective and efficient communications between the Web and Web users. Currently most search engines can only use queries rather than Web user profiles due to the difficulty of automatically acquiring Web user profiles. The first reason for this is that Web users may not know how to represent the topics with respect to their own interests. The second reason is that Web users may not wish to invest a great deal of effort to dig out a few relevant pages from hundreds of thousands of candidates provided by search engines.

Web user profiles were classified into two diagrams [35], the data diagram and the information diagram. The former diagram is the discovery of interesting registration data and customer profile portfolios. In general, the registration data or customer profile portfolio can be described as a database or a set of transactions, e.g., user log data. The meaning of data values in each record (or transaction) is understandable. The latter diagram is the discovery of interesting topics for Web user information needs. Compared to the data diagram, there are two significant differences on the data:

- There are many duplicates in the data;
- The meaning of data values (terms) is ambiguous since there may exist “synonymy” or “hyponymy” relations between terms.

The difficult problem related to information diagram is how to describe what users want. Usually, users themselves are easily acquainted with interesting Web pages while they read through contents of the Web pages. The rationale behind this is that the users implicitly use a concept model (a kind of ontology) based on their knowledge about a specified topic; even though they do not know how to represent it.

Based on the traditional knowledge acquiring method, we may ask Web users to provide descriptions and narratives for the topics with respect to their own interests while we try to represent user profiles. It is difficult, however, for general users to write adequate descriptions and narratives. Although linguists can provide tolerable descriptions and suitable narratives, the corresponding concept model is still incomplete. First, the linguists may ignore some important terms. Dictionaries usually are not very useful for expanding the set of terms since we do not know authors’ writing styles. It is also quite often that the linguists and the dictionaries may ignore some relationships between subtopics [35].

Another alternative way is to ask users to provide a training set (users feedback) in which positive documents are the relevant information and negative ones are non-relevant information. Therefore, the main objective of describing what users want is to understand the discovered knowledge in the training set and apply them to respond to what Web users want.

3 Architecture of Ontology Based Web Mining

The basic architecture of ontology-based Web problem solving is illustrated in Figure 1. It is integrated with cycles and consists of four basic phases: data mining, representations of discovered knowledge, data reasoning, and knowledge evolution.

The objective of data mining phase is to obtain discovered knowledge from databases based on users constraints and meta-data. In this phase, users constraints are used firstly to clean the original data by removing some useless attributes or editing some data values. The meta-data knowledge is used to generate a training set, where a training set may be an information table or a decision table. The association between data items at the same level is discovered and represented as association rules or patterns.

Knowledge representation in knowledge engineering provides many methods for describing experts' knowledge in terms of computable formats. Different to experts' knowledge, usually data mining provides very larger amount of discovered knowledge. Also the discovered knowledge includes a lot of noise and uncertain information, which cannot be detected in the phase of data mining. The objective of the representation phase is to conduct an efficient way (ontology) to accumulate the discovered knowledge in order to access the right discovered knowledge for producing satisfactory results in the third phase.

The third phase is data reasoning, which attempts to use the right discovered knowledge through the ontology to response users' inputs. It will decide a common hypothesis space in order to transform the discovered knowledge in a reasoning model. It also provides a set of outputs as the result of data reasoning.

The last phase is about knowledge evolution. The objective of this phase is to refine the discovered knowledge based on users' feedback. It will prune some discovered knowledge if that makes an unsatisfactory result. It may also revise some discovered knowledge by updating its associated information that was added in the second phase.

An ontology $(\mathbb{T}, \mathbb{O})$ is used as a vehicle between these phases. It consists of two basic parts: an info-set and a backbone, where info-set $\mathbb{T}$ is a vocabulary of selected terms; and backbone $\mathbb{O}$ describes the discovered patterns and relationships between them. The ontology will be constructed in phase 1 and phase 2. In phase 3, the ontology will be used to answer queries. The ontology can also be revised in the last phase if the system can obtain the user feedback.

4 Data Mining in Web Documents

In this section, we discuss two kinds of methods for knowledge discovery in text documents: association rule mining and granule mining.

4.1 Association Mining

Formally the association discovery from text documents can be described as an information table $(\mathbb{D}, V^{\mathbb{D}})$, where $\mathbb{D}$ is a set of documents in which each document is a set of terms (the duplicate terms are removed here); and $V^{\mathbb{D}} = \{t_1, t_2, \dots, t_n\} \subseteq \mathbb{T}$, which is a set of selected terms for all documents in $\mathbb{D}$. Usually $\mathbb{D}$ consists of a set of positive documents $\mathbb{D}^+$, and a set of negative documents $\mathbb{D}^-$.

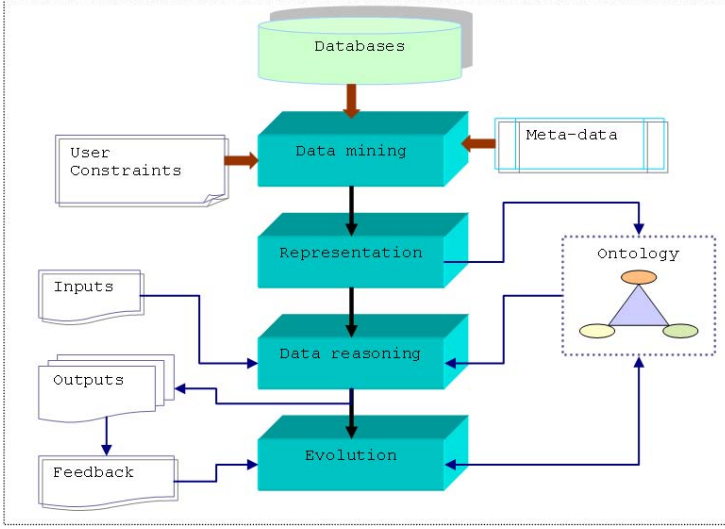


Fig. 1. The architecture of onto-based problem solving

Definition 1. A set of terms X is referred to as a *termset* if $X \subseteq V^{\mathbb{D}}$. Let X be a termset, we use $[X]$ to denote the covering set of X , which includes all positive documents d such that $X \subseteq d$, i.e., $[X] = \{d | d \in \mathbb{D}^+, X \subseteq d\}$.

Given a termset X , its support is $||[X]||/|\mathbb{D}^+|$. A termset X is called frequent pattern if its *support* $\geq \text{min\_sup}$, a minimum support. The confidence of a frequent pattern is the fraction of the documents including the pattern that are positive. Given a frequent pattern X , its confidence is defined as $||[X]||/N$, where $N = \{d | d \in \mathbb{D}, X \subseteq d\}$. The confidence shows the percentage of the pattern's occurrence frequency in the positive documents. A frequent pattern is called an interesting pattern if its *confidence* $\geq \text{min\_conf}$, a minimum confidence.

Table 1 lists part of an information table, where $V^{\mathbb{D}} = \{t_1, t_2, \dots, t_7\}$, $\mathbb{D} = \{d_1, \dots, d_6, d_7, d_8\}$, $\mathbb{D}^+ = \{d_1, d_2, \dots, d_6\}$ and $\mathbb{D}^- = \{d_7, d_8\}$. Let $\text{min\_sup} = 50\%$ and $\text{min\_conf} = 75\%$, we can get 10 interesting patterns. Table 2 shows these interesting patterns and their covering set.

There are some noise patterns in Table 2. For example, pattern $\{t_3, t_4\}$ always occurs with term t_6 in $\mathbb{D}^+$. Therefore, we expect to keep only the larger patterns $\{t_3, t_4, t_6\}$ and prune the noise one, the shorter one $\{t_3, t_4\}$.

Definition 2. Given a termset X , we know its covering set $[X]$, which is a subset of positive documents. Similarly, given a set of positive documents Y , we can define its termset, which satisfies

$$\text{termset}(Y) = \{t | t \in V^{\mathbb{D}}, \forall d \in Y \Rightarrow t \in d\}.$$

Table 1. An information table

<i>Documents</i>	<i>Terms</i>	<i>POS</i>
d_1	$t_1\ t_2$	<i>yes</i>
d_2	$t_3\ t_4\ t_6$	<i>yes</i>
d_3	$t_3\ t_4\ t_5\ t_6$	<i>yes</i>
d_4	$t_3\ t_4\ t_5\ t_6$	<i>yes</i>
d_5	$t_1\ t_2\ t_6\ t_7$	<i>yes</i>
d_6	$t_1\ t_2\ t_6\ t_7$	<i>yes</i>
d_7	$t_1\ t_2$	<i>no</i>
d_8	$t_3\ t_4$	<i>no</i>

Table 2. Interesting patterns and their covering sets

<i>Interesting Pattern</i>	<i>Covering Set</i>
$\{t_3, t_4, t_6\}$	$\{d_2, d_3, d_4\}$
$\{t_3, t_4\}$	$\{d_2, d_3, d_4\}$
$\{t_3, t_6\}$	$\{d_2, d_3, d_4\}$
$\{t_4, t_6\}$	$\{d_2, d_3, d_4\}$
$\{t_3\}$	$\{d_2, d_3, d_4\}$
$\{t_4\}$	$\{d_2, d_3, d_4\}$
$\{t_1, t_2\}$	$\{d_1, d_5, d_6\}$
$\{t_1\}$	$\{d_1, d_5, d_6\}$
$\{t_2\}$	$\{d_1, d_5, d_6\}$
$\{t_6\}$	$\{d_2, d_3, d_4, d_5, d_6\}$

Figure 2 shows the connection between terms and positive documents for these interesting patterns.

Definition 3. Given an interesting pattern X , its closure $Cls(X) = termset([X])$. An interesting pattern X is closed if and only if $X = Cls(X)$.

Based on the above definition, there are only three closed patterns in Table 2. They are $\{t_3, t_4, t_6\}$, $\{t_1, t_2\}$ and $\{t_6\}$.

Each closed pattern can be actually viewed as an association rule, e.g., the closed pattern $\{t_1, t_2\}$ means

$$(t_1 \wedge t_2) \rightarrow (POS = yes);$$

because it satisfies the definition about traditional association rules in data mining.

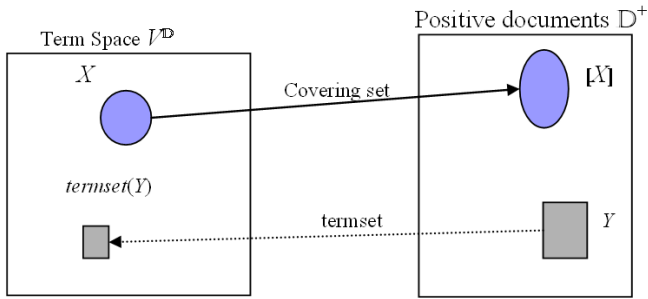


Fig. 2. Connection between Terms and Documents

4.2 Granule Mining

Syntactically topics interesting for the user are constructed from some primitive objects (e.g., terms) in ontology mining technique. According to this assumption, patterns consist of primitive patterns and compound patterns. The primitive patterns are the smallest concepts that cannot be assembled from other patterns. However they may be inherited by some derived concepts or their children (e.g., sub-terms). The compound patterns are the interesting topics, which can be constructed from a set of primitive patterns.

In this sub-section, we use granules to represent compound patterns. The training set $\mathbb{D}$ is now characterized by a decision table. Table 3 illustrates a binary decision table about the documents in Table 1, where N_g is the number of documents that are in the same granule; terms $t_1, t_2, \dots, t_7$ are called condition attributes and POS is called a decision attribute. Compare to Table 1, a granule in Table 3 not only describes a feature in a document, it also shows the number of documents with the same feature.

Table 3. A Binary Decision Table

Granule	t_1	t_2	t_3	t_4	t_5	t_6	t_7	POS	N_g
g_1	1	1	0	0	0	0	0	yes	1
g_2	0	0	1	1	0	1	0	yes	1
g_3	0	0	1	1	1	1	0	yes	2
g_4	1	1	0	0	0	1	1	yes	2
g_5	1	1	0	0	0	0	0	no	1
g_6	0	0	1	1	0	0	0	no	1

Formally let $A = \{a_1, \dots, a_k, \dots, a_m\}$ be the set of attributes, and B be a subset of A . B determines a binary relation $I(B)$ on $\mathbb{D}$ such that $(d_i, d_j) \in I(B)$ if and only if $a(d_i) = a(d_j)$ for every $a \in B$, where $a(d)$ denotes the value of attribute a for $d \in \mathbb{D}$. It is easy to prove that $I(B)$ is an equivalence relation, and the family of all equivalence classes of $I(B)$, that is a partition determined by B , is denoted by $\mathbb{D}/B$. The classes in $\mathbb{D}/B$ are referred to B -granules.

A B -granule g which contains d is called B -granule induced by d , and is denoted by $B(d)$ (or $B(g)$), the set of documents that have the same attribute values for all attributes in B .

The attributes can be divided into two groups: condition attributes and decision attributes, respectively. We call the tripe $(\mathbb{D}, A_C, A_D)$ a decision table if $A_C \cap A_D = \emptyset$ and $A_C \cup A_D = A$. For example, in Table 3 we have $A = \{t_1, \dots, t_7, POS\}$, $A_C = \{t_1, \dots, t_7\}$ and $A_D = \{POS\}$. We also call $\{a | a \in A_C, a(g) \neq 0\}$ the *termset* of granule g , denoted as $termset(g)$. For example, $termset(g_1) = \{t_1, t_2\}$ in Table 3.

Decision tables provide an alternative way to represent the discovered knowledge in a database. For example, each granule in Table 3 can be mapped into a decision rule: either a positive decision rule (the conclusion is $POS = yes$) or a negative decision rule (the conclusion is $POS = no$).

Formally, each granule g determines a sequence $t_1(g), \dots, t_7(g), POS(g)$. The sequence can determine a decision rule:

$$t_1(g) \wedge \dots \wedge t_7(g) \rightarrow POS(g).$$

Its strength is defined as $N_g/|\mathbb{D}|$, and its certainty factor is defined as N_g/K , where $K = \sum_{g_i, \forall t \in V^{\mathbb{D}} \Rightarrow t(g_i)=t(g)} N_{g_i}$.

For example, g_1 in Table 3 (or called A -granule, $A(d_1)$, see Table 1) can be read as the following positive rule:

$$t_1 \wedge t_2 \rightarrow POS = yes$$

or in short $g_1(A_C) \rightarrow g_1(A_D)$ or $d_1(A_C) \rightarrow d_1(A_D)$. Its strength is $1/8$ and the certainty factor is $1/2$.

Binary decision tables only represent high level abstractions of documents. To consider term frequencies in documents, we now use decision tables with multiple attribute values to represent the discovered patterns. Table 4 shows a decision table with multiple attribute values, where the numbers are term frequencies in the corresponding documents.

Based on the above discussion, any document d can be viewed as a condition granule (or a pattern)

$$d(A_C) = \{(t_1, f_1), (t_2, f_2), \dots, (t_m, f_m)\} \quad (1)$$

where f_i is the term frequency of t_i in d .

5 Representations

In this section, we discuss two methods of representation of the discovered knowledge: the pattern taxonomy model and the backbone of the ontology in the ontology mining technique.

Table 4. Original Documents

<i>Granule</i>	t_1	t_2	t_3	t_4	t_5	t_6	t_7	<i>POS</i>	N_P
P_1	2	1	0	0	0	0	0	yes	1
P_2	0	0	2	1	0	1	0	yes	1
P_3	0	0	3	1	1	1	0	yes	1
P_4	0	0	1	1	1	1	0	yes	1
P_5	1	1	0	0	0	1	1	yes	1
P_6	2	1	0	0	0	1	1	yes	1
P_7	2	2	0	0	0	0	0	no	1
P_8	0	0	1	1	0	0	0	no	1

5.1 Pattern Taxonomy Model

In general, we believe that patterns carry more semantic information than individual words. Although patterns contain less ambiguous and narrower meanings than individual words, the likely reasons (also see [57]) for the discouraging performance from the use of patterns are:

- Patterns have inferior statistical properties to words.
- They have low frequency of occurrence.
- There are a large number of redundant and noisy patterns among them.

In order to solve the above-mentioned problems, new studies have been focusing on finding better text representatives from a set of documents. One of them is the pattern taxonomy model (PTM) which has shown a certain degree of improvement in effectiveness [62].

The PTM considered the relations of patterns by creating a taxonomy to show the *is_a* (term “subset”) relation. It also adopted the concept of closed patterns and pruned non-closed patterns from the representation in an attempt to reduce the size of the feature set and lower the impact of noisy patterns.

Figure 3 illustrated an example of PTM for the example in Table 1.

5.2 Backbone of the Ontology

For the ontology mining technique, a pattern is a condition granule that has the form of Eq. (1). When we use such patterns, however, there exists ambiguities for determining other relevance information for a specified topic because there may be many patterns that have the same termset. For this reason, we need to compose patterns with the same termset into one if they are positive.

The composition of two patterns P_1 and P_2 can be processed by using the following expression:

$$\begin{aligned}
 P_1 \oplus P_2 &= \{(t, f_1 + f_2) | (t, f_1) \in P_1, (t, f_2) \in P_2\} \cup \\
 &\quad \{(t, f) | (t, f) \in P_1 \cup P_2, \\
 &\quad t \in (\text{termset}(P_1) \cup \text{termset}(P_2) - \text{termset}(P_1) \cap \text{termset}(P_2))\} \quad (2)
 \end{aligned}$$

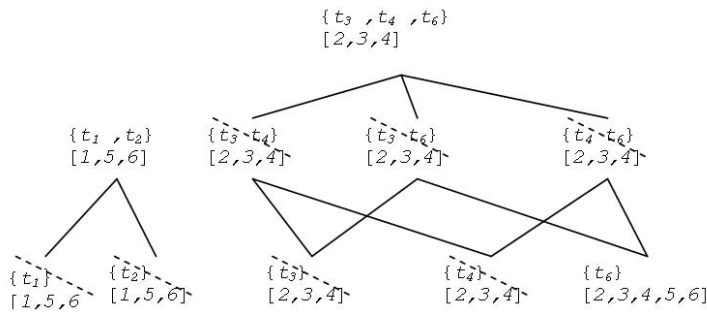


Fig. 3. Pruning Non-closed Patterns

For example, $P_a = \{(t_1, 1), (t_2, 1)\}$ and $P_b = \{(t_2, 1), (t_3, 1)\}$ respectively. The composition of these two patterns

$$P_a \oplus P_b = \{(t_1, 1), (t_2, 2), (t_3, 1)\}.$$

Table 5 illustrates the result of using composition operation to the patterns in Table 4, where $G_1 = P_1$, $G_2 = P_2$, $G_3 = P_3 \oplus P_4$, and $G_4 = P_5 \oplus P_6$ for all positive granules.

Table 5. Compositions of patterns

Granule	t_1	t_2	t_3	t_4	t_5	t_6	t_7	POS	N_G
G_1	2	1	0	0	0	0	0	yes	1
G_2	0	0	2	1	0	1	0	yes	1
G_3	0	0	4	2	2	2	0	yes	2
G_4	3	2	0	0	0	2	2	yes	2
G_5	2	2	0	0	0	0	0	no	1
G_6	0	0	1	1	0	0	0	no	1

Figure 4 shows the relationships between patterns, where an arrow denotes an “is_a” relation that is also a subset relation between patterns’ termsets, and *oplus* means a composition. In [35], the structure in Figure 4 was also called a top backbone of the ontology in describing what users want.

6 Data Reasoning

In general, the concept of relevance is subjective. People normally describe the relevance of a specified topic in two dimensions, specificity and exhaustivity. It is easy for human experts to subjectively assess objects using several scales. For example, we may use 0 to denote not specific, 1 to denote marginally specific, 2 to denote fairly specific, and 3 to denote highly specific.

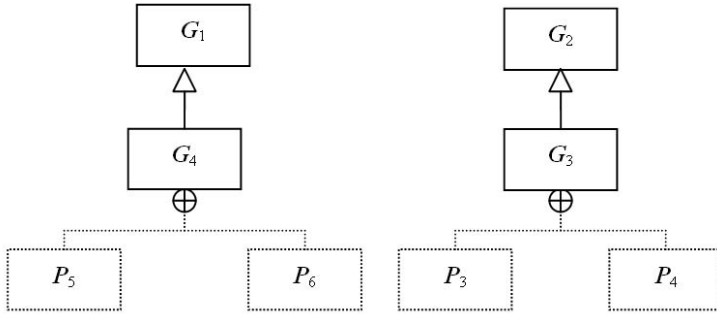


Fig. 4. Top Backbone of the Ontology

The objective of data reasoning is to use the right discovered knowledge to response users' inputs. For this purpose, the relevance of a topic in the ontology is assessed according to the following two dimensions: (1) Exhaustivity (*exh* for short), which describes the extent to which the pattern (or topic) discusses what users want; and (2) Specificity (*spe* for short), which describes the extent to which the pattern (or topic) focuses on what users want.

A pattern in the ontology can be assessed as highly exhaustive relatively even though it is not specific to what users want. Similarly, a pattern can be assessed as highly specific relatively even though it discusses many or only a few aspects of what users want. However, a pattern that does not discuss what users want at all must have the lowest specificity.

6.1 Pattern Deploying Method

In PTM, usually the shorter a pattern is, the less its specificity (greater its exhaustivity); and the larger a pattern is, the greater its specificity. However, in the phase of data mining, the shorter a pattern is, the greater its support.

To give a suitable assessment method to patterns in PTM, the data reasoning model is to deploy interesting patterns into a feature space (or a common space). Formally, for each positive document d_k , we deploy its patterns on $\mathbb{T}$ in order to obtain the following vector:

$$\vec{d}_k = \langle (t_{k_1}, n_{k_1}), (t_{k_2}, n_{k_2}), \dots, (t_{k_m}, n_{k_m}) \rangle \quad (3)$$

where t_{k_i} in pair (t_{k_i}, n_{k_i}) denotes a single term and n_{k_i} is its support in d_k which is the number of patterns that contain t_{k_i} .

The detailed deploying process is presented in Algorithm 1. Note that the SPMin-ing algorithm [62] is used in line 3 for generating frequent sequential patterns. The main process of pattern deploying is between line 5 and line 7 (the nested for loop) inclusively. The normalization operation in line 8 is normally used to assume the contribution of each document is equal. The result of this algorithm is a set of document vectors, which can be expressed as follows.

$$\Theta = \{\vec{d}_1, \vec{d}_2, \dots, \vec{d}_n\} \quad (4)$$

We also need to determine the weight for each term in $\mathbb{T}$ when we use the discovered knowledge in Θ . The weighting scheme for a given term t_i is denoted as the following function.

$$weight_1(t_i) = \sum_{\vec{d}_k \in \Theta, (t_i, n_{k_i}) \in \vec{d}_k} \left(\frac{n_{k_i}}{\sum_{(t,w) \in \vec{d}_k} w} \right) \quad (5)$$

where w is the support of term t indicating the number of patterns that contain t in a document d .

In order to efficiently deploy discovered patterns and acquire deployed support of each term in these patterns, we use the pattern composition operation (see Eq. (2)) to join two patterns, where we assume each pattern a set of term number pairs, for instance, $P = \{t_1, t_2, t_3\}$ would be viewed as $\{(t_1, 1), (t_2, 2), (t_3, 1)\}$.

6.2 Rough Association

Give a granule G in Table 5, it can also be separated into two parts, condition granule $A_C(G)$ and decision granule $A_D(G)$. We also need to distinguish positive condition granules and negative condition granules because they would provide different semantic meanings for information gathering. Therefore, we can use a 2-tier structure to show the relation between condition granules and decision granules.

For example, using Table 4, we can get a set of positive condition granules (also a set of negative condition granules)

$$\mathbb{D}^+/A_C = \{\{P_1\}, \{P_2\}, \{P_3, P_4\}, \{P_5, P_6\}\}$$

and the set of decision granules,

$$\mathbb{D}/A_D = \{POS = yes, POS = no\} = \{\{P_1, P_2, P_3, P_4, P_5, P_6\}, \{P_7, P_8\}\}$$

, respectively. In the following we let $\mathbb{D}^+/A_C = \{cg_1, cg_2, cg_3, cg_4\}$ and $\mathbb{D}/A_D = \{dg_1, dg_2\}$.

Algorithm 1. PDM($\mathbb{D}$, min_sup)

input : a list of documents, $\mathbb{D}$; minimum support, min_sup .

output: a set of document vectors Θ .

$\Theta \leftarrow \emptyset$;

for document d in $\mathbb{D}$ **do**

$SP = \text{SPMining}(PL, min\_sup)$; //pattern discovery in a set of paragraphs

$P \leftarrow \emptyset$;

for pattern P_i in SP **do**

$P \leftarrow P \oplus P_i$; //pattern composition

end

 normalize P ;

$\Theta \leftarrow \Theta \cup \{P\}$;

end

We also need to consider the weight distribution of terms for the condition granules in order to consider the factor of the frequencies of terms in documents. Let cg_i 's *coverset* be $\{P_{i1}, P_{i2}, \dots, P_{im}\}$, that is $cg_i = P_{i1} \oplus P_{i2} \oplus \dots \oplus P_{im}$, we can obtain a weight distribution for terms t_j in these documents using the following equation:

$$weight(t_j) = \frac{t_j(cg_i)}{\sum_{t \in A_c} t(cd_i)} \quad (6)$$

where we use the composition operation to assign a value to condition granules' attributes, which satisfies:

$$t(cg_i) = t(P_{i1}) + t(P_{i2}) + \dots + t(P_{im}) \quad (7)$$

for all $t \in A_C$.

Table 6 illustrates granules in the 2 tiers (condition granules and decision granules) and their *coversets* we obtain from Table 4 according to previous definitions, where each condition granule consists of both a termset and a weight distribution. For example, $cg_1 = \langle \{t_1, t_2\}, (2/3, 1/3, 0, 0, 0, 0, 0) \rangle$ or in short $cg_1 = \{(t_1, 2/3), (t_2, 1/3)\}$.

Table 6. Granules and their coversets

<i>Condition granule</i>	t_1	t_2	t_3	t_4	t_5	t_6	t_7	<i>coverset</i>	<i>POS</i>
cg_1	2/3	1/3	0	0	0	0	0	$\{P_1\}$	<i>yes</i>
cg_2	0	0	1/2	1/4	0	1/4	0	$\{P_2\}$	<i>yes</i>
cg_3	0	0	2/5	1/5	1/5	1/5	0	$\{P_3, P_4\}$	<i>yes</i>
cg_4	1/3	2/9	0	0	0	2/9	2/9	$\{P_5, P_6\}$	<i>yes</i>
cg_5	1/2	1/2	0	0	0	0	0	$\{P_7\}$	<i>no</i>
cg_6	0	0	1/2	1/2	0	0	0	$\{P_8\}$	<i>no</i>

(a) Condition granules

<i>Decision granule</i>	<i>POS</i>	<i>coverset</i>
dg_1	<i>yes</i>	$\{P_1, P_2, P_3, P_4, P_5, P_6\}$
dg_2	<i>no</i>	$\{P_7, P_8\}$

(b) Decision granules

Using the associations between condition granules and decision granules, we can rewrite the decision rules in Table 3 as follows:

$$\begin{aligned}
 termset(g_1) &\rightarrow \{((dg_1, cg_1), 1/2), ((dg_2, cg_5), 1/2)\} \\
 termset(g_2) &\rightarrow \{((dg_1, cg_2), 1)\} \\
 termset(g_3) &\rightarrow \{((dg_1, cg_3), 1)\} \\
 termset(g_4) &\rightarrow \{((dg_1, cg_4), 1)\} \\
 termset(g_5) &\rightarrow \{((dg_2, cg_6), 1)\}
 \end{aligned}$$

Formally the association can be represented as the following mapping:

$$\Gamma : (2^{V^{\mathbb{D}}}) \rightarrow 2^{((\mathbb{D}/A_D) \times (\mathbb{D}/A_C)) \times [0,1]}$$

where $\Gamma(\text{termset}(g_i))$ satisfies:

$$\sum_{((dg, cg), f) \in \Gamma(\text{termset}(g_i))} f = 1$$

Now we consider the support degree for each condition granule. The obvious way is to use frequencies in the decision table, that is,

$$NC(cg_i) = \sum_{P \in \text{cover set}(cg_i)} N_P.$$

for every condition granule cg_i . By normalizing, we can get a support function sup on $(\mathbb{D}/A_C)$ such that

$$sup(cg_i) = \frac{NC(cg_i)}{\sum_{cg \in (\mathbb{D}/A_C)} NC(cg)}$$

for all $cg_i \in (\mathbb{D}/A_C)$. It is obvious that sup is a probability function on $\mathbb{D}/A_C$.

Definition 4. Let $\Gamma(\text{termset}(g_i)) = \{((dg_{i,1}, cg_{i,1}), f_{i,1}), \dots, ((dg_{i,k}, cg_{i,k}), f_{i,k})\}$. We call “ $cg_{i,j} \rightarrow (dg_{i,j}, f_{i,j})$ ” ($1 \leq j \leq k$) a rough association rule, its strength is $sup(cg_{i,j}) \times f_{i,j}$ and its certainty factor is $f_{i,j}$.

Figure 5 illustrates a data structure for representations of rough associations between condition granules and decision granules.

<u><i>termset</i></u>		Γ	
$\{t_1, t_2\}$	$\rightarrow$	$((dg_1, cg_1), 0.5)$	$((dg_2, cg_5), 0.5)$
$\{t_3, t_4, t_6\}$	$\rightarrow$	$((dg_1, cg_2), 1)$	
$\{t_3, t_4, t_5, t_6\}$	$\rightarrow$	$((dg_1, cg_3), 1)$	
$\{t_1, t_2, t_6, t_7\}$	$\rightarrow$	$((dg_1, cg_4), 1)$	
$\{t_3, t_4\}$	$\rightarrow$	$((dg_2, cg_6), 1)$	

Fig. 5. Rough association rules

To use rough associations, we use the following equation to evaluate a weight for each term according to the discovered rough association rules *Rule*:

$$\text{weight}_2(\text{term}) = \sum_{(cg \rightarrow (dg_1, f)) \in \text{Rule}, (\text{term}, w) \in cg} sup(cg) \times w \quad (8)$$

where dg_1 is the short form of $(POS = yes)$.

In addition, given a testing document d , we use the following equation to determine its relevance in both PTM and rough association mining model:

$$rel(d) = \sum_{term \in V^d, term \in d} weight_i(term) \quad (9)$$

where $i = 1$ or 2 .

7 Knowledge Evolution

The consequential result of using Eq. (9) is that many irrelevant documents may be marked in relevance in Web information gathering.

For PTM, the problem might be that strongly supported patterns (usually they are shorter ones) are more general concepts even they have high confidence. On the other hand, some weakly supported patterns (they are usually larger patterns) may focus on what users want. Therefore, it would be necessary to consider Web user's search intention based on both positive and negative documents in the training sets in order to improve the performance of PTM [75]. The focus of this research is on finding out user search intention based on the form of taxonomy of discovered patterns. Compared with the original PTM method, the experiments based on the new method showed a considerable improvement in terms of search effectiveness regarding precision and recall. Although the overall improvement is modest, it was consistent in all experiments.

For the ontology mining technique, the problem could be that we used incorrect weigh distributions if we only consider positive documents in the training set. It is desirable to update the positive rough association rules based on some interesting negative rough association rules.

Give a negative rough association rule $\tilde{c}g \rightarrow (\tilde{d}g, \tilde{f})$, where $\tilde{d}g = dg_2$ is the short form of $(POS = no)$ and $termset(\tilde{c}g) = \{t_1, t_2, \dots, t_m\}$. We call it an interesting negative rule if

$$rel(termset(\tilde{c}g)) \geq \min\{rel(d) | d \in \mathbb{D}^+\}. \quad (10)$$

The approach we used here is to trace the cause of the occurrences of interesting negative rules. For a given interesting negative one, $\tilde{c}g \rightarrow (\tilde{d}g, \tilde{f})$, we check which positive rules have been used to give rise to such error. We call these rules its offenders. The offenders can be determined by the following equation

$$\{(cg \rightarrow (dg_1, f)) \in Rule | termset(cg) \cap termset(\tilde{c}g) \neq \emptyset\}. \quad (11)$$

The basic idea of updating some positive rough association rules are described in Algorithm 2. There are two kinds of offenders: total conflict offenders (see the condition of the first "if statement" - $(termset(cg) \subseteq termset(\tilde{c}g))$), and partial conflict offenders (see the condition of the second "if statement" - $(termset(cg) \cap termset(\tilde{c}g) \neq \emptyset)$). We reduce the supports for all total conflict offenders. For partial conflict offenders, we reshuffle their term weight distributions to evaporate the uncertainties contained in the condition granule.

The details of the reshuffle operation can be found in [35] and its performance was significant.

8 Related Work

As mentioned in the introduction, currently Web mining can be classified into four categories: Web usage mining, Web structure mining, Web user profile mining, and Web content mining [14,28,59,46]. The obvious difference between Web mining and data mining is that the former is based on Web-related data sources, such as unstructured documents (e.g., HTML), semi-structured documents (e.g., XML), log, services and user profiles; and the latter is based on databases [21,28].

Data mining, which is also referred to as knowledge discovery in database is a process of nontrivial extraction of implicit, previously unknown and potentially useful information (patterns) from data in databases [8,10]. To discover the potential useful knowledge (or patterns), several typical approaches have been presented. They are data classification, data clustering, and mining association rules.

Data classification and clustering have been used in Web log mining to discover new and interesting user behavior patterns (see [43] for more details). Association mining technique has been used in many Web usage mining systems to find correlations among Web pages and interesting access patterns [9,58,60,1]. This technique has also been used for Web pre-fetching [64] and Web personalization [50].

Also association mining has been used in Web text mining, which refers to the process of searching through unstructured data on the Web and deriving meaning from it [11,12,18]. The main purposes of text mining were association discovery, trends discovery, and event discovery [6]. The association between a set of keywords and a predefined category (e.g., a term) can be described as an association rule. The trends discovery means the discovery of phrases, a sort of sequence association rules. The event discovery is the identification of stories in continuous news streams. Usually clustering based mining techniques can be used for such a purpose. It was also necessary to combine association rule mining with the existing taxonomies in order to determine useful patterns [13,4].

To compare with IR-based models, data mining-based Web text mining models do not use term independent assumption [2,41]. Also, Web mining models try to discover

Algorithm 2. Updating Positive Rough Association Rules

input : a set of rough association rules, *Rule*, and an interesting negative association rule, $\tilde{c}g \rightarrow (\tilde{d}g, \tilde{f})$.
output: the updated set of rough association rules
//update all offenders.
for each positive rule $(cg \rightarrow (dg_1, f)) \in Rule$ **do**
 if ($termset(cg) \subseteq termset(\tilde{c}g)$) **then**
 reduce *cg*'s support;
 else
 if ($termset(cg) \cap termset(\tilde{c}g) \neq \emptyset$) **then**
 reshuffle *cg*'s term weight distribution by shifting part of weights from all terms in $termset(cg) \cap termset(\tilde{c}g)$ to *cg*'s rest terms;
 end
 end
end
end

some unexpected useful data [43,42,6]. The disadvantage of association rule mining is that there are too many discovered patterns that make the application of the discovered knowledge inefficient. Also there are many noise patterns that make the discovered knowledge contains much uncertainties. Although pruning non-closed patterns that can improve the quality of association mining in text mining in some extents [62], the performance of text mining systems are still ineffectively.

Granule mining [34] can be an alternative solution to specify association rules, where a granule is kind of representation of a group of objects (transactions) that satisfy user constraints, e.g., all objects have the same attributes' values. Decision tables [54,29] can be a basic structure for granule mining in which attributes are divided into two groups, condition attributes and decision attributes. This study is also related to formal concept analysis and rough set analysis (see [65]).

However, there exists ambiguities whilst we use the decision rules for Web information gathering. We have demonstrated in the previous sections that rough association rule mining [33,36,38] can be used to overcome these disadvantages.

The basic architecture we used to implement the above idea is to automatically construct and maintain an ontology for representation, application and updating of discovered knowledge. The related work for this architecture is about ontology learning algorithms. Several ontology learning algorithms have been presented such as pattern matching [17,51], hierarchical clustering and pattern taxonomy [49,62]. The taxonomic backbone of the ontology used in these algorithms only included "is-a" relation. Non-taxonomic relation was also motioned in [48] by using association rules.

9 Conclusions

We presented the basic idea of ontology-based approach for Web problem solving in [28]. We also formalized an abstract Web mining model for this research in [31]. The abstract Web mining model represented knowledge on user profiles by using an ontology which consisted of both "part-of" and "is-a" relations. We also described the details of using the abstract Web mining model for information gathering. In addition, the ontology mining technique was proposed that can be used to automatically discover ontologies and refine discovered ontologies for Web information gathering in [35].

Discovered knowledge is the primary source of the intellectual power of Web information gathering systems. We discuss two ways of finding discovered knowledge from Web documents for ontology-based Web data mining. The basic PTM model used up-to-date techniques of association mining for this question (see [62]). However, the performance of the basic PTM model was not satisfactory. Currently there are two methods to improve the basic PTM model, the pattern deploying method [63] and the model [75] for considering user search intension. The former got a significant improvement for PTM.

Granule mining was another way that we formally presented in [34] recently. The initial idea came from decision rules in [55]. We presented an efficient algorithm to represent decision rules in [29]. We also presented a new concept, rough association mining, to make decision rule suitable for Web information gathering in [33,36]. In addition, we clarified the relation between association mining and granule mining, and

proved that a granule is a sort of closed pattern in [38]. We also evaluated the evolution method for rough association mining and that can obtain a significant performance (see [37]).

Data mining based approaches have not had the impact of the expert systems produced by the usual knowledge engineering methods. To overcome this disadvantage, we are now planning to introduce domain ontologies for ontology-based Web data mining. The first one was DDC (Dewey Decimal Code) based QUT Library ontology [19] that has been partially tested for the problem of collection selection. We are also considering how to automatically acquire training sets from the Web by using surrounding knowledge about the Web [61].

Acknowledgements

This chapter was partially supported by Grant DP0556455 from the Australian Research Council. The authors also wish to thank Dr. Yue Xu, Dr. Richi Nayak, Dr. Raymond Lau, Prof. Peter Bruza, Prof. Chengqi Zhang, Dr. Shichao Zhang, Prof. Yiyu Yao, Prof. Jiming Liu, Mrs. Xujuan Zhou, Mr. Sheng-Tang Wu, Mr. John King, Mr. Wangzhong Yang, Mr. Xiaohui Tao, Mr. Ben Murphy, and Mr. Jia Hu.

References

1. Abraham, T., de Vel, O.: Investigative profiling with computer forensic log data and association rules. In: Proc. 2002 IEEE International Conference on Data Mining, pp. 11–18. IEEE Press, Los Alamitos (2002)
2. Antonie, M.L., Zaiane, O.R.: Text document categorization by term association. In: Proc. 2002 IEEE International Conference on Data Mining, pp. 19–26. IEEE Press, Los Alamitos (2002)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
4. Bi, Y., Anderson, T., McClean, S.: A rough set model with ontologies for discovering maximal association rules in document collection. *Knowledge-Based Systems* 16, 243–251 (2003)
5. Brusilovsky, P., Maybury, M.: From adaptive hypermedia to the the adaptive Web. *Communications of the ACM* 45(5), 30–33 (2002)
6. Chang, G., Healey, M.J., McHugh, J.A.M., Wang, J.T.L.: Mining the World Wide Web: an information search approach. Kluwer Academic Publishers, Dordrecht (2001)
7. Chen, M.S., Park, J., Yu, P.S.: Data mining for path traversal patterns in a Web environment. In: Proc. 16th International Conference on Distributed Computing Systems, Hong Kong, pp. 385–392 (1996)
8. Chen, M.S., Han, J., Yu, P.S.: Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8(6), 866–883 (1996)
9. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: information and pattern discovery on the World-Wide Web. In: 9th IEEE International Conference on Tools with Artificial Intelligence, pp. 558–567. IEEE Press, Los Alamitos (1997)
10. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in knowledge discovery and data mining. AAAI Press/The MIT Press, Menlo Park, California (1996)

11. Feldman, R., Hirsh, H.: Mining associations in text in presence of background knowledge. In: Proc. 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 343–346. AAAI Press/The MIT Press (1996)
12. Feldman, R., Dagen, I., Hirsh, H.: Mining text using keywords distributions. *Journal of Intelligent Information Systems* 10(3), 281–300 (1998a)
13. Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., Zamir, O.: Text mining at the term level. In: Żytkow, J.M. (ed.) PKDD 1998. LNCS, vol. 1510, pp. 65–73. Springer, Heidelberg (1998b)
14. Garofalakis, M.N., Rastogi, R., Seshadri, S., Shim, K.: Data mining and the Web: past, present and future. In: Proc. WIDM 1999 conference, Kansas City, Missouri, pp. 43–47 (1999)
15. Grossman, D.A., Frieder, O.: Information retrieval algorithms and heuristics. Kluwer Academic Publishers, Dordrecht (1998)
16. Guan, J.W., Bell, D.A., Liu, D.Y.: The rough set approach to association rules. In: Proc. 2003 IEEE International Conference on Data Mining, Melbourne, Florida, USA, pp. 529–532. IEEE Computer Society Press, Los Alamitos (2003)
17. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. 14th International Conference on Computational Linguistics, Nantes, France (1992)
18. Holt, J.D., Chung, S.M.: Multipass algorithms for mining association rules in text databases. *Knowledge and Information Systems* 3, 168–183 (2001)
19. King, J., Li, Y., Tao, X., Nayak, R.: Mining world knowledge for analysis of search engine content, *Web Intelligence and Agent Systems*, 5 (in press)
20. Jones, K.S.: Information retrieval and artificial intelligence. *Artificial Intelligence* 114(1-2), 257–281 (1999)
21. Lee, K.C., Kim, J.S., Chung, N.H., Kwon, S.J.: Fuzzy cognitive map approach to Web mining inference amplification. *Expert System with Applications* 22, 197–211 (2002)
22. Lesser, V., et al.: BIG: an agent for resource-bounded information gathering and decision making. *Artificial Intelligence* 118, 197–244 (2000)
23. Li, Y., Zhang, C., Swan, J.R.: An information filtering model on the Web and its application in JobAgent. *Knowledge-based Systems* 13(5), 285–296 (2000)
24. Li, Y.: Information fusion for intelligent agent-based information gathering. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) WI 2001. LNCS (LNAI), vol. 2198, pp. 433–437. Springer, Heidelberg (2001)
25. Li, Y., Yao, Y.Y.: User profile model: a view from artificial intelligence. In: Proc. 3rd International Conference on Rough Sets and Current Trends in Computing, Malvern, PA, USA, pp. 493–496 (2002)
26. Li, Y.: Extended random sets for knowledge discovery in information systems. In: Proc. 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, China, pp. 524–532 (2003)
27. Li, Y., Zhang, C., Zhang, S.: Cooperative strategy for Web data mining and cleaning. *Applied Artificial Intelligence* 17(5-6), 443–460 (2003)
28. Li, Y., Zhong, N.: Ontology-based Web Mining Model: representations of user profiles. In: Proc. 2003 IEEE/WIC International Conference on Web Intelligence, pp. 96–103 (2003a)
29. Li, Y., Zhong, N.: Interpretations of association rules by granular computing. In: Proc. 2003 IEEE International Conference on Data Mining, Melbourne, Florida, USA, pp. 593–596 (2003b)
30. Li, Y., Zhong, N.: Capturing evolving patterns for ontology-based Web mining. In: Proc. IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, pp. 256–263 (2004a)
31. Li, Y., Zhong, N.: Web mining model and its applications on information gathering. *Knowledge-Based Systems* 17(5-6), 207–217 (2004b)

32. Li, Y., Murphy, B., Zhong, N.: Mining Interesting Topics for Web Information Gathering and Web Personalization. In: Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence, France, pp. 305–308 (2005)
33. Li, Y., Zhong, N.: Rough Association Mining and Its Application in Web Information Gathering. In: Zhang, S., Jarvis, R. (eds.) AI 2005. LNCS (LNAI), vol. 3809, pp. 1005–1008. Springer, Heidelberg (2005)
34. Li, Y., Yang, W., Xu, Y.: Multi-Tier Granule Mining for Representations of Multidimensional Association Rules. In: Proc. 6th IEEE International Conference on Data Mining, Hong Kong, pp. 953–958. IEEE Computer Society Press, Los Alamitos (2006)
35. Li, Y., Zhong, N.: Mining ontology for automatically acquiring Web user information needs. *IEEE Transactions on Knowledge and Data Engineering* 18(4), 554–568 (2006a)
36. Li, Y., Zhong, N.: Mining Rough Association from Text Documents. In: Proc. 5th International Conference on Rough Sets and Current Trends in Computing, Kobe, Japan, pp. 368–377 (2006c)
37. Li, Y., Zhong, N.: Rough Association Rule Mining in Text Documents for Acquiring Web User Information Needs. In: Proc. 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, pp. 226–232 (2006d)
38. Li, Y., Zhong, N.: Mining Rough Association from Text Documents for Web Information Gathering. In: *Transactions on Rough Sets VII*. LNCS, vol. 4400, pp. 103–119. Springer, Heidelberg (2007)
39. Li, X., Liu, B.: Learning to classify texts using positive and unlabeled data. In: Proc. IJCAI 2003, Mexico, pp. 587–592 (2003)
40. Lin, T.Y.: Database mining on derived attributes. In: Proc. 3rd International Conference on Rough Sets and Current Trends in Computing, Malvern, PA, USA, pp. 14–32 (2002)
41. Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S.: Building text classifiers using positive and unlabeled examples. In: Proc. 2003 IEEE International Conference on Data Mining, Melbourne, Florida, USA, pp. 179–186. IEEE Computer Society Press, Los Alamitos (2003)
42. Liu, B., Ma, Y., Yu, P.S.: Discovery unexpected information from your competitor's Web sites. In: Proc. 7th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (2001)
43. Liu, B., Ma, Y., Yu, P.S.: Discovering business intelligence information by comparing company Web sites. In: Zhong, N., Liu, J., Yao, Y.Y. (eds.) *Web Intelligence*, pp. 105–127. Springer, Heidelberg (2003)
44. Liu, J.: Web Intelligence (WI): What Makes Wisdom Web? In: Proc. Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003), pp. 1596–1601 (2003)
45. Liu, J., Zhang, S., Yang, J.: Characterizing Web usage regularities with information foraging agents. *IEEE Transactions on Knowledge and Data Engineering* 16(5), 566–584 (2004)
46. Lu, Z.Y., Yao, Y.Y., Zhong, N.: Web log mining. In: Zhong, N., Liu, J., Yao, Y.Y. (eds.) *Web Intelligence*, pp. 174–194. Springer, Heidelberg (2004)
47. Madria, S.M., Bhowmick, S.S., Ng, W.K., Lim, E.P.: Research issues in Web data mining. In: Proc. 1st International Conference on Data Warehousing and Knowledge Discovery, pp. 303–312 (1999)
48. Maedche, A.: *Ontology learning for the semantic Web*. Kluwer Academic Publishers, Dordrecht (2003)
49. Manning, C., Schuetze, H.: *Foundations of statistical natural language processing*. The MIT Press, Cambridge (1999)
50. Mobasher, B., Dai, H., Luo, T., Sun, Y., Zhu, J.: Combining Web usage and content mining for more effective personalization. In: Proc. International Conference on Ecommerce and Web Technologies, Greenwich, UK (2000)

51. Morin, E.: Automatic acquisition of semantic relations between terms from technical corpora. In: Proc. 5th International Congress on Terminology and Knowledge Engineering (1999)
52. Mostafa, J., Lam, W., Palakal, M.: A multilevel approach to intelligent information filtering: model, system, and evaluation. *ACM Transactions on Information Systems* 15(4), 368–399 (1997)
53. Pal, S.K., Talwar, V.: Web mining in soft computing framework: relevance, state of the art and future directions. *IEEE Transactions on Neural Networks* 13(5), 1163–1177 (2002)
54. Pawlak, Z.: Rough sets and intelligent data analysis. *International J. of Information and Sciences* 147, 1–12 (2002a)
55. Pawlak, Z.: In pursuit of patterns in data reasoning from data - the rough set way. In: Proc. 3rd International Conference on Rough Sets and Current Trends in Computing, Malvern, PA, USA, pp. 1–9 (2002b)
56. Perkowitz, M., Etzioni, O.: 'Adaptive Web sites. *Communications of the ACM* 43(8), 152–158 (2000)
57. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
58. Shen, L., Cheng, L., Ford, J., Makedon, F., Megalooikonomou, V., Steinberg, T.: Mining the most interesting Web access associations. In: Proc. 5th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, pp. 145–154 (1999)
59. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage pattern from Web data. *SIGKDD Explorations* 1(2), 1–12 (2000)
60. Tan, P.N., Kumar, V.: Discovery of indirect associations from Web usage data. In: Zhong, N., Liu, J., Yao, Y.Y. (eds.) *Web Intelligence*, pp. 128–152. Springer, Heidelberg (2003)
61. Tao, X., Li, Y., Zhong, N., Nayak, R.: Automatically Acquiring Training Sets for Web Information Gathering. In: Proc. 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, pp. 532–535 (2006)
62. Wu, S.T., Li, Y., Xu, Y., Pham, B., Chen, P.: Automatic pattern taxonomy extraction for Web mining. In: Proc. 2004 IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, pp. 242–248 (2004)
63. Wu, S.T., Li, Y., Xu, Y.: Deploying Approaches for Pattern Refinement in Text Mining. In: Proc. 6th IEEE International Conference on Data Mining, Hong Kong, pp. 1157–1161 (2006)
64. Yang, Q., Zhang, H., Tian, I., Li, Y.: Mining Web logs for prediction models in WWW caching and prefetching. In: Proc. 7th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, pp. 473–478 (2001)
65. Yao, Y.Y.: A Comparative Study of Formal Concept Analysis and Rough Set Theory in Data Analysis. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) *RSCTC 2004. LNCS (LNAI)*, vol. 3066, pp. 59–68. Springer, Heidelberg (2004)
66. Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.): *WI 2001. LNCS (LNAI)*, vol. 2198, pp. 1–17. Springer, Heidelberg (2001)
67. Zhong, N., Liu, J., Yao, Y.Y., Ohsuga, S.: Web Intelligence (WI). In: Proc. 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC 2000), pp. 469–470. IEEE Computer Society Press, Los Alamitos (2000)
68. Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.): *WI 2001. LNCS (LNAI)*, vol. 2198. Springer, Heidelberg (2001)
69. Zhong, N., Liu, J., Yao, Y.Y.: In search of the wisdom Web. *IEEE Computer* 35(11), 27–31 (2002)
70. Zhong, N.: Representation and construction of ontologies for Web intelligence. *International Journal of Foundation of Computer Science* 13(4), 555–570 (2002)

71. Zhong, N., Liu, J., Yao, Y.Y. (eds.): Web Intelligence. Springer, Heidelberg (2003)
72. Zhong, N.: Impending Brain Informatics (BI) Research from Web Intelligence (WI) Perspective. *International Journal of Information Technology and Decision Making* 5(4), 713–727 (2006)
73. Zhong, N., Liu, J. (eds.): Intelligent Technologies for Information Analysis. Springer, Heidelberg (2004)
74. Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies (iIT) from the Stand-Point of Web Intelligence (WI). *Communications of the ACM* 50(3), 89–94 (2007)
75. Zhou, X., Wu, S.T., Li, Y., Xu, Y., Lau, R.Y.K., Bruza, P.: Utilizing Search Intent in Topic Ontology-based User Profile for Web Mining. In: *Proc. 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Hong Kong, pp. 558–561 (2006)

A Reasonable Rough Approximation for Clustering Web Users

Duoqian Miao, Min Chen, Zhihua Wei, and Qiguo Duan

¹ Department of Computer Science and Technology,
Tongji University, Shanghai, 201804, China

² The Key Laboratory of Embedded System and Service Computing,
Ministry of Education, China
`miaoduoqian@163.com`

Abstract. Due to the uncertainty in accessing Web pages, analysis of Web logs faces some challenges. Several rough k -means cluster algorithms have been proposed and successfully applied to Web usage mining. However, they did not explain why rough approximations of these cluster algorithms were introduced. This paper analyzes the characteristics of the data in the boundary areas of clusters, and then a rough k -means cluster algorithm based on a reasonable rough approximation (RKMrra) is proposed. Finally RKMrra is applied to Web access logs. In the experiments RKMrra compares to Lingras and West algorithm and Peters algorithm with respect to five characteristics. The results show that RKMrra discovers meaningful clusters of Web users and its rough approximation is more reasonable.

1 Introduction

Web usage mining [1] can be viewed as the application of data mining techniques to any collection of Web access logs. It is a promising research field because Web user information needs are acquired by mining Web access logs. In recent years, it has also become a subtopic of Web Intelligence (WI) [2, 3]. Clustering as an important data mining technique is generally used in Web usage mining. User profiles can be established by clustering Web access logs based on some sort of similarity measures. Clustering is done so that Web users within the same cluster behave more similarly than those in different clusters. Therefore, it is very useful for Web applications, such as personalized recommendation [4], business intelligence [13], and other Web based applications [1].

However, clustering faces some challenges in Web usage mining compared to traditional data mining. Due to the uncertainty in accessing Web pages and the ease of movement from one Web page to another, the clusters tend to have vague or imprecise boundaries. Rough set theory [5, 14] as a kind of tool dealing with imprecision and incomplete knowledge is widely used in clustering Web logs. Several rough k -means cluster algorithms have been proposed, for example the rough k -means cluster algorithm introduced by Lingras and West [6] and the

refined one by Peters [7]. Some other algorithms for clustering Web users have also been proposed in the literature [8, 9, 10, 11].

The concept of the rough approximation in clusters was presented by Lingras and West. In rough clustering each cluster has two approximations, namely the lower and upper approximations. Then Peters analyzed Lingras and West cluster algorithm and pointed out some refinements. However, both of them did not illustrate why these rough approximations in clusters were introduced.

The objective of this paper is to analyze the data objects in the boundary areas of clusters. Based on the analysis, a reasonable rough approximation will be suggested. Then the rationality of the rough approximation will also be explained.

The structure of the paper is as follows. In Section 2 we introduce two rough k -means algorithms, which are Lingras and West algorithm and Peters algorithm. Then these algorithms are analyzed in Section 3. In Section 4 we suggest a more reasonable rough approximation. Based on the rough approximation, a k -means cluster algorithm is proposed. To evaluate the performance of the algorithm, experiments are presented in Section 5. Finally, the paper concludes with a summary in Section 6.

2 Review of Existing Rough k -Means Cluster Algorithms

2.1 Rough Properties of the Cluster Algorithms

Rough set is a kind of mathematical tool for dealing with uncertainty. All the previous rough k -means cluster algorithms use this characteristic of rough set theory. A cluster is represented by a rough set based on a lower approximation and an upper approximation. Although the rough k -means algorithms do not verify all the properties of rough set theory, they have some basic properties as follows:

- Property 1: A data object $\mathbf{X}$ belongs to one lower approximation at most.
- Property 2: For a cluster (set) C and a data object $\mathbf{X}$, if $\mathbf{X}$ belongs to the lower approximation of C , then it also belongs to the upper approximation of C .
- Property 3: If a data object $\mathbf{X}$ does not belong to any lower approximation, then $\mathbf{X}$ belongs to two or more upper approximations. That means $\mathbf{X}$ lies in two or more boundary areas of clusters.

2.2 Existing Rough k -Means Cluster Algorithms

Let $\mathbf{X}_n$ represents the n th data object which is a multidimensional vector. C_k is the k th cluster (set), and its upper and lower approximation are $\overline{C}_k$ and $\underline{C}_k$ respectively. $C_k^B = \overline{C}_k - \underline{C}_k$ is the boundary area of the cluster. $\mathbf{m}_k$ represents the centroid of cluster C_k .

K -means clustering is a process of finding centroids for all clusters, and assigns objects to each cluster based on their distance from the centroids. This process

is done iteratively until stable centroid values are found. Rough k -means cluster algorithms incorporate rough sets into k -means clustering, which requires the addition of the concept of lower and upper bounds, such as Lingras and West algorithm and Peters algorithm.

Lingras and West algorithm use Eq. (1) to calculate the centroids of clusters that is modified to include the effects of lower as well as upper bounds.

$$m_k = \begin{cases} \omega_l \sum_{\mathbf{X}_n \in \underline{C}_k} \frac{\mathbf{X}_n}{|\underline{C}_k|} + \omega_b \sum_{\mathbf{X}_n \in \overline{C}_k^B} \frac{\mathbf{X}_n}{|\overline{C}_k^B|} & \text{for } C_k^B \neq \phi \\ \omega_l \sum_{\mathbf{X}_n \in \underline{C}_k} \frac{\mathbf{X}_n}{|\underline{C}_k|} & \text{otherwise} \end{cases} \quad (1)$$

where ω_l is the lower weight and ω_b is the boundary weight.

The next step in Lingras and West algorithm is to design criteria to determine whether an object belongs to the upper or lower bound of a cluster. When assigning the data object $\mathbf{X}_n$ to the lower or upper approximation, we look for the centroid $\mathbf{m}_s$ closest to $\mathbf{X}_n$ firstly, and then the following set T must be determined first (see Eq. (2)).

$$T = \{t : d(\mathbf{X}_n, \mathbf{m}_k) - d(\mathbf{X}_n, \mathbf{m}_s) \leq \varepsilon \wedge k \neq s\} \quad (2)$$

- If $T \neq \phi$, then $\mathbf{X}_n \in \overline{C}_t$, $\forall t \in T$.
- Else $\mathbf{X}_n \in \underline{C}_s$.

where ε is the threshold.

Lingras and West algorithm, described above, depends on three parameters ω_l , ω_b and ε . Experimentation with various values of the parameters is able to develop a reasonable rough set clustering and it also delivers meaningful results. However, there exist some problems in the algorithm as presented by Lingas and West, such as its numerical instability and its instability in computing the number of clusters. Therefore, Peters made some improvement for the rough cluster algorithm to resolve these problems.

The rough cluster algorithm proposed by Peters use Eq. (3) to calculate the centroids of clusters.

$$m_k = \omega_l \sum_{\mathbf{X}_n \in \underline{C}_k} \frac{\mathbf{X}_n}{|\underline{C}_k|} + \omega_u \sum_{\mathbf{X}_n \in \overline{C}_k} \frac{\mathbf{X}_n}{|\overline{C}_k|} \quad \text{with} \quad \omega_l + \omega_u = 1 \quad (3)$$

where ω_l is the lower weight and ω_u is the upper weight.

The next step is to forces a data object as a lower approximation for each cluster (see Eq. (4)). Then, in order to assign any one of other data objects $\mathbf{X}_n$, except the data objects satisfying Eq. (4), to the lower or upper approximation, looking for the centroid $\mathbf{m}_s$ closest to $\mathbf{X}_n$, so the set T' is determined (see Eq. (5)).

$$d(\mathbf{X}_l, \mathbf{m}_s) = \min_{n,k} d(\mathbf{X}_n, \mathbf{m}_k) \Rightarrow \mathbf{X}_l \in \underline{C}_s \wedge \mathbf{X}_l \in \overline{C}_s \quad (4)$$

$$T' = \left\{ t : \frac{d(\mathbf{X}_n, \mathbf{m}_k)}{d(\mathbf{X}_n, \mathbf{m}_s)} \leq \zeta \wedge k \neq s \right\} \quad (5)$$

- If $T' \neq \phi$, then $\mathbf{X}_n \in \overline{C_t}$, $\forall t \in T'$.
- Else $\mathbf{X}_n \in \underline{C_s}$.

where ζ is the threshold.

3 Comments on Existing Rough k -Means Algorithms

Peters analyzed Lingras and West cluster algorithm from several aspects and then put forward some problems existing in the algorithm. Based on his analysis a refined rough k -means cluster algorithm was proposed. However, there still exist some improvements to be made.

1. Computation of Centroid

As can be seen from Eq. (2), the importance of the lower and upper approximations are defined by the weight ω_l and ω_u respectively. Moreover, Peters suggested a limitation $\omega_l + \omega_u = 1$. Obviously, the weights are determined by end users and not related to the data objects in the lower or boundary area.

2. Numerical stability

Lingras and West algorithm is numerical instable since there are data constellations where $|\underline{C}| = 0$. When $|\underline{C}| = 0$, the cluster C seems to have no sure representative according to the definition of the lower approximation in rough set theory. To avoid such kind of case, Peters suggested that each cluster has at least one lower member and was forced to have a lower member in the initial cluster assignment. Therefore, it is better for a cluster algorithm to assure $|\underline{C}| \neq 0$ whether it forces the lower member for each cluster or not.

3. Interpretation issues and objective function

Peters gave two (extreme) examples of data constellation to illustrate that the objective function of relative distance between data objects was better than that of absolute distance. However, if the objective function is taken into account from other aspects, the data objects in the lower and boundary areas may be explained more intuitively.

4 A Rough k -Means Algorithm Based on a Reasonable Rough Approximation

4.1 Analysis of the Data Objects in the Boundary Areas of Clusters

The data objects that the cluster algorithms deal with are usually multidimensional data sets. Suppose that the data objects are in the multidimensional space now. After a rough k -means cluster algorithm is performed, several cluster means (centroids) that are the representatives of clusters are generated. These centroids are also multidimensional vectors. Note that an arbitrary object A and two cluster centroids (C and B) in the multidimensional space form a triangle, which also decide a plane (see Fig. 1). Moreover, the data objects E, D and A are in the same plane.

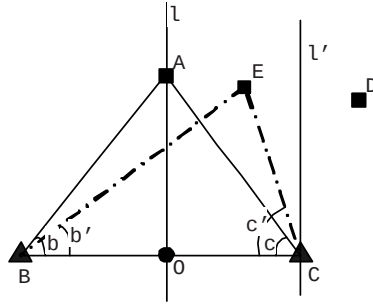


Fig. 1. The boundary area

Suppose

- A is an arbitrary data object. B and C are two cluster centroids. They are all in the multidimensional space. Moreover, C is the cluster centroid closet to A.
- O is the mid-point of the line segment from B to C.
- The straight line l is perpendicular to the line segment from B to C, so is the straight line l' .
- Angle b represents $\angle ABC$ and angle c represents $\angle ACB$. Similarly, Angle b' represents $\angle EBC$ and angle c' represents $\angle ECB$.

Given A is not an outlier. Intuitively, it is so hard to make clear whether the object A belongs to the cluster C or the cluster B when A lies in the straight line l , since the distance from A to C equals that from A to B. In this case, it is easy to find that $c=b$. The closer A gets to C, namely the further A is from B, the easier it is for A to be distinguished. Similarly, the larger the value of $c-b$ is, the further A is from the boundary area (such as the data object E). Obviously, it is more intuitive to assign A to the lower approximation of the cluster C when A is in the right side of l' (such as the data object D). It can be explained that the boundary area between cluster C and cluster B gets to its largest area when c equals the maximum 90. Therefore, a more reasonable rough approximation is suggested in the following to replace the distance measure for determining the set T''' :

- If $(c - b) \leq \varepsilon'$ and $c \leq \vartheta$, then $A \in \overline{C}$ and $A \in \overline{B}$
- Else $A \in \underline{C}$

Accordingly, the weights ω_l and ω_u are changed as follows:

$$\omega_l = \frac{360 - 2\vartheta}{360} \quad \text{and} \quad \omega_u = \frac{2\vartheta}{360}$$

where ε' and ϑ are two given thresholds. Note that $0 \leq \varepsilon'$, $\vartheta \leq 90$. The threshold ε' defines the biggest difference between from A to C and from A to another cluster centroid B. The threshold ϑ determines the weights and reflects the maximum of the boundary area. Note that ϑ must be selected from 0 to 90.

4.2 The Proposed Algorithm

The outline of the rough k -means cluster algorithm based on a reasonable rough approximation (RKMrra) can be stated as follows:

Step 1. Initialization. Randomly assign each data object to exactly one lower approximation. By definition (Property 2, Section 2.1) the data objects in the same cluster belong to both the lower and the upper approximations of the cluster.

Step 2. Calculation of the new cluster centroids according to Eq. (2).

Step 3. Assign the data objects to the lower and upper approximations.

(i) For a given data object $\mathbf{X}_n$ determine its closest centroid $\mathbf{m}_s$:

$$d_{n,s}^{min} = d(\mathbf{X}_n, \mathbf{m}_s) = \min_{k=1,\dots,K} d(\mathbf{X}_n, \mathbf{m}_k) \quad (6)$$

Assign $\mathbf{X}_n$ to the upper approximation of the cluster s : $\mathbf{X}_n \in \overline{C}_s$.

(ii) Determine whether $\mathbf{X}_n$ belongs to other approximations:

– Calculation of the set T''' :

Step 3.1 Initialization. The set T''' is set to ϕ . The set L is set to $\{1, 2, \dots, K\}$.

Step 3.2 $L = L - \{s\}$. For a centroid $\mathbf{m}_j$ ($j \in L$) calculate two angles as follows:

$$\theta_s = \arccos \frac{d(\mathbf{X}_n, \mathbf{m}_s)^2 + d(\mathbf{m}_s, \mathbf{m}_j)^2 - d(\mathbf{X}_n, \mathbf{m}_j)^2}{2d(\mathbf{X}_n, \mathbf{m}_s)d(\mathbf{m}_s, \mathbf{m}_j)}$$

$$\theta_j = \arccos \frac{d(\mathbf{X}_n, \mathbf{m}_j)^2 + d(\mathbf{m}_s, \mathbf{m}_j)^2 - d(\mathbf{X}_n, \mathbf{m}_s)^2}{2d(\mathbf{X}_n, \mathbf{m}_j)d(\mathbf{m}_s, \mathbf{m}_j)}$$

Step 3.3 If $(\theta_s - \theta_j) \leq \varepsilon'$ and $\theta_s \leq \vartheta$ ($0 \leq \vartheta \leq 90^\circ$), then $T''' = T''' \cup \{j\}$. Where ε' and ϑ are two given thresholds.

Step 3.4 $L = L - \{j\}$. If $L \neq \phi$, continue with Step 3.2.

– If $T''' \neq \phi$, then $\mathbf{X}_n \in \overline{C}_t, \forall t \in T'''$

– Else $\mathbf{X}_n \in \underline{C}_s$

(iii) Update the weights ω_l and ω_u according to the following equations:

$$\omega_l = \frac{360 - 2\vartheta}{360} \quad \text{and} \quad \omega_u = \frac{2\vartheta}{360} \quad (7)$$

Step 4. Check convergence of the algorithm.

- If the algorithm has not converged, continue with Step 2.
- Else STOP.

4.3 Analysis of the Rationality of the Proposed Rough Approximation

We analyze the rationality of the proposed rough approximation from the following three aspects:

1. Computation of Centroid

As can be seen from Step 3.3, the threshold ϑ defines the width of the boundary area. Furthermore, it also decides the weights ω_l and ω_u (see Eq. (7)). Therefore the weights are closely related to the boundary area. This leads to the ease of decision made by end users or experts for the parameters ω_l and ω_u .

2. Numerical stability

Unlike the algorithms proposed by Lingras et al. or by Peters, the algorithm proposed above is numerical stable since there doesn't exist that $|\underline{C}| = 0$. Therefore, no data object need to be forced as lower members of clusters. Moreover, each cluster has definite representatives.

3. Interpretation issues and objective function

The objective function (see Step 3.3) is taken into account from the angle aspect instead of from the distance aspect. Moreover, the data objects in the lower and boundary areas are explained more intuitively.

5 Experiments and Discussion

Experiments were conducted on the Web access logs of the introductory first year course in computing science at Saint Mary's University. Lingras and West showed that the visits from students attending these courses could fall into one of the following three categories (for more details see [6]):

1. **Studious:** These students always download the current set of notes regularly.
2. **Crammers:** These students download a large set of notes just before the exam for a pre-test cramming.
3. **Workers:** These group of students are more interested in doing class and lab assignments than downloading the notes.

Since the students in the courses are of different educational backgrounds. Lingras and West decided to use the following five attributions representing each visitor:

1. On campus/Off campus access
2. Day time/Night time access
3. Access during lab/class days or non-lab/class days
4. Number of hits
5. Number of notes downloaded

The values for the first three attributes were either 0 or 1. The last two values were normalized to the interval $[0,1]$ and the last attribute was the most important for clustering visitors.

The total access logs (AllData) have a total size of 21637. We selected 3000 data records (D1) randomly out of the total access logs. Similarly, we got other nine data sets (D2, D3, D4 and so on) with a size of 3000 respectively. The

eleven data sets in all were used for the following experiments. Furthermore the performance of RKMrra is compared to that of two other rough k -means cluster algorithms, which are Lingras and West algorithm and Peters algorithm. To exclude any influence of different selections of the weights, we consider these algorithms with $\omega_l = 0.7$ and $\omega_u = 0.3$, which corresponds to $\vartheta = 54$ in the following experiments. Each algorithm is repeated i -times (iteration factor). When the clustering result doesn't change any more, the cluster algorithm gets to the maximum number of iterations (i_{max}). Among the final results, the experiment with the minimal Davies-Bouldin index (D-B Index) (for more details see [12]) is considered as best.

We focus on the following aspects to evaluate the performance of RKMrra:

- In Section 5.1 we analyze the convergence speed of the algorithm.
- In Section 5.2 we investigate the selections of the thresholds.
- In Section 5.3 we analyze the stability of the algorithm.
- In Section 5.4 we discuss the initial cluster assignment of the algorithm.
- In the last analysis (Section 5.5) we compare the clustering quality of RKMrra with that of the other two algorithms.

5.1 Convergence Speed

In order to evaluate the convergence speed of the cluster algorithms, we conduct 10 experiments on 10 data sets (D1, . . . , D10) for each algorithm. The thresholds are chosen $\varepsilon = 0.5, 0.6, 0.7$ for Lingras and West algorithm, $\zeta = 1.1$ for Peters algorithm and $\varepsilon' = 0.7$ for RKMrra. Table 1 shows the final number of iterations when the result of a cluster algorithm remains stable. The last column is the average number of iterations on 10 data sets for each algorithm.

There are some slightly differences among the average number of iterations of the three algorithms. The average number of iterations of Peters algorithm is the smallest, while that of RKMrra is larger and Lingras and West algorithm has the largest average number of iterations.

In general, these three algorithms have similar convergence speeds.

Note that the threshold ε is chosen as three different values (0.5, 0.6 and 0.7) for Lingras and West algorithm. However, the threshold ζ is chosen as a definite value for Peters algorithm, so is the threshold ε' for RKMrra.

We explain why we choose three different values for the threshold ε for Lingras and West algorithm from the following two aspects:

1. Lingras and West algorithm is very sensitive to the threshold ε

For example, the number of the objects in the boundary area increases too much when the threshold ε changes from 0.63 to 0.64 (see Table 2). The same

Table 1. The number of iterations

Algorithm	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	AverIter
Lingras	40	24	26	36	27	21	28	34	30	32	29.8
Peters	30	29	26	30	29	19	26	26	34	27	27.6
RKMrra	30	26	25	25	26	34	33	28	30	26	28.3

Table 2. The sensitivity to the threshold ε

Data set	The number of the data objects in the boundary area	ε
D1	8	0.61,0.62
	12	0.63
	2598	0.64,0.65,...

cases happen when experiments are conducted on other data sets for Lingras and West algorithm.

2. To maintain the consistency

In order to maintain the consistency of the analysis when comparing with other algorithms, we limit the number of the data objects in the boundary area within 150.

In conclusion, because of the analysis above, the threshold ε for Lingras and West algorithm must be adjusted for different data sets.

5.2 Selections of the Thresholds

To evaluate the influence of the threshold on the algorithm we use the data set (AllData) with the largest size. We analyze the selections of the thresholds from the following two aspects:

1. The range of the threshold ε'

Lingras and West algorithm and Peters algorithm suggested the selections of the threshold ε or ζ respectively. However, they did not point out the range within which ε or ζ should be selected. Therefore, the selections of ε or ζ depend on the decisions of experts or end users. Here we discuss the dependency of the thresholds of the three algorithms. The results are illustrated in Fig. 2.

Lingras and West algorithm shows good performance and similar linear characteristic with RKMrra when the value of the threshold (ε or ε') ranges from 0.2 to 1.0. However, the number of data objects in the boundary area increases too much suddenly so that Lingras and West algorithm delivers no meaningful result for $\varepsilon > 1.0$, so does the same case for Peters algorithm for $\zeta > 1.7$. In contract to Lingras and West algorithm and Peters algorithm, RKMrra still work well even though ε' becomes very large.

In order to better illustrate the range of ε' , we consider four extreme cases as follows:

- When $\varepsilon' = 90$ and $\vartheta = 90$, there exist that $|\underline{C}| > 0$ and $|C^B| > 0$.
- When $\varepsilon' = 90$ and $\vartheta = 90$, there exist that $|\underline{C}| > 0$ and $|C^B| = 0$.
- When $\varepsilon' = 0$ and $\vartheta = 90$, there exist that $|\underline{C}| > 0$ and $|C^B| = 0$.
- When $\varepsilon' = 0$ and $\vartheta = 0$, there exist that $|\underline{C}| > 0$ and $|C^B| = 0$.

where $|\underline{C}|$ and $|C^B|$ are the number of data objects in the lower and boundary areas of clusters respectively.

From the four extreme cases above, we conclude that there must exist data objects in the lower area for any ε' from 0 to 90. Therefore, RKMrra needn't

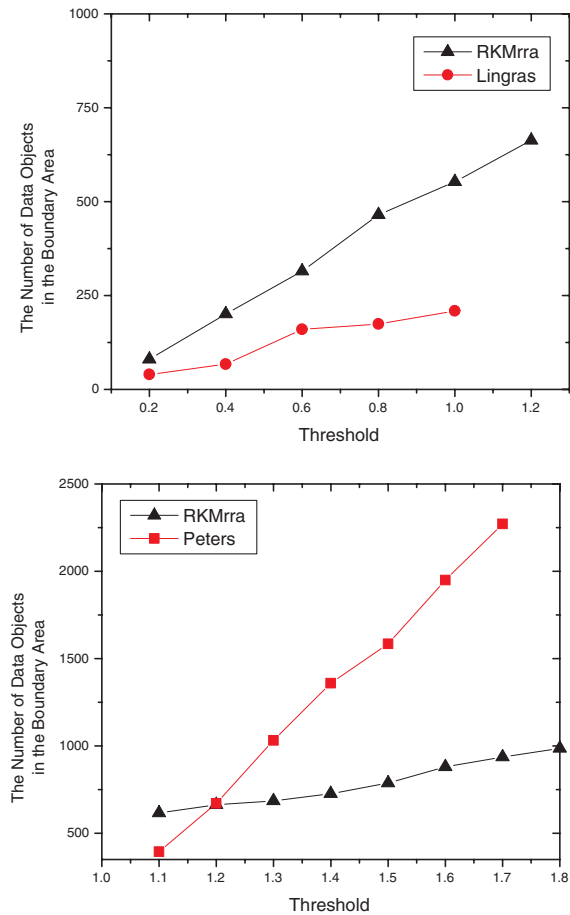


Fig. 2. Dependency on the threshold ε'

force the data objects as the lower members of clusters during the initial cluster assignment. In contrast to RKMrra, it is a necessary step for Peters algorithm (see Eq. (5)).

2. The rationality of the threshold ϑ

The threshold ϑ is used to do two things:

- (1) Computing the weights. (see Eq. (7))
- (2) Controlling the boundary area. (see Step 3.3)

To evaluate the dependency on the threshold ϑ of RKMrra, we remove the threshold ϑ from Step 3.3 and replace the weights (ω_l and ω_u) in Eq. (7) with the weights in Eq. (2). Table 3 illustrates how the results are influenced by the threshold ϑ . The experiment is conducted on the data set AllData.

When ε' is small, the two algorithms have the same clustering results no matter whether the threshold ϑ is removed. As ε' becomes very large, such as the maximum value (90), the Davies-Bouldin indexes of both algorithms change a

Table 3. The dependency of the threshold ϑ

Algorithm	Whether removing the parameter ϑ	ω_l	ω_u	ε'	The number of the boundary objects	D-B index
$RKMrra$	No	0.7	0.3	1.1	616	0.626
$RKMrra'$	yes					
$RKMrra$	No	0.7	0.3	1.2	663	0.628
$RKMrra'$	yes					
...						
$RKMrra$	No	0.7	0.3	90	12433	1.175
$RKMrra'$	yes				90	16532

lot. Moreover, the algorithm of removing the threshold ϑ (*RKMrra'*) has a larger Davies-Bouldin index comparatively. Therefore, it is reasonable to suggest the threshold ϑ to compute the weights and control the boundary area for *RKMrra*.

In general, in comparison to Lingras and West algorithm and Peters algorithm, the parameter ϑ is reasonable suggested by *RKMrra*. At the same time, the thresholds (ε' and ϑ) of *RKMrra* are selected within a reasonable range. As far as the selection of the thresholds within the specified range is concerned, the setting of the thresholds has actually been relaxed. Furthermore, during the initial cluster assignment, it isn't a necessary step to force the data objects as the lower members of clusters for *RKMrra*.

5.3 Stability

We use 10 data sets (*D1*, ..., *D10*) to conduct the experiments. The thresholds are chosen as in Section 5.1. Since the algorithms adjust the assignment of the data objects gradually, we find that the Davies-Bouldin index of each algorithm changes a lot in the experiments. Before the clustering result remains stable, the Davies-Bouldin index sometimes increases and sometimes decreases. That means the Davies-Bouldin index does not always monotonously increase or decrease. For example, the Davies-Bouldin index (D-B index) of an algorithm increases from $i = 5$ to $i = 10$ firstly, then decreases from $i = 10$ to $i = 15$ and increases again from $i = 15$ to $i = i_{max}$. If we use the number of jumps to record the change of the D-B index, then the number of jumps of the example equals 3. Here we analyze the stability of the algorithm from the number of jumps of the D-B index in relation to iteration factors.

Table 4 shows the jumps of the D-B index of the algorithms run on ten different data sets. The iteration factor is set to $i = 5, 10, 15, 20, \dots, i_{max}$ respectively to

Table 4. The number of jumps of the D-B index

Algorithm	<i>D1</i>	<i>D2</i>	<i>D3</i>	<i>D4</i>	<i>D5</i>	<i>D6</i>	<i>D7</i>	<i>D8</i>	<i>D9</i>	<i>D10</i>	<i>AvJump</i>
<i>Lingras</i>	1	2	3	4	1	1	2	2	2	3	$\frac{21}{10} = 2.1$
<i>Peters</i>	1	1	2	3	1	1	2	1	2	3	$\frac{17}{10} = 1.7$
<i>RKMrra</i>	1	2	1	1	4	1	1	2	2	1	$\frac{16}{10} = 1.6$

calculate the number of jumps of the D-B index. The last column of the table is the average number of jumps.

Obviously, RKMrra with the smallest average value has the best stability. Peters algorithm has a close value with the proposed one and Lingras and West algorithm is the most instable.

5.4 The Initial Cluster Assignment

The number of the data objects in the boundary area (the boundary objects) changes a lot for different iteration factors (see Table 5). We use the ratio of the boundary objects for $i = 5(i_{min})$ to those for $i = i_{max}$ to demonstrate this change. Figure 3 shows the ratios of different algorithms run on five data sets (D1, ..., D5). The ratios on other five data sets (D6, ..., D10) are similar to the ratios on the five data sets (D1, ..., D5).

As can be seen from Fig. 3, there is a significant difference between RKMrra and the other two algorithms. The ratio of RKMrra is either greater or less than one, while the ratios of the other two algorithms are both less than or equal one. This shows that the number of the boundary objects of RKMrra for i_{max} is greater than those for i_{min} on some data sets. However, there don't exist such cases for Lingras and West algorithm and Peters algorithm.

Table 5. The ratio of the boundary objects for i_{max} to those for i_{min}

Algorithm	D1		D2		D3		D4		D5	
	i_{max}	i_{min}	i_{max}	i_{min}	i_{max}	i_{min}	i_{max}	i_{min}	i_{max}	i_{min}
Lingras	12	109	9	98	14	77	4	12	48	117
Peters	40	40	47	98	41	63	47	85	70	84
RKMrra	52	69	55	39	44	78	57	94	66	102

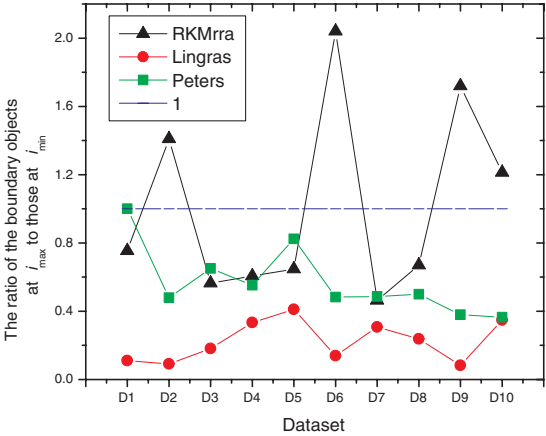


Fig. 3. The ratio of the boundary objects for i_{max} to those for i_{min}

The reason is that RKMrra just adjusts the assignment of the data objects in the boundary area for each iteration factor. In contrast to RKMrra, Lingras and West algorithm and Peters algorithm need to restrict the boundary area within a certain range for the initial cluster assignment firstly, then select the data objects around the limited boundary area.

5.5 Clustering Quality

As introduced above, among the rough k -means cluster algorithms, the one with the minimal Davies-Bouldin index is considered as best. In order to evaluate the clustering quality, We use 10 data sets (D1, . . . , D10) to conduct the experiments. The thresholds are chosen as in Section 5.1. An interesting phenomenon is found among the results of the experiments: The Davies-Bouldin index increases with the number of data objects in the boundary area. That shows the clustering quality of an algorithm is better as the data objects in the boundary area decrease.

The boundary area in rough k -means clustering is also referred to as the security zone [7]. Because the data objects in the boundary area have the possibilities to belong to more than one clusters and require a second look before making a final decision. Hence, the cluster algorithm with the maximum number of data objects in the boundary area indicates the highest security requirements.

Strictly speaking, unlike the classical k -means cluster algorithms, the rough k -means can be interpreted as two layer interval clustering approaches with lower and upper approximations. Therefore, there isn't a kind of cluster validity criterion for the rough k -means cluster algorithms virtually.

Taken into consideration two factors (the D-B index and the number of the boundary objects) analyzed above, here we use the ratio of the Davies-Bouldin index to the number of data objects in the boundary area to evaluate the clustering quality. The results are depicted in the Fig. 4. The algorithm with the minimum ratio is considered as best.

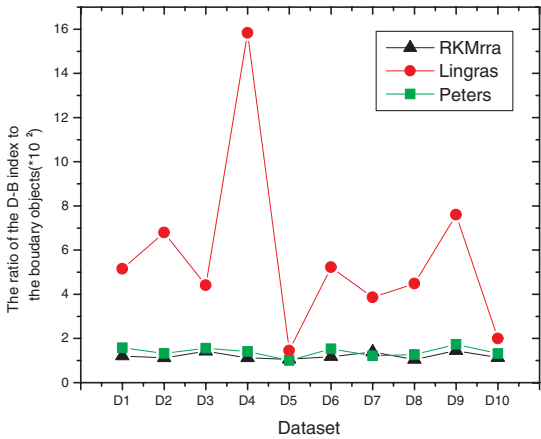


Fig. 4. The clustering quality

Obviously, RKMrra has the similar clustering quality with Peters algorithm. In contract to Lingras and West algorithm, the ratios of RKMrra and that of Peters algorithm change smoothly and have small ratios. Therefore, RKMrra and Peters algorithm have better clustering quality.

6 Conclusion

In this paper we introduce Lingras and West rough cluster algorithm and Peters refined one at first and then comment on them to put forward some problems. In order to solve these problems, the characteristics of the data objects in the boundary area are analyzed. This led to the suggestion of a reasonable rough approximation. The reasonable rough approximation is proposed from the angle aspect, instead of from the distance aspect, although there are some relationships between them. Based on the reasonable rough approximation suggested, a rough k -means cluster algorithm is proposed.

A challenge of the rough k -means is resolved to some extent: the selection of the initial parameters ω_l and ω_u . Since the parameters are limited within a reasonable range according to the threshold ϑ . Furthermore, they are closely related with the width of the boundary area. By tuning the initial parameters, experts can interpret the clustering results according to the given width of the boundary area.

At last, RKMrra is applied to Web logs. The paper describes the design of the experiments to compare RKMrra with Lingras and West algorithm and Peters algorithm with respect to five characteristics. The results show that RKMrra discovers meaningful clusters of Web users and its rough approximation is more reasonable.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No.60475019) and the Ph.D. programs Foundation of Ministry of Education of China (No.20060247039). We are very grateful to Pawan Lingras for supplying with the Web access logs.

References

1. Cooley, R., Mobasher, B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. Tools with Artificial Intelligence. In: Proceedings of the Ninth IEEE International Conference, pp. 558–567. IEEE-CS Press, Los Alamitos (1997)
2. Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.): WI 2001. LNCS (LNAI), vol. 2198, pp. 1–17. Springer, Heidelberg (2001)
3. Zhong, N., Liu, J., Yao, Y.Y. (eds.): Special issue on Web Intelligence (WI). IEEE Computer 35(11) (2002)
4. Ji, J., Liu, C., Sha, Z., Zhong, N.: Online Personalized Recommendation Based on a Multilevel Customer Model. International Journal of Pattern Recognition and Artificial Intelligence 19(7), 895–917 (2005)

5. Pawlak, Z.: Rough Set Theory and Its Applications to Data Analysis. *Cybernetics and Systems. An International Journal* 29, 661–688 (1998)
6. Lingras, P., West, C.: Interval Set Clustering of Web Users with Rough K-means. *Journal of Intelligent Information System* 23(1), 5–16 (2004)
7. Peters, G.: Some Refinement of K-means Clustering. *Pattern Recognition* 39, 1481–1491 (2006)
8. De Kumar, S., Radha Krishna, P.: Clustering Web Transactions Using Rough Approximation. *Fuzzy Set and Systems* 148, 131–138 (2004)
9. Mitra, S.: An Evolutionary Rough Partitive Clustering. *Pattern Recognition Letters* 25, 1439–1449 (2004)
10. Asharaf, S., Murty, M.N., Shevade, S.K.: Rough Set Based Incremental Clustering of Interval Data. *Pattern Recognition Letters* 27, 515–519 (2006)
11. Hogo, M., Snorek, M., Lingras, P.: Temporal Versus Latest Snapshot Web Usage Mining Using Kohonen Som and Modified Kohonen Som Based on the Properties of Rough Sets Theory. *International Journal on Artificial Intelligence Tools* 13(3), 569–592 (2004)
12. Bezdek, J.C., Pal, N.R.: Some New Indexes of Cluster Validity. *IEEE Trans. Systems Man Cybernet Part-B* 28, 301–315 (1998)
13. Kohavi, R.: Mining e-Commerce Data: the Good, the Bad, and the Ugly. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 1(1), pp. 5–32 (2001)
14. Yao, Y.Y.: Information Granulation and Rough Set Approximation. *International Journal of Intelligent Systems* 16(1), 87–104 (2001)

E-Business Intelligence Via MCMP-Based Data Mining Methods

Yi Peng¹, Yong Shi^{2,3,4,*}, Xingsen Li³, Zhengxin Chen⁴, and Gang Kou⁵

¹ School of Management

University of Electronic Science and Technology of China

² CAS Research Center on Fictitious Economy and Data Sciences

Beijing 100080, China

³ School of Management

Graduate University of the Chinese Academy of Sciences

Beijing, 100080, China

⁴ College of Information Science and Technology

University of Nebraska at Omaha

Omaha, NE 68182, USA

⁵ Thomson Co., 610 Opperman Dr, Eagan, MN 55123, USA

Abstract. Organizations gain competitive advantages and benefits through e-Business Intelligence (e-BI) technologies at all levels of business operations. E-BI gathers, processes, and analyzes tremendous relevant data to help enterprises make better decisions. Data mining, which utilizes methods and tools from various fields to extract useful knowledge from large amount of data, provides significant support to e-BI/BI applications. This paper presents an overview of a data mining approach: Multiple Criteria Mathematical Programming (MCMP); describes a real-life application using MCMP; and explains how business users at different levels can benefit from the results of MCMP. Then three application models were presented for efficient implementation of e-BI/BI by MCMP models.

1 Introduction

Business Intelligence (BI) includes a collection of technologies that store, retrieve, process, analyze, report and communicate data and helps business users to better devise and implement business strategies [5]. To put it simple, BI technologies are designed to provide the right information to the right user at the right time [15]. According to Gartner Inc.'s survey of 1300 CIOs, BI is the number two technology priority in IT spending [2]. Furthermore, the recent trend of conducting BI via Internet and grid computing environment, which is so-called e-Business Intelligence (e-BI), makes BI and e-BI an integrating area for both researchers and practitioners [16].

Data mining, which utilizes methods and tools from various fields to extract useful information and knowledge from large volumes of data, provides significant

* Corresponding author.

support to BI applications. Data mining is established on the basis of many disciplines, such as machine learning, databases, statistics, computer science, and mathematics. Each field comprehends data mining from its own perspective and makes its distinct contributions.

The objective of this paper is to provide an overview of several Multiple Criteria Mathematical Programming (MCMP)-based data mining approaches; describe a real-life application using MCMP; and explain how business users, who especially conduct e-BI/BI activities, at different levels can benefit from the results of MCMP.

To achieve these goals, this paper is organized as follows: the first part introduces the basic notions and mathematical formulation of a general MCMP classification model. The second part presents a real-life application of MCMP: credit cardholders' behavior analysis. The third part describes how to use MCMP results in e-BI/BI platforms at different levels of organizations.

2 Multiple Criteria Mathematical Programming-Based Classification Models

This section introduces a generalized multi-criteria programming method for classification. This method classifies observations into distinct groups based on two criteria for data separation. The following models represent this concept mathematically:

Given an r -dimensional attribute vector $a = (a_1, \dots, a_r)$, let

$$A_i = (A_{i1}, \dots, A_{ir}) \in \mathbf{R}^r \quad (1)$$

be one of the sample records of these attributes, where $i = 1, \dots, n$; n represents the total number of records in the dataset. Suppose two groups, G_1 and G_2 , are predefined. A boundary scalar b can be selected to separate these two groups. A vector $X = (x_1, \dots, x_r)^T \in \mathbf{R}^r$ can be identified to establish the following linear inequations [1,14,11]:

$$\begin{aligned} A_i X &< b, \forall A_i \in G_1 \\ A_i X &\geq b, \forall A_i \in G_2. \end{aligned}$$

To formulate the criteria and complete constraints for data separation, some variables need to be introduced. In the classification problem, $A_i X$ is the score for the i^{th} data record. Let α_i be the overlapping of two-group boundary for record A_i (external measurement) and β_i be the distance of record A_i from its adjusted boundary (internal measurement). The overlapping α_i means the distance of record A_i to the boundary b if A_i is misclassified into another group. For instance, in Figure 1 the black dot located to the right of the boundary b belongs to G_1 , but it was misclassified by the boundary b to G_2 . Thus, the distance between b and the dot equals α_i . Adjusted boundary is defined as $b - \alpha^*$ or $b + \alpha^*$, while α^* represents the maximum of overlapping. Then, a mathematical

function $f(\alpha)$ can be used to describe the relation of all overlapping α_i while another mathematical function $g(\beta)$ represents the aggregation of all distances β_i . The final classification accuracies depend on simultaneously minimize $f(\alpha)$ and maximize $g(\beta)$. Thus, a generalized bi-criteria programming method for classification can be formulated as:

(Generalized Model) Minimize $f(\alpha)$ and Maximize $g(\beta)$. Subject to:

$$A_i X - \alpha_i + \beta_i - b = 0, \forall A_i \in G_1 \quad (2)$$

$$A_i X + \alpha_i - \beta_i - b = 0, \forall A_i \in G_2 \quad (3)$$

where $A_i, i = 1, \dots, n$ are given, X and b are unrestricted, and $\alpha = (\alpha_1, \dots, \alpha_n)^T$, $\beta = (\beta_1, \dots, \beta_n)^T$, $\alpha_i, \beta_i \geq 0, i = 1, \dots, n$.

Figure 1 presents a two-group classification scenario with overlapping. Black dots indicate G_1 data objects and stars indicate G_2 data objects. There is one misclassified data object from each group if the boundary scalar b is used to classify the two groups, whereas adjusted boundaries $b - \alpha^*$ and $b + \alpha^*$ separate two groups without misclassification.

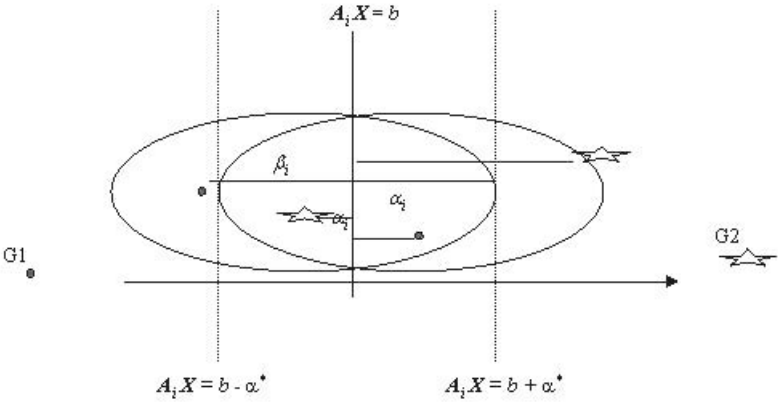


Fig. 1. Two-group classification model

Based on the generalized model, other Multiple Criteria models, such as Multiple Criteria Linear Programming (MCLP) model and Multiple Criteria Quadratic Programming (MCQP) model, can be formulated. For details about these models, refer to: [4,12,13].

3 An Experimental Study: Credit Cardholders' Behavior Analysis

In credit scoring, classification methods can be used to assess applicants' risk and manage existing customers' risk. The goal of this experiment is to classify existing customers into "Good" and "Bad" risk classes. The research procedure

used to conduct the experiment consists of five steps. The first step is *data cleaning*. Missing data and outliers are removed from the dataset. The second step is *data transformation*. The dataset is transformed in accord with the format requirements of corresponding methods or software. The third step is *datasets selection*. Training dataset and testing dataset are selected according to a heuristic process. The fourth step is *model formulation and classification*. A two-group MCLP model was applied to the training dataset to obtain optimal solutions. The solutions were then applied to the testing dataset. Based on these scores, each record is predicted as either Bad (Bankrupt account) or Good (Current account). By comparing the predicted labels with original labels of records, the classification accuracies of multiple-criteria models can be determined. Since the predictive accuracy and stability of a learned model that tests on only one testing dataset is not as good as a model that tests on multiple testing datasets, multiple training and testing datasets were selected to do the classifications. This study uses k-fold cross validation for training and testing datasets selections. The fifth step is *results' presentation*. The acceptable classification results are summarized in tables or figures and presented to end users.

The credit card dataset is provided by a major US bank. It contains 5000 records and 102 variables. The data were collected from June 1995 to December 1995 and the cardholders were from twenty-eight States of USA. Each record has a class label to indicate its credit status: either Good or Bad. Bad indicates a bankruptcy credit card account and Good indicates a good status account. Among these 5000 records, 815 are bankruptcy accounts and 4185 are good status accounts. The variables provide information about credit cardholders' behaviors and can be divided into four categories: balance, purchase, payment, and cash advance. The classification results were summarized in Table 1 [10].

The columns "Bad" and "Good" refer to the number of records that were correctly classified as "Bad" and "Good", respectively. The column "Accuracy" was calculated using correctly classified records divided by the total records in that class. For instance, 80.43% accuracy of Dataset 1 for Bad record in the training dataset was calculated using 563 divided by 700 and means that 80.43% of Bad records were correctly classified. The average predictive accuracies for Bad and Good groups in the training dataset are 79.79% and 78.97% and the average predictive accuracies for Bad and Good groups in the testing dataset are 68% and 74.39%. The results demonstrated that a good separation of bankruptcy and good status credit card accounts is observed with this method.

4 Applications of MCMP Models in e-Business Intelligence/Business Intelligence

4.1 Application Levels

In today's competitive business world, organizations expect e-BI/BI technologies to help optimize decision-makings at all levels. Manglik [8] described the adoption of general BI technologies at four levels of business unites (Figure 2). At

Table 1. MCLP credit card accounts classification

Cross Validation	Training Set (700 Bad +700 Good)			
	Bad	Accuracy	Good	Accuracy
DataSet 1	563	80.43%	557	79.57%
DataSet 2	546	78.00%	546	78.00%
DataSet 3	564	80.57%	560	80.00%
DataSet 4	553	79.00%	553	79.00%
DataSet 5	548	78.29%	540	77.14%
DataSet 6	567	81.00%	561	80.14%
DataSet 7	556	79.43%	548	78.29%
DataSet 8	562	80.29%	552	78.86%
DataSet 9	566	80.86%	557	79.57%
DataSet 10	560	80.00%	554	79.14%
Testing Set (115 Bad +3485 Good)				
	Bad	Accuracy	Good	Accuracy
DataSet 1	78	67.83%	2575	73.89%
DataSet 2	75	65.22%	2653	76.13%
DataSet 3	75	65.22%	2550	73.17%
DataSet 4	78	67.83%	2651	76.07%
DataSet 5	78	67.83%	2630	75.47%
DataSet 6	79	68.70%	2576	73.92%
DataSet 7	77	66.96%	2557	73.37%
DataSet 8	79	68.70%	2557	73.37%
DataSet 9	83	72.17%	2588	74.26%
DataSet 10	80	69.57%	2589	74.29%

the operational level, BI provides detailed information to line managers and help them make day-to-day decisions. The users at the knowledge level are knowledge workers and they may use BI for process improvement, research, or modeling. The management level requires information and knowledge across departments or region and the BI data need to be aggregated and integrated. To support executives to make strategic decisions, BI data are normally historic, highly aggregated, and presented in a clear and understandable way.

The results of data mining can be utilized to support each level of the above BI value chain. Take the credit cardholders' behavior analysis as an example. The classification results provide a list of predicted Bad and Good accounts. At the operational level, this list can be used by credit card issuers to take proactive steps to reduce charge-off loss and promote the right products to the right customers at a timely manner. At the knowledge level, the classification results allow knowledge workers to compare traditional credit card analysis methods with MCMP and find out advantages and disadvantages of each approach. Insights about these tools can improve the practice of selecting the appropriate models to analyze operational data in the future. At the management level, the characteristics of Bad and Good accounts extracted from classification results can help management to design credit products and customer retention plans. At the strategic level, historic classification results can be aggregated and integrated to reveal the general trends of credit cardholders' behaviors. This information can be used by executives to make strategic decisions. Figure 3, adapted from [3], shows the possible business intelligence related levels by a data mining architecture.

Therefore, from data exploration, data mining to knowledge and business intelligence, optimization, including MCMP, can do much work.

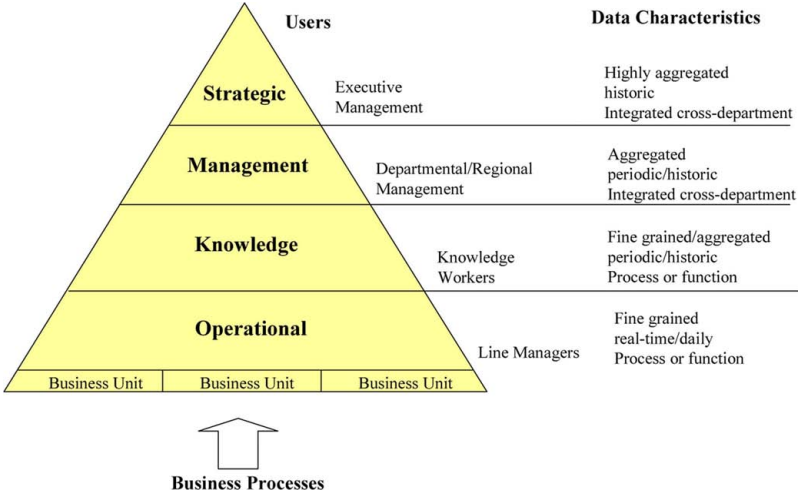


Fig. 2. The business intelligence value chain [8]

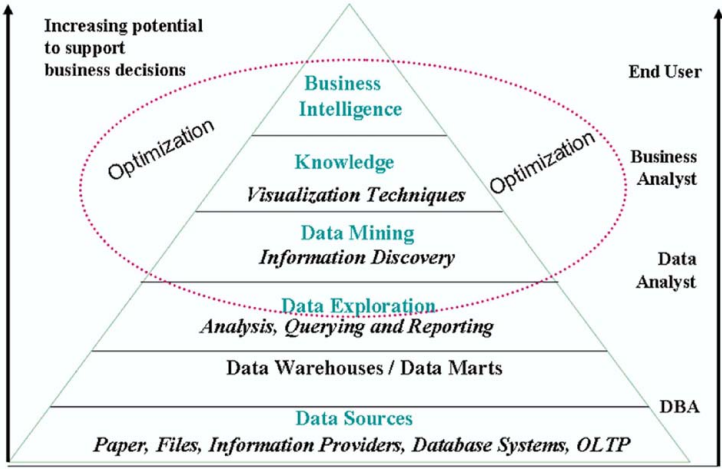


Fig. 3. The business intelligence levels (adapted from Han & Kamber, 2001 [3])

4.2 Application Models for e-BI/BI

Multiple Criteria Mathematical Programming (MCMP)-based data mining can support e-BI/BI in various ways. Here we illustrate the main ideas of three modeling frameworks that we have applied in the real-life applications as below:

Independent Enterprise Internal Application Model. MCMP-based data mining can be developed as software tools and be used independently for knowledge acquisitions. As shown in Figure 4, as long as raw data such as internet data source, offline network data and business source data mostly from MIS are

integrated and processed as input data of data mining, through training and testing several times, a classification model can be obtained to present or visualized the attributes weights. If new data without labels is inputted, it can be first clustered and then classified as a useful list with labels. Both the resulting classifier and the classification results are saved into a knowledge base, through knowledge management platform. All business units will be benefited from them in the processes of planning, execution and monitoring. The benefit will also generate more high-quality data for creating a next level of knowledge discovering circle.

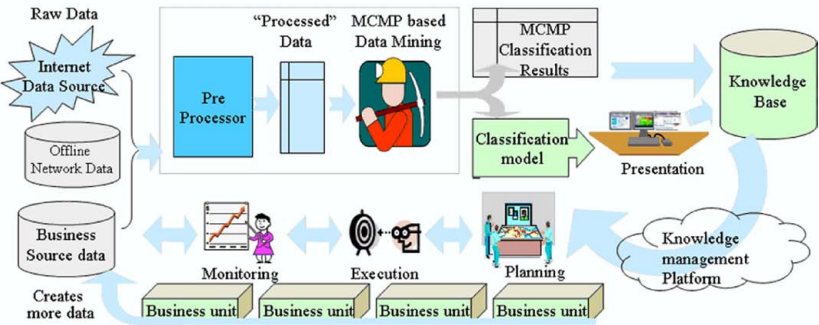


Fig. 4. Application model of MCMP-based data mining for e-BI

This model recently had been implemented at a well-known Chinese web service corporation for its email user churn analysis. Behaviors of the email users include: receiving letters, sending out letters, log on, payment and complaints. An MCMP scoring system scored its customers based on user behavior. Those with higher score were taken as ordinary customer, while low score as churn customers. By reasonable explanation of the model, we have helped the enterprise to improve the customer relationship and reduce the churn of customers [9].

Component-Based Internal Application Model. If the classification model is proved efficiently in the application of the above model, it can be developed as a business software component which can be integrated in business operation systems. Its working process is shown in Figure 5.

Here, the classification model in business working process is combined with data extraction component. When necessary data is created, it can be processed as input data for the model. The model component then classifies the records and produces real time-alert information, controlled information, knowledge or real time suggestion for real time management such as alerting, monitoring, decision making and other wise actions. Through such a knowledge management platform, all business units can work more wisely. These will also make just-in-time knowledge management possible.

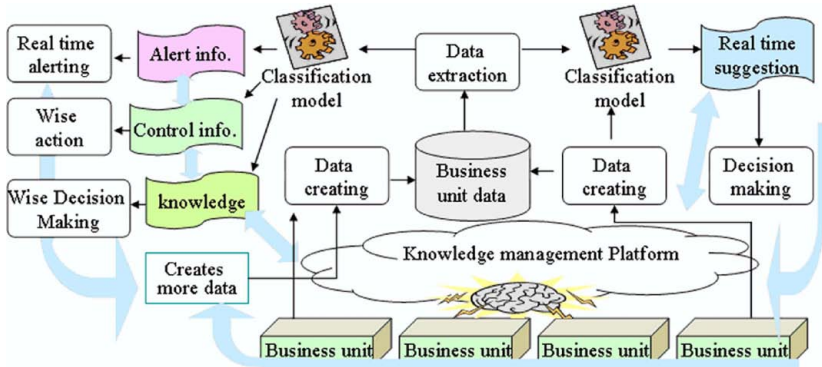


Fig. 5. Component-based internal application model

Web-Based Application Service Model. An application service provider (ASP) is a firm that offers individuals or enterprises access over the Internet to applications and related services that would otherwise have to be located in their own personal computers and enterprise servers. ASP is responsible for the management of the application system from building and maintenance to upgrade of the application. As the revolution of Internet is pushing the society from the global economy environment to e-Business era, ASP mode may re-direct future development of data mining applications.

In this model, as shown in Figure 6, the MCMP based data mining software can be distributed on the Internet and can employ application service for other users outside the enterprise. Through a web page, users can register, make payment by using a user id and password, when login correctly. Users can also

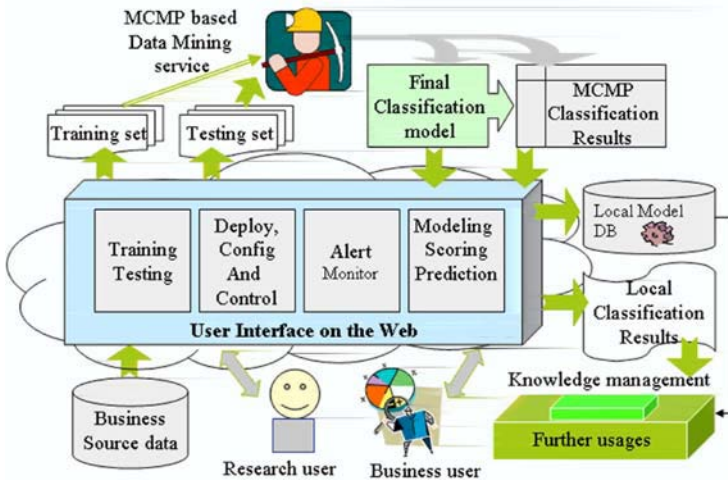


Fig. 6. Web-based application service model

extract data from business database, form needed training data set and testing data set, and obtain the best classifier and classification results which can be saved in their local machine for further usages according to their authority. This ASP model can make data mining tools available for many users, such as researcher or practitioners, over a grid computing environment. Especially, it will be useful for the middle and small businesses who cannot offer the expensive data mining software [6].

4.3 Strategy Implementation

Knowledge is the key factor for e-BI/BI. Based on our research project on various data mining methods and real life practice of management application, a serious of knowledge management platform has been presented [7]. Referred to these models, we designed an e-BI implement strategy as described in the following:

- Step 1.* Communicate with the directors or CTO/CIO of the business so that they can realize the value of e-BI and make a long-term e-BI strategy.
- Step 2.* Apply data mining in a specific field and achieve a valued business target. Accumulate experience and knowledge then train work teams.
- Step 3.* Make data mining with a common sense in most business units and form a knowledge management platform on intranet for component based integration of business intelligence. The platform collects the experts' experience in daily works with the knowledge from data mining and accumulates knowledge day after days.
- Step 4.* Provide out-sourcing services to other companies that have no data mining experiences. Such as conducting data mining consulting project, marketing e-BI proposals or providing certain knowledge or data to those organizations that need it. Finally, the ASP mentioned above can become a new profit source.

Our applications in some Chinese corporations have shown that it is effective to implement of an e-BI application by MCMP-based data mining tools.

5 Conclusion

Producing e-business intelligence/business (e-BI/BI) intelligence by using data mining has been attracting an increasing attention from organizations. It is important to realize that data mining results can be useful at different levels of e-BI/BI architecture. This paper has described a data mining approach: Multiple Criteria Mathematical Programming (MCMP) and its application in credit cardholders' behavior analysis. Based on the BI value chain proposed by [8], this paper has explained how the results of MCMP can be utilized by different levels of organizations. Although the models presented in this paper have been approved to be useful, there is a lot of work to be done for the improvement of the e-BI/BI platforms that can be available for various businesses.

Acknowledgments

A portion of this paper has been published as an invited chapter in *Advanced Topics in Data Warehousing and Mining 2007* by the IGI Publisher. This research has been partially supported by grants #70621001, #70531040, #70472074, National Natural Science Foundation of China; 973 Project #2004CB720103, Ministry of Science and Technology, China; and BHP Billiton Co., Australia.

References

1. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7, 179–188 (1936)
2. Gartner, EXP. Delivering IT's Contribution: The 2005 CIO Agenda, Gartner Executive Programs (January 14, 2005)
3. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA (2001)
4. Kou, G., Peng, Y., Shi, Y., Wise, M., Xu, W.: Discovering Credit Cardholders Behavior by Multiple Criteria Linear Programming. *Annals of Operations Research* 135(1), 261–274 (2005)
5. Kudyba, S., Hoptroff, R.: *Data Mining and Business Intelligence: A Guide to Productivity*. Idea Group Publishing, USA (2001)
6. Li, X., Liu, Y., Li, J., Shi, Y., Zhang, Y.: A Knowledge Management Model for Middle and Small Enterprises. In: *Proceedings of 2006 International Symposium on Distributed Computing and Applications for Business, Engineering, and Sciences*, Hangzhou, China (October 2006a)
7. Li, X., Shi, Y., Liu, Y., Li, J., Li, A.: A Knowledge Management Platform for Optimization-based Data Mining. In: *Optimization-based Data Mining Techniques with Applications Workshop at Sixth IEEE International Conference on Data Mining*, Hong Kong, China (December 2006b)
8. Manglik, A.: Increasing BI Adoption: An Enterprise Approach. *Business Intelligence Journal*, Second Quarter 11(2), 44 (2006)
9. Nie, G., Zhang, L., Li, X., Shi, Y.: The Analysis on the Customers Churn of Charge Email Based on Data Mining – Take One Internet Company for Example. In: *Optimization-based Data Mining Techniques with Applications Workshop at Sixth IEEE International Conference on Data Mining*, Hong Kong, China (December 2006)
10. Peng, Y., Kou, G., Chen, Z., Shi, Y.: Cross-validation and Ensemble Analyses on Multiple-Criteria Linear Programming Classification for Credit Cardholder Behavior. In: *Craw, S., Preece, A.D. (eds.) ECCBR 2002. LNCS (LNAI), vol. 2416*, pp. 931–939. Springer, Heidelberg (2002)
11. Olson, D., Shi, Y.: *Introduction to Business Data Mining*. McGraw-Hill, Irwin (2007)
12. Shi, Y., Peng, Y., Kou, G., Chen, Z.: Classifying Credit Card Accounts for Business Intelligence and Decision Making: A Multiple-Criteria Quadratic Programming Approach. *International Journal of Information Technology and Decision Making* 4(4) (2005)
13. Shi, Y., Peng, Y., Kou, G., Chen, Z.: Introduction to Data Mining Techniques via Multiple Criteria Optimization Approaches and Applications. In: *Advanced Topics in Data Warehousing and Mining*, IGI Publisher (2006)

14. Shi, Y., Wise, M., Luo, M., Lin, Y.: Multiple Criteria Decision Making in Credit Card Portfolio Management. In: Koksalan, M., Zionts, S. (eds.) Multiple Criteria Decision Making in New Millennium, pp. 427–436. Springer, Berlin (2001)
15. Voe, L.D., Neal, K.: When Business Intelligence Equals Business Value. *Business Intelligence Journal* 10(3), 57 (2005)
16. Zhong, N., Liu, J.M., Yao, Y.Y.: Envisioning Intelligent Information Technologies from the Stand-Point of Web Intelligence. *Communications of the ACM* 50(3), 89–94 (2007)

Intelligence Metasynthesis in Building Business Intelligence Systems*

Longbing Cao¹, Chengqi Zhang¹, Dan Luo¹, and Ruwei Dai²

¹Faculty of Information Technology, University of Technology, Sydney, Australia

²Institute of Automation, Chinese Academy of Sciences, China
{lbcao, chengqi, dluo}@it.uts.edu.au

Abstract. In our previous work, we have analyzed the shortcomings of existing business intelligence (BI) theory and its actionable capability. One of the works we have presented is the ontology-based integration of business, data warehousing and data mining. This way may make existing BI systems as user and business-friendly as expected. However, it is challenging to tackle issues and construct actionable and business-friendly systems by simply improving existing BI framework. Therefore, in this paper, we further propose a new framework for constructing next-generation BI systems. That is *intelligence metasynthesis*, namely the next-generation BI systems should to some extent synthesize four types of intelligence, including data intelligence, domain intelligence, human intelligence and network/web intelligence. The theory for guiding the intelligence metasynthesis is *metasynthetic engineering*. To this end, an appropriate intelligence integration framework is substantially important. We first address the roles of each type of intelligence in developing next-generation BI systems. Further, implementation issues are addressed by discussing key components for synthesizing the intelligence. The proposed framework is based on our real-world experience and practice in designing and implementing BI systems. It also greatly benefits from multi-disciplinary knowledge dialog such as complex intelligent systems and cognitive sciences. The proposed theoretical framework has potential to deal with key challenges in existing BI framework and systems.

1 Introduction

Business intelligence (BI) and business intelligence systems (BIS) [27] play more and more important roles in business operational analysis and decision support. In practice, business intelligence has emerged as a relatively mature field with plenty of tools and techniques available. It has been widely used in areas such as telecommunications, banking, retail and education for reporting and data analysis. However, real experience in building business intelligence for telecommunications also shows limitations of traditional business intelligence techniques, which mainly generate reports through integrating data sources into data warehouse

* This work is sponsored by Australian Research Council Discovery and Linkage Grants (DP0773412, LP0775041, DP0667060), and UTS internal grants.

[7,10]. As discussed in Section 2, some major weaknesses have been reflected in client feedbacks from intelligence development perspective, for instance, incomplete intelligence involvement and insufficient intelligence emergence.

These issues request further cutting-edge and applied research on intelligence development in BIS. In our previous work, we have studied some of them and corresponding solution initiatives. For instance, in [10,14], we proposed *ontology-based business intelligence system* that may integrate business ontologies, data warehouse ontologies and enterprise information system ontologies. In developing actionable knowledge discovery, we proposed *domain driven data mining* [15,16] to involve domain intelligence into data intelligence mining. To support the human-machine interaction and human intelligence involvement in BI development, we studied *agents and multiagent system* techniques [35].

In particular, the methodology of *metasynthesis from qualitative to quantitative* [31,32] and the techniques for *metasynthetic engineering of intelligent systems* [21] offer a powerful framework for us to metasynthesize complementary intelligence in BI development. Following these methodologies and techniques, an effort on developing *Integrated Business Intelligence Reporting Systems*¹ was led and investigated by Dr Longbing Cao when he was Chief Technical Officer with a BI company in China.

Following all above efforts, this paper further summarizes and proposes a framework for intelligence metasynthesis in building BIS. This will lead to the development of next-generation BIS, called Metasynthetic Business Intelligence Systems (MBIS). The MBIS integrates data intelligence, domain intelligence, human intelligence and web intelligence to complement existing data and business intelligence. It further encloses services and intelligences for developing actionable decision-making support and services. The benefits from this approach consist of enhanced BIS with capability of tackling some of major issues in existing BIS as aforementioned.

The remaining sections are organized as follows. Section 2 discusses some of major issues in existing BI framework. In Section 3, we briefly introduce the methodology of metasynthesis from qualitative to quantitative and techniques for metasynthetic engineering of intelligent systems. A framework of intelligence metasynthesis for building BIS is presented in Section 4. Section 5 further discusses corresponding implementation strategies for intelligence metasynthesis. We conclude and present future work in Section 6.

2 Issues in Building Business Intelligence Systems

In our previous work [5,10,14], we discussed some issues of existing BIS from information integration and integration of data warehousing, reporting and data mining. Here we further summarize some of major problems from the perspective of intelligence integration and operational use in BI development and business decision making.

¹ Funded by China Innovation Fund for Small Technology Based Firms (04C26211100957).

- Incomplete intelligence involvement, one of the major issues in existing BIS is that the intelligence available from BIS is actually very limited. The existing BIS mainly concentrate on data, while data intelligence has not been fully discovered. However, BI development may greatly benefit from other types of intelligence such as domain intelligence, human intelligence and web intelligence which are closely related to BI.
- Insufficient intelligence emergence, the emergence capability of existing BIS is actually very limited. In most cases, the reports, which are the major intelligence carrier, repeatedly present business users content in terms of predefined attributes. New intelligence, for instance new customer demographics, rarely or hardly emerges.
- Inconvenient intelligence interaction and delivery, the delivery of intelligence from BIS to business people is not user friendly and business friendly, and inflexible. This is embodied in aspects such as unfriendly human-BIS interaction, dispersed BI components from varying vendors, and lack of automated and/or interactive intelligence integrator that combines inputs of data intelligence, domain intelligence, human intelligence and web intelligence. For instance, there is no decision-support report generator to integrate finding tables, graphs, as well as user inputs into a uniform and neat decision-support report.
- Lack of integrative workflow support, the existing BIS do not offer visually interactive workflow and process management support that may be helpful for users to build up and embed BI processes with underlying enterprise information systems into a one-stop lifecycle.
- Limited actionable knowledge for effective operational analysis and decision support, existing BIS are still lack of providing knowledge so that business users can take actions straightway. Very often many reports generated but few of them really capture hidden and significant intelligence.

To tackle these problems, we propose the following strategic solutions:

- expanding existing BI framework by involving other complementary intelligence,
 - designing an intelligence metasynthesis framework to synthesize relevant intelligence, and
 - developing corresponding techniques to support intelligence metasynthesis.
- We outline them in the following sections.

3 Metasynthetic Engineering

In 1991, the most renowned Chinese scientist Dr Xueseng Qian and his colleagues proposed a new science field named Open Complex Giant Systems (OCGS) [31,32]. They further proposed a powerful theory for dealing with OCGS, that is the Metasynthesis from Qualitative to Quantitative (MQQ), namely Metasynthetic Engineering (ME) [31,32,21]. They further pointed out that a practical roadmap for building problem-solving systems for OCGS is to construct a Hall

for Workshop of Metasynthetic Engineering (HWME) [31,32,21]. With regard to HWME, please refer to the following references [2,3,4,5,6,8] for more information. Following this feat, a wave of research and development on OCGS and the metasynthesis methodology have been emerged in China. Some typical actions include a monograph on Metasynthesis of Intelligent Systems (MIS) [21], a large NSF China grant on "*Research of Human Computer Cooperated Metasynthetic Engineering for National Macroeconomic Decision Support*"² led by Prof Ruwei Dai, and preliminary research on multi-agent-based open giant intelligent systems [3,5,6,8] and open complex agent systems [11,12,13].

The OCGS present main characteristics such as openness, emergence, giant, distribution, societal feature, and human involvement [31,32,21,3,6,10]. A typical example of OCGS is the Internet [9]. As an effective methodology for dealing with such challenging problems, the theory of MQQ and HWME was summarized based on Dr Qian's luxuriant experience and expertise in building real OCGS such as the very famous giant nuclear-weapons and space engineering, so-called "LiangDan YiXing" Program. It also benefits from Dr Qian's metasynthetic intelligence in multi-disciplines crossing a large portion of his systems of modern science and technology consisting of basic sciences, technical sciences and applied techniques [30], as well as broad interaction and integration in the system, in particular, artificial intelligence and pattern recognition, cognitive sciences, system sciences, information sciences as well as social sciences.

The methods effective for dealing with OCGS are beyond reductionism. Studies and practice have clearly proven that the most feasible and effective way is the ME techniques, which are extracted, generalized and abstracted from practical studies. The theory of MQQ and HWME is as follows. Based on the characteristics of the human mind, the human cognitive process and the current computer mechanism, we can find out that an autonomous intelligent system simply composed of computers cannot be the problem-solver of OCGS. A feasible design methodology for tackling OCGS-like problems is to combine human qualitative intelligence, for instance intuition, inspiration and human pattern recognition ability, which cannot be simulated by current computer, with quantitative intelligence, such as analysis, logical reasoning, etc., which can be efficiently simulated by computers [21,3,8]. This is to say that problem-solving systems of OCGS-like problems should be based on *human-computer cooperation*, where both human intelligence and machine intelligence are complementary and collaborative. Further, the problem-solving systems of OCGS problems are essentially human-centered, indicating that human intelligence plays a major role in dealing with OCGS. This is determined by not only the limited capability of machine intelligence in solving OCGS, but the essence of OCGS in which there could be no hints for what to do and how to do regarding the problem-solving. As a result, problem-solving systems tackling OCGS-like problems are actually "human-technology symbionts" or "natural-born cyborgs" [18,19].

² Large Grant from the National Science Foundation of China (NSFC), 79990580, 1999-2003.

The engineering approach of ME is to build an HWME for specific problems. The essence of the MQQ is to organically unite the expert group, data and all sorts of information, and to unite scientific theories of various disciplines and human experience and knowledge [31,32,21,3]. This makes a workspace in itself. In this workspace with friendly human-computer interfaces, human experts discuss or even debate with each other, invoke or create models or methods to prove their own opinions, evaluate and analyze requirements and aims, in proper sequences. An encyclopedical chairman is in charge of the problem-solving process. Figure 1 illustrates the above principle. A successful application of the ME theory depends on a full play and the interplay of the synergetic advantages of the above three factors.

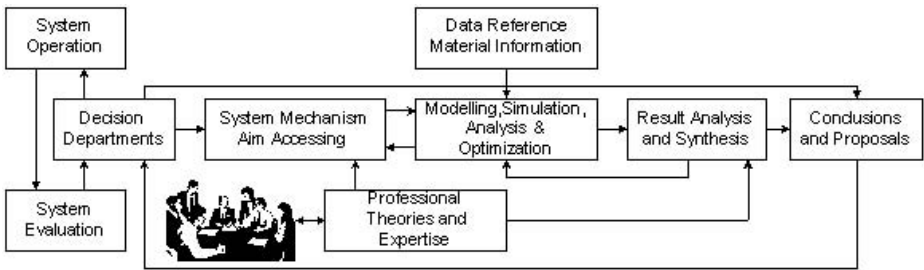


Fig. 1. Process of Metasynthetic Engineering

Its key idea for the problem-solving of OCGS-like problems is to follow the roadmap of "human-centered human-machine cooperation" [32,21,3,6,8] through constructing a Hall for Workshop of Metasynthetic Engineering(HWME) [2,5], which is substantially a human-computer-cooperated intelligent information system. Figure 2 illustrates layered model of a network-based HWME. In Figure 3, major services and functions for implementing metasynthesis in the HWME are listed [5].

It is possible to design such HWME [3,2,5,6,8]. Figure 4 further illustrates major services and functionalities of a HWME-based problem-solving system. It consists of five centers.

- A Metasynthetic Support center, this is the place that synthesizes information and intelligence. It includes domains-specific modeling and method systems, consensus-building applications, knowledge development environment, visual modeling system, and information search engines retrieving information inside or outside the hall portal.
- A Seminar center, this is the place that supports the workshop and seminars in the problem-solving. It provides different forms of discussion templates, script builders, distributed sub-seminars integrator, and management of discussion flow, system status, roles, discussion record and concluding report.
- A Data and Knowledge center, this is the place to connect, manage and transform data, information and knowledge required in the problem-solving.

It may be composed of distributed data sources, system databases, metadata, meta-knowledge, and database management systems.

- A Security center provides services for firewalls, proxies, intrusion detection, access control, privacy processing and protection in distributed heterogeneous environment.
- An Information Cooperation center, which supports internal and external message passing and communication, including mailing services, notification services, schedule system, and office support system, etc.

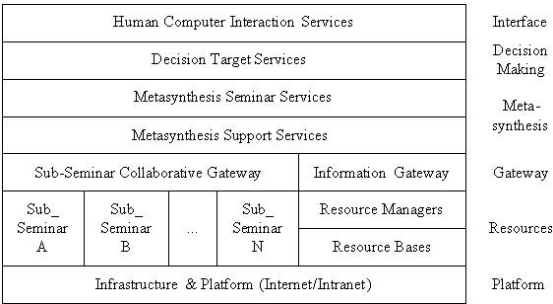


Fig. 2. Layered Model of A Network-based HWME

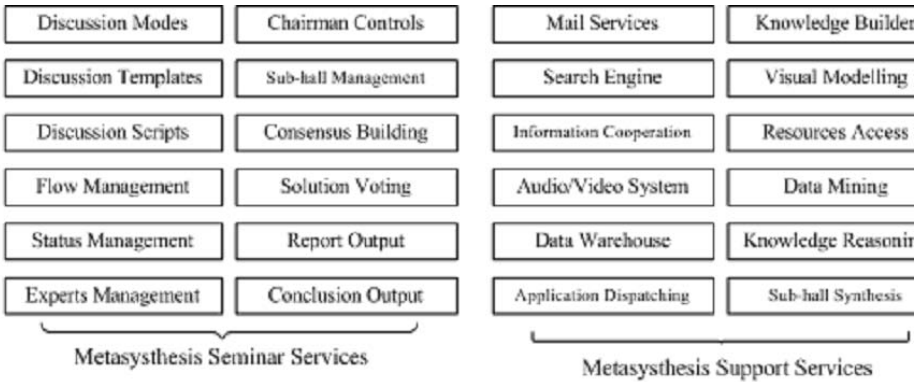


Fig. 3. Metasynthesis services

4 A Framework for Intelligence Metasynthesis

Real-world experience tells us that it is significant and essential to involve complementary intelligence into existing BIS in order to solve some of fundamental issues. Further, it is necessary for us to develop an appropriate framework to integrate such intelligence. In this section we attempt to answer these questions by presenting a framework based on the metasynthetic engineering theory.

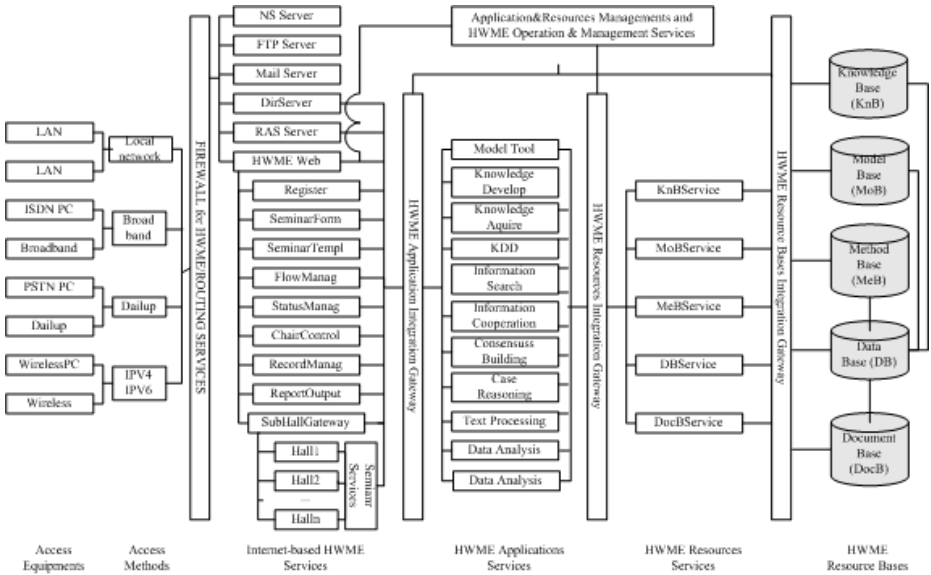


Fig. 4. Structure of the Network-based HWME

4.1 Intelligence Metasynthesis

To enhance existing BIS, we highlight the following four types of intelligence: data intelligence, domain intelligence, web intelligence and human intelligence, as well as the interaction and metasynthesis between them. Their major components are illustrated in Figure 5. We explain them in the following sections in detail. In particular, we believe the intelligence metasynthesis of these intelligences will promote the paradigm shift from the existing BI framework to next generation BIS, which highlights the roles and involvement of as well as the interaction and intelligence emergence between the four types of intelligences.

4.2 Discovering Data Intelligence

Traditionally, BI mainly focuses on developing reports including predefined tables, ad-hoc tables and OLAP reports. We emphasize the concept of *data intelligence*, that is to let data tell a story about a problem. This is not achievable simply using reports. Major effort should be on developing appropriate data mining algorithms to understand and analyze varying data structures and patterns. The extracted patterns have potential to disclose hidden causes and principles helpful for interpreting business problems. Therefore, major efforts in data intelligence development are on designing effective and efficient algorithms that serve better understanding of the data and the problem.

Actions on discovering data intelligence can be conducted in many aspects, in particular developing methodologies and methods in terms of data-centered aspects.

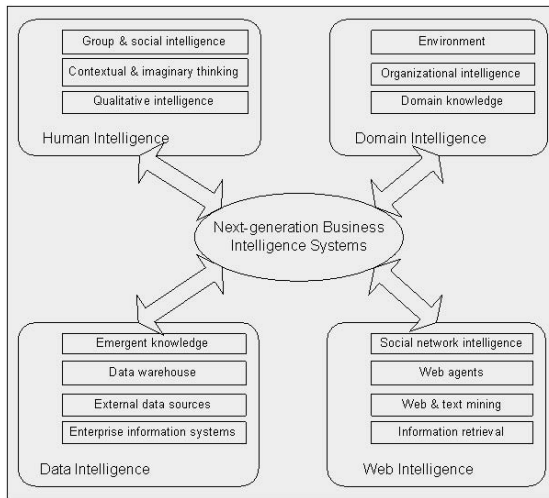


Fig. 5. Intelligence metasynthesis in building next-generation BIS

- Data type such as numeric, categorical, XML, multimedia, composite
- Data timing such as temporal, time-series and sequential
- Data space such as spatial and temporal-spatial
- Data speed such as data stream
- Data frequency such as high frequency data
- Data dimension such as multi-dimensional data
- Data relation such as multi-relations and links

Data intelligence development has been a major effort of data mining community. We do not extend discussions on this issue. Rather, as addressed in Section 5.2, actionable knowledge discovery has been an emerging challenge of data mining. We present a new research topic namely domain driven data mining which aims to involve domain intelligence and human intelligence into data mining to enhance the actionability of discovered knowledge. Compared with traditional data mining, domain driven in-depth data intelligence mining is to discover knowledge that not only presents hidden patterns in the data, but supports actions taken for smart business decisions. For instance, in [29], we illustrate how in-depth data mining can discover deep patterns from generic trading rules. The identified patterns can satisfy not only technical significance but business expectations such as "beat market".

4.3 Developing Domain Intelligence

The power of data intelligence is limited if it is not considered with the involvement of domain intelligence. *Domain intelligence* refers to domain resources that not only wrap a problem and its target data but also assist in the understanding and problem-solving of BI development. Domain intelligence may extremely

complement data intelligence by providing necessary environment and scenario information useful for deep understanding of the data and the problem.

Domain intelligence consists of qualitative intelligence and quantitative intelligence. Both qualitative and quantitative intelligence is instantiated in terms of aspects such as domain knowledge, background information, constraints, organization factors, business process, and so on. In a summary, we categorize domain intelligence in terms of the following major aspects.

- (1) Domain knowledge
 - Including domain knowledge, background and prior information,
- (2) Constraint intelligence
 - Including constraints from system, business process, data, knowledge, deployment, etc
 - Privacy
 - Security
- (3) Organizational intelligence
 - Organizational factors
 - Business process, workflow, project management and delivery
 - Business rules, law, trust
- (4) Environment intelligence
 - Relevant business process, workflow
 - Linkage systems
- (5) Deliverable intelligence
 - Profit, benefit
 - Cost
 - Delivery manner
 - Business expectation and interestingness
 - Embedding into business system and process

Relevant work needs to be studied such as user modeling, domain knowledge acquisition, representation and fusion with data mining, semantic and ontological information management, the embodiment of business rules and processes, etc. We will further implementation strategic issues in Section 5.

4.4 Involving Human Intelligence

The role of humans is proved to be significant in BI development. More generally, human intelligence may play essential and irreplaceable roles in BI development. Possible roles that can better played by human intelligence [31,32,21,18,19] include but are not limited to the following aspects.

- Explicit or direct involvement of humans, referring to intelligence information acquired from human sources through overt collection techniques into a system
 - Empirical knowledge, for example, selecting favorite models to deal with specific problem scenario

- Belief, intention and expectation, for example, setting specific interest range as business expectation benchmark
- Run-time supervision, for instance, setting and tuning parameters and models at run time
- Evaluating
- Expert group
- Implicit or indirect involvement of human intelligence, referring to intelligence information acquired from human sources through covert collection techniques into a system
 - Imaginary thinking
 - Emotional intelligence
 - Inspiration
 - Brainstorm, learning from group intelligence
 - Reasoning inputs
 - Social and contextual intelligence adapting to environment

4.5 Supporting Network Intelligence

Network/web intelligence is a new research direction, which explores the fundamental roles as well as practical impacts of artificial intelligence and advanced information technology for the next generation of network/web-empowered systems, services, and environments [37]. Network/web intelligence presents as an external data and intelligence sources to BIS. In this aspect, the following network/web intelligence techniques may be useful for supporting BI.

- Information retrieval and text mining for actively searching and discovering related information to complement BI decision making.
- Web mining for mining BI crossing opponent web pages, and information about competitive products, services, opponents, etc.
- Semantic web and ontological engineering techniques for specifying and transforming terms and relationships crossing multiple intelligence domains.
- Web knowledge management for crossing language BI-related knowledge representation, sharing, transformation and mapping on the web.
- Web agents for services recommendation, message communications, information collection, customer caring, etc.
- Social network intelligence for specific BI problem-targeted link mining, network and community analysis, association and correlation analysis, etc.
- Agent-based infrastructure for building up BIS by involving web intelligence discovery, information retrieval, recommendations, services, and so on.

5 Implementing Intelligence Metasynthesis

Certainly there are many things to be researched in developing each of specific intelligence, for instance, link analysis algorithms for serving business intelligence. On a higher level, a big issue is to implement the intelligence metasynthesis in building next-generation BIS. To this end, both local and global levels of issues

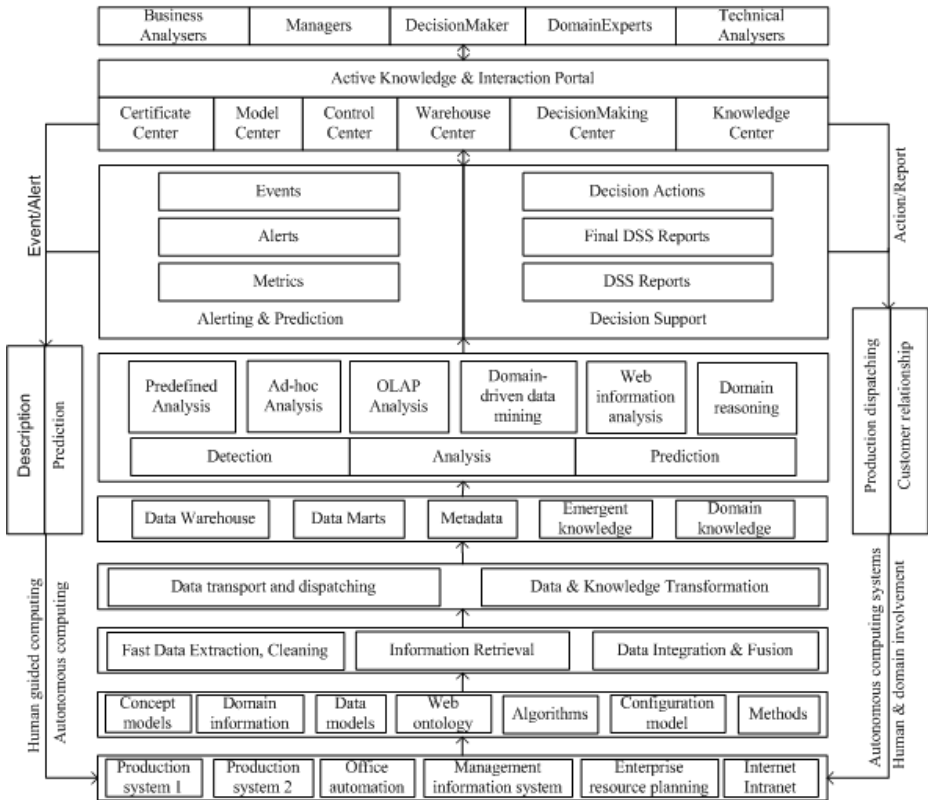


Fig. 6. Metasynthetic business intelligence systems

must be addressed. In the following, we first present a system framework that integrates the aforementioned intelligences. We then highlight four aspects of developing each type of intelligence in BIS.

5.1 Metasynthetic Business Intelligence Systems

Besides some of functionalities and services addressed in HWME supporting human intelligence involvement, a MBIS may consist of the following services and support. Figure 6 exhibits a metasynthetic system framework for intelligence metasynthesis in designing and implementing Metasynthetic BIS.

- Knowledge Specification and Transformation services, which represent and transform domain knowledge and discovered knowledge, and further involve the knowledge into BIS;
- Seminar services, which support the dialogue between human and computer systems, as well as between humans;
- Data and Knowledge services;

- Security services;
- Coordination services, which support the coordination, communication, dispatching and planning between system components/modules, and manage system workflow and processes;
- Reporting services, which present data in terms of reports, tables, graphs;
- Human-Machine Interaction services, which provide interfaces and services supporting human-machine interaction, especially ad-hoc and run-time dialogue;
- Web Information and Knowledge services, which look after web information retrieval, manage web knowledge, implement web services.

Many things should be done to support this architecture. In the remaining of this section, we emphasize the following points that are essentially important for MBIS.

- Developing ontology services to represent, transform, map and discover intelligence intra or inter business domain, problem-solving domain, and data source domain.
- Studying domain driven data mining to discover in-depth data intelligence, and involve domain intelligence, human intelligence, and web intelligence into BI systems for more actionable knowledge discovery.
- Discovering web intelligence to complement BI and supply additional decision-making resources.
- Implementing agent-based MBIS infrastructure to conduct intelligence metasynthesis in BIS.

5.2 Ontology Services for Intelligence Representation and Transformation

One of the major issues in designing MBIS is the representation and transformation of intelligence from one domain to another, for instance, transferring domain knowledge from humans into terms understandable to the system. Ontological engineering offers a suitable technique for dealing with intelligence representation and transformation in BI area [23,17,24,28,10,14,34]. In particular, we need to develop ontology services to support an ontology-based MBIS. Ontology services may serve the following functions.

- Representing domain ontologies for domain knowledge and business concepts in business domain
- Representing problem-solving ontological items
- Representing semantic and structural relationships in or across multiple domains
- Representing meta-ontology items, as well as semantic and structural meta-relationships
- Transforming ontological items crossing ontological domains
- Mapping business terms to problem-solving ontological items
- Aggregating concepts and ontological items

- Discovering ontology patterns intra and inter ontological domains

To develop the above ontology services, appropriate ontological engineering techniques must be studied. This may include the following aspects.

- Ontology-based system architecture, which deals with an ontological framework to manage relevant concept namespaces and ontology transformation and mapping among domains;
- Developing appropriate ontology representation mechanisms;
- Specifying semantic and structural relationships in BI area for a specific problem domain;
- Designing aggregation rules for ontology and relationship aggregation intra and inter ontology domains;
- Designing transformation rules for ontology and relationship transformation intra and inter ontology domains; and
- Ontology mining aiming to ontology patterns intra and inter domains.

5.3 Domain Driven Data Mining Discovering In-Depth Intelligence

Data mining is used as one of major components and means of BI. One of the problems is that a gap between academia and business emerges in the real-world data mining. The performance and actionable capability of mining data intelligence hasnt been satisfactory as expected. To this end, we propose domain driven data mining for discovering in-depth intelligence [15,16]. The idea is to involve domain intelligence into knowledge discovery to complement existing data-driven data mining framework towards domain driven actionable knowledge discovery. Correspondingly, a series of major work needs to be studied to support domain driven data mining. For instance, the following lists some of such tasks.

- Definition of domain intelligence
- Representation of domain knowledge
- Ontological and semantic representation of domain intelligence
- Domain intelligence transformation between business and data mining
- Human role, modeling and interaction in domain driven mining
- Theoretical problems in involving domain intelligence into KDD
- Metasynthesis of domain intelligence in knowledge discovery
- Human-cooperated data mining
- Constraint-based data mining
- Privacy, security in data mining
- Open environment in data mining
- In-depth data mining
- Knowledge actionability
- Objective and subjective interestingness
- Gap resolution between statistical significance and business expectation
- Domain-oriented knowledge discovery process model
- Profit mining and cost-benefit mining.

5.4 Mining Network Intelligence Supporting Business Decision Making

Additional useful information and knowledge from the Web may greatly supplement existing BI and BI-based operational analysis and decision making. In this case, the key is to develop web intelligence that can be synthesized and integrated into BIS. To this end, the following services may be studied.

- Retrieving information that can supplement the reporting and the data mining findings in BI, for instance, developing information searching engines, content-based searching, etc.
- Analyzing target or opponents web content or web usage to support BI analysis for instance opponent analysis, for instance, analyzing opponents accounting information and listing announcements to investigate opponents production and sales performance.
- Developing web agents to implement information searching, task dispatching, message passing, content and service recommendation, and so on, for instance, designing personal mobile agents to classify and circulate decision-support reports.
- Studying and utilizing social network intelligence to enhance domain intelligence and human intelligence in BI development. For instance, analyzing links in specific social networks may recommend suitable people or knowledge to be presented in the seminar-oriented problem-solving.

5.5 Agent Service Based MBIS Infrastructure

Multi-agent systems and agent-based computing have good potential to model, analyze, design and implement complex intelligent systems, either automated or interactive. On the other hand, service-oriented computing can complement multi-agent systems in integrating distributed enterprise applications, constructing network-based system architecture, etc. Therefore, the integration of agent-based computing with service-oriented computing may offer greater advantages to complex intelligent systems. This way generates an interesting research task, namely agent service-based computing [11]. Agent service-based computing can be very promising and effective for developing MBIS. This is due to the following factors:

- Agent-based computing handling human-machine interaction, developing automated and flexible system modules, implementing self-organization and intelligence interaction, searching web-based information, analyzing opponents web access, etc;
- Service-oriented computing tackling enterprise application integration, distributed system infrastructure, business logic processes, workflow management, as well as knowledge management and services.

The following tasks may be studied in developing agent service-based MBIS.

- Designing agent service based computing model to combine agents and services;
- Information and knowledge sharing in agent service based computing;
- Representing ontologies in agent services;
- Constructing agent service-based system architecture;
- Studying strategies for enterprise application integration;
- Implementing agent services for user modeling, user-computer interaction, run-time interaction;
- Implementing agent services for external information searching, web knowledge management;
- Developing coordination mechanisms for agent services;
- Conducting mechanisms supporting agent service communications and transport;
- Implementing agent service-based decision support services;
- Implementing agent services for generating and managing decision-making reports, and so on.

The evaluation of performance of the next-generation BI systems will be subject to many aspects. However, the basic framework is to check both functional and non-functional requirements, and both technical and business significance from objective and subjective perspectives. More information is available from the reference [15,16].

6 Conclusions

Business intelligence techniques have been widely used for operational analysis and smart decision making. Unfortunately, real-world deployment of existing BI techniques has disclosed many issues that cannot be handled well. With the development of intelligence sciences, in particular, the newly emerging studies on human intelligence, domain intelligence and web intelligence, business intelligence development may greatly benefit from these areas. On the other hand, it is the real-world challenges that determine the advantageous roles of involving and integrating data intelligence, domain intelligence, web intelligence and human intelligence into BI system development. This motivates the research of intelligence metasyntesis in building up next-generation business intelligence systems.

In this paper, we have proposed a framework for BI-oriented intelligence metasyntesis, which follows the practical and powerful theory — Metasyntesis from Qualitative to Quantitative, namely Metasyntetic Engineering. We first summarize the main ideas of metasyntetic engineering theory. Further, a framework is proposed for BI-oriented intelligence metasyntesis. We also discuss the implementation issues in building up metasyntetic business intelligence systems.

Intelligence metasynthesis in constructing business intelligence systems has potential in dealing with complexities and issues in real-world business intelligence development. Its strength comes from the complementary roles of in-depth data intelligence, domain intelligence and web intelligence. It also benefits from human intelligence, and in particular, human-centered problem-solving in tackling open complex problems with human involvement.

Our current work is on developing individual components following the meta-synthetic framework. Further work will be on studying and integrating human intelligence into machine intelligence from both technical and cognitive perspectives.

References

1. Brachman, R.J., Levesque, H.J. (eds.): Readings in Knowledge Representation. Morgan Kaufmann, San Francisco
2. Cao, L.B., Dai, R.W.: Architecture of Internet-based hall for workshop of meta-synthetic engineering. *Computer Science (in Chinese)* 29(6), 63–66 (2002)
3. Cao, L.B.: Studies on some problems in multi-agents-based open giant intelligent systems. PhD thesis, Chinese Academy of Sciences (2002)
4. Dai, R., Cao, L.: Research of Hall for Workshop of Metasynthetic Engineering. *Journal of Management Sciences, China* 5(3), 10–16 (2002)
5. Cao, L.B., Dai, R.W.: Software Architecture of the Hall for Workshop of Metasynthetic Engineering. *Journal of Software* 13(8), 1430–1435 (2002)
6. Cao, L.B., Dai, R.W.: Human-Computer Cooperated Intelligent Information System Based on Multi-Agents. *ACTA AUTOMATICA* 29(1), 86–94 (2003)
7. Cao, L.B., et al.: Systematic engineering in designing architecture of telecommunications business intelligence system. In: *Proceedings of HIS 2003*, pp. 1084–1093. IOS press, Amsterdam (2003)
8. Cao, L.B., Dai, R.W.: Agent-Oriented Metasynthetic Engineering for Decision Making. *Int. J. of Information Technology and Decision Making* 2(2), 197–215 (2003)
9. Dai, R., Cao, L.: Internet—An Open Complex Giant System, *Science in China (Series E)*. *Sciences In China Series E* 33(4), 289–296 (2003)
10. Cao, L.B., et al.: Integration of Business Intelligence Based on Three-Level Ontology Services. In: *Proceedings of WI 2004*, pp. 17–23. IEEE Computer Society Press, Los Alamitos (2004)
11. Cao, L.B., Zhang, C.Q., Ni, J.R.: Agent Services-Oriented Architectural Design of Open Complex Agent Systems. In: *IAT 2005* (2005)
12. Cao, L.B., Zhang, C.Q., Dai, R.W.: The OSOAD Methodology for Open Complex Agent Systems. *Int. J. on Intelligent Control and Systems* (2005)
13. Cao, L.B., Zhang, C.Q., Dai, R.W.: Organization-Oriented Analysis of Open Complex Agent Systems. *Int. J. on Intelligent Control and Systems* 10(2), 114–122 (2005)
14. Cao, L.B., Zhang, C.Q., Liu, J.: Ontology-Based Integration of Business Intelligence. *Int. J. on Web Intelligence and Agent Systems* 4(4), 1–14 (2006)
15. Cao, L.B., Zhang, C.Q.: Domain-Driven Data Mining, a Practical Methodology. *International Journal of Data Warehousing and Mining* 2(4), 49–65 (2006)
16. Cao, L.B., Zhang, C.Q.: The evolution of KDD: Towards domain-driven data mining. *International Journal of Pattern Recognition and Artificial Intelligence* (2007)

17. Chalupsky, H.: OntoMorph: A translation system for symbolic logic. In: KR 2000: Principles of Knowledge Representation and Reasoning, pp. 471–482. Morgan Kaufmann, San Francisco (2000)
18. Clark, A.: *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford University Press, Oxford (2000)
19. Clark, A.: *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, Oxford (2003)
20. Dai, R.W., Wang, J.: Research on giant intelligent systems 19(6), 645–655 (1993)
21. Dai, R.W., Wang, J., Tian, J.: *Metasynthesis of Intelligent Systems* (in Chinese), Zhejiang Science and Technology Publishing House (1995)
22. Fayyad, U.M.: Tutorial report. Summer school of DM. Monash Uni, Australia (July 2003)
23. Fensel, D.: *Ontologies: a silver bullet for knowledge management and electronic commerce*, 2nd edn. Springer, Heidelberg (1998)
24. Gomez-perez, A., et al.: *Ontological engineering*. Springer, Heidelberg (2004)
25. Han, J.: Towards Human-Centered, Constraint-Based, Multi-Dimensional Data Mining. An invited talk at Univ. Minnesota, Minneapolis, Minnesota (1999)
26. Han, J.W., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco (2006)
27. Inmon, W.H.: *Building the data warehouse*, 3rd edn. Wiley, Chichester (2002)
28. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18(1), 1–31 (2003)
29. Lin, L., Cao, L.: Mining In-Depth Patterns in Stock Market. *Int. J. Intelligent System Technologies and Applications* (to appear, 2007)
30. Qian, X.S.: Modern science and technology structure — restudies on the system of sciences and technologies, *Chinese J. of Zhexue Yanjiu* (3) (1982)
31. Qian, X.S., Yu, J.Y., Dai, R.W.: A new discipline of science — the study of open complex giant system and its methodology. *Chinese J. of Nature* 13(1), 3–10 (1990)
32. Qian, X.S.: Restudies on open complex giant systems. *Chinese J. of Pattern Recognition and Artificial Intelligence* 4(1), 5–8 (1991)
33. Wang, S.Y., Dai, R.W., et al.: *Open complex giant systems*, Zhejiang science and technology publishing house (1996)
34. Storey, V.C.: Understanding semantic relationships. *The very large data bases Journal* 2(4), 455–488 (1993)
35. Wooldridge, M., Jennings, N.: *Intelligent agents: theory and practice*. *Knowledge Engineering Review* 10(2), 115–152 (1995)
36. Zhong, N., Liu, J.M., Yao, y.y.: *Web Intelligence*. Springer, Heidelberg (2003)
37. China Mobile, China mobile operational analysis system technical specification (Chinese) (2002)

Risk Mining in Medicine: Application of Data Mining to Medical Risk Management

Shusaku Tsumoto¹, Yuko Tsumoto², Kimiko Matsuoka²,
and Shigeki Yokoyama³

¹ Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo 693-8501 Japan
tsumoto@computer.org

² Department of Fundamental Nursing, Shimane University, School of Nursing
89-1 Enya-cho, Izumo 693-8501 Japan

³ Osaka Prefectural General Hospital, Osaka, Japan

⁴ Department of Medical Information, Kodan Industry, Tokyo, Japan

Abstract. Organizations in our modern society grow larger and more complex to provide advanced services due to the varieties of social demands. Such organizations are highly efficient for routine work processes but known to be not robust to unexpected situations. According to this observation, the importance of the organizational risk management has been noticed in recent years. On the other hand, a large amount of data on the work processes has been automatically stored since information technology was introduced to the organizations. Thus, it has been expected that reuse of collected data should contribute to risk management for large-scale organizations. This paper proposes risk mining, where data mining techniques were applied to detection and analysis of risks potentially existing in the organizations and to usage of risk information for better organizational management. We applied this technique to the following three medical domains: risk aversion of nurse incidents, infection control and hospital management. The results show that data mining methods were effective to detection of risk factors.

1 Introduction

It has passed about twenty years since clinical information are stored electronically as a hospital information system since 1980's. Stored data include from accounting information to laboratory data and even patient records are now started to be accumulated: in other words, a hospital cannot function without the information system, where almost all the pieces of medical information are stored as multimedia databases. Especially, if the implementation of electronic patient records is progressed into the improvement on the efficiency of information retrieval, it may not be a dream for each patient to benefit from the personal database with all the healthcare information, "from cradle to tomb". However, although the studies on electronic patient record has been progressed rapidly, reuse of the stored data has not yet been discussed in details, except

for laboratory data and accounting information to which OLAP methodologies are applied. Even in these databases, more intelligent techniques for reuse of the data, such as data mining and classical statistical methods has just started to be applied from 1990's [1,2].

Human data analysis is characterized by a deep and short-range investigation based on their experienced "cases", whereas one of the most distinguished features of computer-based data analysis is to enable us to understand from the different viewpoints by using "cross-sectional" search. It is expected that the intelligent reuse of data in the hospital information system provides us to grasp the all the characteristics of university hospital and to acquire objective knowledge about how the hospital management should be and what kind of medical care should be served in the university hospital.

This paper focuses on application of data mining to medical risk management. To err is human. However, medical practice should avoid as many errors as possible to achieve safe medicine. Thus, it is a very critical issue in clinical environment how we can avoid the near misses and achieve the medical safety. Errors can be classified into the following three type of errors. First one is systematic errors, which occur due to problems of system and workflow. Second one is personal errors, which occur due to lack of expertise of medical staff. Finally, the third one is random error. The important point is to detect systematic errors and personal errors, which may be prevented by suitable actions, and data mining is expected as a tool for analysis of those errors.

For this purpose, this paper proposes *risk mining* where data including risk information is analyzed by using data mining methods and mining results are used for risk prevention. We assume that risk mining consists of three major processes: risk detection, risk clarification and risk utilization, as shown in Section 2.

We applied this technique to the following three medical domains: risk aversion of nurse incidents, infection control and hospital management. The results show that data mining methods were effective to detection of risk factors.

This paper is organized as follows. Section 2 shows background of our studies. Section 3 proposes three major processes of risk mining. Sections 4 to 6 give illustrative applications of risk mining. Finally, Section 7 concludes this paper.

2 Background

A hospital is a very complicated organization where medical staff, including doctors and nurses give a very efficient and specialized service for patients. However, such a complicated organization is not robust to rapid changes. Due to rapid advances in medical technology, such as introduction of complicated chemotherapy, medical workflow has to be changed in a rapid and systematic way. Such rapid changes lead to malpractice of medical staff, sometimes a large-scale accident may occur by chain reaction of small-scale accidents.

Medical accidents include not only careless mistakes of doctors or nurses, but also prescription errors, intrahospital infections or drug side-effects. The

cause for such accidents may not be well investigated and it is unknown whether such accidents can be classified into systematic errors or random errors. Since the occurrence of severe accidents is very low, case studies are used for their analysis. However, in such investigations, personal errors tend to be the cause of the accidents. Thus, it is very important to discover knowledge about how such accidents occur in a complicated organization and knowledge about the nature of systematic errors or random errors.

On the other hand, clinical information have been stored electronically as a hospital information system(HIS). The database stores all the data related with medical actions, including accounting information, laboratory examination, treatment and patient records described by medical staffs. Incident or accident reports are not exception: they are also stored in HIS as clinical data. Thus, it is now expected that mining such combined data will give a new insight to medical accidents.

3 Risk Mining

In order to utilize information about risk extracted from information systems, we propose risk mining which integrates the following three important process: risk detection, risk clarification and risk utilization.

3.1 Risk Detection

Patterns or information unexpected to domain experts may be important to detect the possibility of large scale accidents. So, first, mining patterns or other types of information which are unexpected to domain experts is one of the important processes in risk mining. We call this process *risk detection*, where acquired knowledge is referred to as *detected risk information*.

3.2 Risk Clarification

Focusing on detected risk information, domain experts and data miners can focus on clarification of modelling the hidden mechanism of risk. If domain experts need more information with finer granularity, we should collect more data with detailed information, and apply data mining to newly collected data. We call this process *risk clarification*, where acquired knowledge is referred to as *clarified risk information*.

3.3 Risk Utilization

We have to evaluate clarified risk information in a real world environment to prevent risk events. If risk information is not enough to prevention, then more analysis is required. Thus, additional data collection is evoked for a new cycle of risk mining process. We call this process *risk utilization*. where acquired knowledge is referred to as *clarified risk information*.

Figure 1 shows the overview of risk mining process.

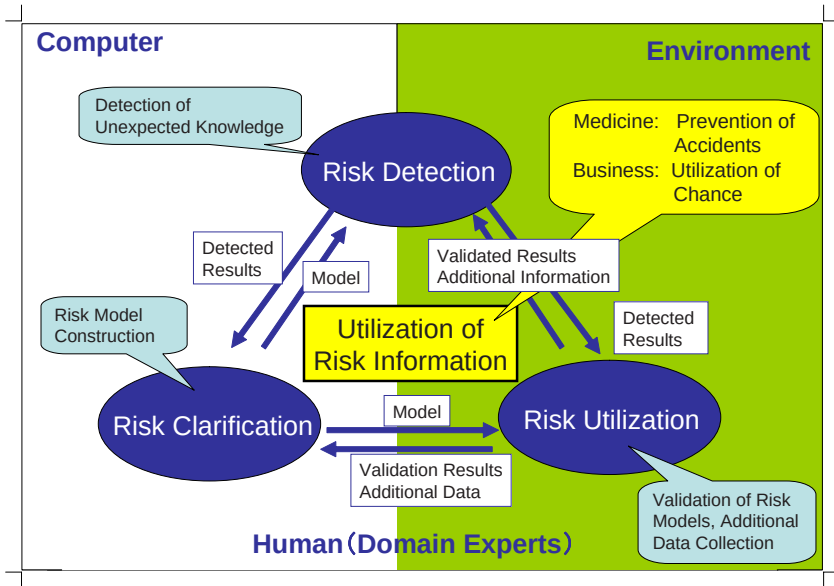


Fig. 1. Risk Mining Proces: Overview

3.4 Elemental Techniques for Risk Mining

Mining Unbalanced Data. A large scale accident rarely occur: usually such it can viewed as a large deviation of small scale accidents, called incidents. Since even the occurence of incidents is very low, the probability of large accidents is nearly equal to 0. On the other hand, most of the data mining methods depend on “frequency” and mining such unbalanced data with small probabilities is one of the difficult problems in data mining research. Thus, for risk mining, techiques for mining unbanced data are very important to detect risk information.

3.5 Interestingness

In convetional data mining, indices for mining patterns are based on frequency. However, to extract unexpected or interesting knowledge, we can introduce measures for unexpectedness or interestingness to extract patterns from data, and such studies have been reported in data mining literature.

3.6 Uncertainty and Granularity: Granular Computing

Since incident reports include information about human actions, these data are described by subjective information with uncertainty, where we need to deal with coarseness and fineness of information (information granularity). Granular computing, including fuzzy sets and rough sets, are closely related with this point.

3.7 Visualization

Visualizing cooccurrence events or items may enable domain experts to detect risk information, to clarify the mechanism of risk, or to utilize risk information.

3.8 Structure Discovery: Graph Mining

Risk may be detected or clarified only by relations between several items in a large network structure. Thus, extracting partial structure from network hidden in data is a very important technique, focusing on risk information based on relations between items.

3.9 Clustering

Similarity may find relations between similar objects which seems not to be similar. Or events which seems to occur independently can be grouped into several “similar” events, which enables us to find dependencies between events. For this purpose, clustering is a very important technique.

3.10 Evaluation of Risk Probability

Since probability is formally defined as a Lebesgue measure on a fixed sample space, its performance is very unstable when the definition of sample space is unstable. Especially, when we collect data dynamically, such instability frequently occurs. Thus, deep reflection on evaluation of risk probability is very important.

3.11 Human Computer Interaction

This process is very important for risk mining process because of the following reasons. First, risk information may be obtained by deep discussions on mining results among domain experts because mining results may show only small part of the total risk information. Since domain experts have knowledge, which is not described in a datasets, they can compensate for insufficient knowledge to obtain a hypothesis or explanation of mining results. Second, mining results may lead to domain experts’ deep understanding of workflow, as shown in Section 4. Interpretation of mining results in risk detection may lead to new data collection for risk clarification. Finally, human computer interaction gives a new aspect for risk utilization. Domain experts can not only performance of risk clarification results, but also look for other possibilities from the rules which seems to be not so important, compared with rules for risk clarification and also evaluate the possibility to design a new data collection.

4 Application I: Prevention of Medication Errors

As an illustrative example, we applied risk mining process to analysis of nurses’ incident data. First, data collected in 6 months were analyzed by rule induction

methods, which detects several important factors for incidents (risk detection). Since data do not include precise information about these factors, we recollect incident data for 6 months to collect precise information about incidents. Then, rule induction is applied to new data. Domain experts discussed all the results obtained and found several important systematic errors in workflow (risk clarification). Finally, nurses changed workflow to prevent incidents and data were recollected for 6 months. Surprisingly, the frequency of medication errors has been reduced to one-tenth (risk utilization).

4.1 Background

A hospital is a very complicated organization where medical staff, including doctors and nurses give a very efficient and specialized service for patients. However, such a complicated organization is not robust to rapid changes. Due to rapid advances in medical technology, such as introduction of complicated chemotherapy, medical workflow has to be changed in a rapid and systematic way. Such rapid changes lead to malpractice of medical staff, sometimes a large-scale accident may occur by chain reaction of small-scale accidents.

Medical accidents include not only careless mistakes of doctors or nurses, but also prescription errors, intrahospital infections or drug side-effects. The cause for such accidents may not be well investigated and it is unknown whether such accidents can be classified into systematic errors or random errors. Since the occurrence of severe accidents is very low, case studies are used for their analysis. However, in such investigations, personal errors tend to be the cause of the accidents. Thus, it is very important to discover knowledge about how such accidents occur in a complicated organization and knowledge about the nature of systematic errors or random errors.

On the other hand, clinical information have been stored electronically as a hospital information system(HIS). The database stores all the data related with medical actions, including accounting information, laboratory examination, treatment and patient records described by medical staffs. Incident or accident reports are not exception: they are also stored in HIS as clinical data. Thus, it is now expected that mining such combined data will give a new insight to medical accidents.

4.2 Risk Detection

Dataset. Nurses' incident data were collected by using the conventional sheet of incident reports during 6 months from April, 2001 to September, 2001 at the emergency room in Osaka Prefectural General Hospital.

The dataset includes the types of the near misses, the patients' factors, the medical staff's factors and the shift (early-night, late-night, and daytime) and the number of items of incidents collected was 245.

We applied C4.5 [3], decision tree induction and rule induction to this dataset.

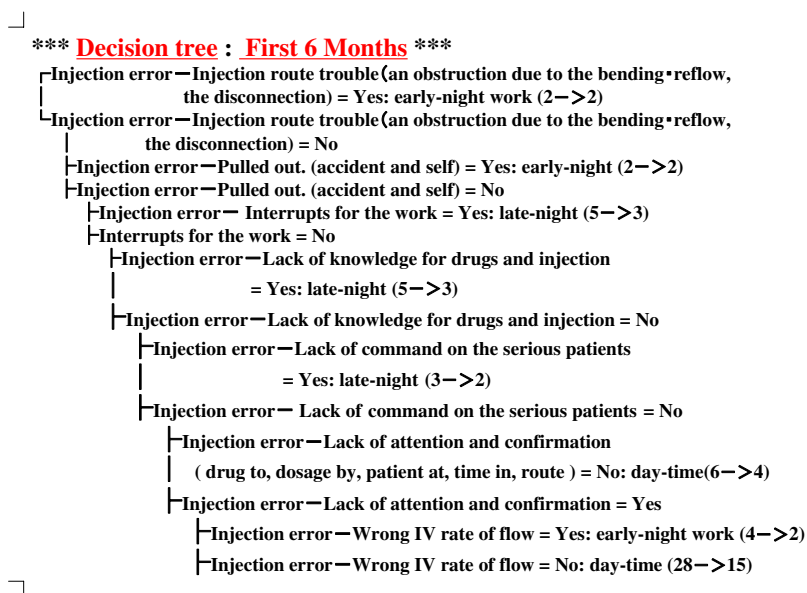


Fig. 2. Decision Tree in Risk Detection

Rule Induction. We obtained a decision tree shown in Fig. 2 and the following interesting rules.

(medication error):

If late-night and lack of checking,
then medication errors occur: probability (53.3%, 8/15).

(injection error):

If daytime and lack of checking,
then injection incidents occur: probability (53.6%, 15/28).

(injection error):

If early-night, lack of checking, and error of injection rate,
then injection incidents occur: probability (50%, 2/4)

Those rules show that the time shift of nurse and lack of checking were the principal factors for medication and injection errors. Interestingly, lack of expertise (personal errors) was not selected. Thus, time shift and lack of checking could be viewed as risk factor for these errors. Since the conventional format of incident reports did not include future information about workflow, we had decided to ask nurses' to fill out new report form for each incident. This is the next step in risk clarification.

Circumstances of incident

Person who noticed the incident

Whether the incident was anticipated or not

Degree of busyness, etc.

Environment of incident

Number of patients in A ward

Number of patients in B ward

Number of patients isolated due to infections

Number of restless patients

Whether there were new arrivals of patients or not

Whether treatment was made or not, etc.

Additional check sheet

ニアス病室には職員が2名（看護士）と医師が1名（医師）といます。看護士の人数は1名です。この表に記入してください。

4口病室には看護士の2名がいます。

ニアス病室の状況

1 飲食・排泄・睡眠・その他は問題ありませんか？ ☐ はい ☐ いいえ ☐ その他

2 ニアスに異変はありましたか？ ☐ はい ☐ いいえ

3 ニアスを予備で止めたか？ ☐ はい ☐ いいえ

4 ニアスを起こす前に観察に気づきましたか？ ☐ はい ☐ いいえ

5 服用・投与の状況はどうか？ ☐ はい ☐ いいえ

6 ニアスが不安定な状態にあるか？ ☐ はい ☐ いいえ

7 ニアスは安定しているか？ ☐ はい ☐ いいえ

8 ニアスに異変はありましたか？ ☐ はい ☐ いいえ

9 気管挿入・気管吸引・気管挿入後 ☐ はい ☐ いいえ

10 呼吸器・呼吸器・呼吸器 ☐ はい ☐ いいえ

11 大量輸液・大量輸液・大量輸液 ☐ はい ☐ いいえ

12 不眠（夜間）を繰り返しているか？ ☐ はい ☐ いいえ

13 外来患者の入室 ☐ はい ☐ いいえ

14 外来患者の入室 ☐ はい ☐ いいえ

15 外来患者の入室 ☐ はい ☐ いいえ

16 外来患者の入室 ☐ はい ☐ いいえ

17 外来患者の入室 ☐ はい ☐ いいえ

18 外来患者の入室 ☐ はい ☐ いいえ

19 外来患者の入室 ☐ はい ☐ いいえ

20 外来患者の入室 ☐ はい ☐ いいえ

21 ニアス前に治療変更がございましたか？ ☐ はい ☐ いいえ

22 はい・いいえ・その他は変更前を継続できる状況でしたか？ ☐ はい ☐ いいえ

Fig. 3. Sheet for Additional Information

4.3 Risk Clarification

Dataset. Just after the first 6 months, we had found that the mental concentration of nurses may be important factors for medical errors. During the next 6 months from October 2001 to March 2002, the detailed interference factors were included in the additional incident report form as the items of "environmental factors".

Figure 3 shows a sheet for additional information. The additional items included the duration of experience at the present ward, the number of nurse, the degree of business, the number of serious patients whether the nursing service was interrupted or not and so on.

We applied C4.5 [3], decision tree induction and rule induction to this dataset.

Rule Induction. The following rules were obtained:

(medication error):

If the number of disturbing patients is one or more,
then medication errors occur: probability (90%, 18/20).

(medication error):

If nurses' work interrupted,
then medication errors occur: probability (80%, 4/5).

By addition of "the environmental factors", these high probability rules of medication errors were extracted.

Rule Interpretation. With these results, the nurses discussed their medication check system.

At the emergency room, the nurses in charge of the shift prepared the medication (identification, quantity of medicines, etc.). The time of preparation before the beginning of the shift was occasionally less than 30 minutes when the liaison conference between shifts took time. In such cases, the sorting of medicines could not be made in advance and must be done during the shift.

If nurses' concentration was disturbed by the restless patients in such situations, double check of the preparation for medicine could not be made, which leads to medication errors.

4.4 Risk Utilization

Therefore, it was decided that two nurses who had finished their shifts would prepare medicines for the next shift, and one nurse in charge of the medication would check the dose and identification of medicines alone (triple check by a total of 3 nurses). (However, heated discussions among domain experts (nurses) needed for this decision, as shown in Section 4.5.) Improvement was applied to the check system as a result of their discussion. During the last 6 months (April 2002 to October 2002), incident reports were collected.

After introducing the triple check system, the total number of the medication errors during the last 6 months decreased to 24 cases. It was considered that the nurses' medication work was improved by the triple check system during the last 6 months.

4.5 Discussion for Case Study

Risk Utilization as Information Sharing. For discussion among domain experts, mining results were presented to medical staffs as objective evidence. Discussion on mining results give a very interactive discussion among the staff of the department of emergency and finally achieve common understanding of the problem on its workflow. Then, it is found that changes in workflow is required for solving the problem: If the staff assigned to the shift cannot prepare medicines, other members who are free should cooperate. However, this idea met a fierce objection in the department at first because of disagreement among nurses about the responsibility of those who prepare medicines. After repeated discussions, it was decided that nurses in charge of medication were responsible for mistakes rather than those who made preparations and nurses in the preceding shift should prepare medicines for the next shift.

During the last 6 months, medication errors were reduced markedly by creating the common perception that liaison (overlapping of shift margins, or paste margins) is important among nurses, and the initial opposition completely subsided. Following this nursing example, we could extend this policy of "paste margins", i.e. mutual support by free staff members, to the entire department.

This process also shows that information granularity is a very important issue for risk clarification.

Items in a conventional report form, such as “lack of checking, lack of attention, etc.” are too coarse for risk clarification. Rather, detailed description of environmental factors are much more important to evoke domain experts’ discussion and their risk utilizaiton.

5 Application II: Infection Control (Risk Detection)

5.1 Background

Intrahospital infection is one of the serious medical accidents where microorganisms are transmitted by doctors or nurses and a patient admitted to the hospital without infection suffer from infectious diseases. One of the most important bacteria is MRSA (methicillin-resistant Staphylococcus aureus), which is resistant to almost all the antibiotics and may cause the immunosuppressed patients to death. Thus, it is very important to extract risk factors of intrahospital infection from data as evidence.

5.2 Methods

Objective. For prevention of MRSA infection, we extract background risk factors of MRSA detection from a clinical database, extracted from a hospital information system, including laboratory examination results and microbial examination data.

Material. The population for this study consisted of 236 patients, comprising 118 patients with MRSA and 118 patients with microorganisms other than MRSA who had been treated between 1995 and 1998. Table 1 shows all the components of this dataset.

Table 1. Attributes in a Dataset on Infectious Diseases

Item	Attributes (63)
Patient’s Profile	ID, Gender, Age
Department	Department, Ward, Diagnosis(3)
Order	Background Diseases, Sampling Date, Sample, No.
Symptom	Fever, Cathether(5), Traheotomy, Endotracheal intubation, Drainage(5)
Examination Data	CRP, WBC, Urin data, Liver/Kidney Function, Immunology
Therapy	Antibiotic agents(3), Steroid, Anti-cancer drug, Radiation Therapy, Lactobacillus Therapy
Culture	Colony count, Bacteria, Vitek biocode, β -lactamase
Susceptibility	Cephems, Penicillins, Aminoglycoside, Macrolides, Carbapenums, Chloramphenicol,Rifanpic, VCM, etc.

Analytic Method. For analysis, we used decision tree and rule induction based on C4.5 [3] and chi-square test. Decision tree was used to capture the global structure of the important features. "If-then" rules were extracted from the decision trees. Finally, chi-square test was applied in order to rank the rules obtained.

5.3 Results

Rules Obtained. The following significant "If-then" rules were extracted from a decision tree for MRSA detection and the various catheters.

Rule-1:

If (transfusion catheter=No) and (catheter=Yes) and (CVP=No),
then (MRSA=detected). Probability was 0.88 (n=7 / 8).

Rule-2:

If (transfusion catheter=No) and (urinary tract catheter=No)
and (intra-arterial catheter=Yes), then (MRSA=detected).
Probability was 1.00 (n=3 / 3).

From these rules, there may be a strong relationship between various catheters and detection of MRSA.

Chi-square Test. Chi-square test was applied to evaluate an association between MRSA detection and various catheter by using a contingency table shown in Table 2. The p -value was 0.00063 and dependence between catheter and MRA is very strong.

Table 2. Contingency table between MRSA detection and Catheters

	Catheter	Non Catheter	Total
MRSA	80	38	118
Others	54	64	118

5.4 Decision Tree

The following decision tree was obtained as a relationship between MRSA detection and medication of the antimicrobial agents, as shown in Fig. 4.

Decision tree showing an association between MRSA detection and Antibiotics in Fig. 4, cephem antibiotic agents were not extracted by the decision tree. In general, there might be a strong relationship between MRSA detection and cephem medication. However, this decision tree showed that Cephem medication has a weak relationship with MRSA detection: there might not be an association between MRSA detection and Cephem medication.

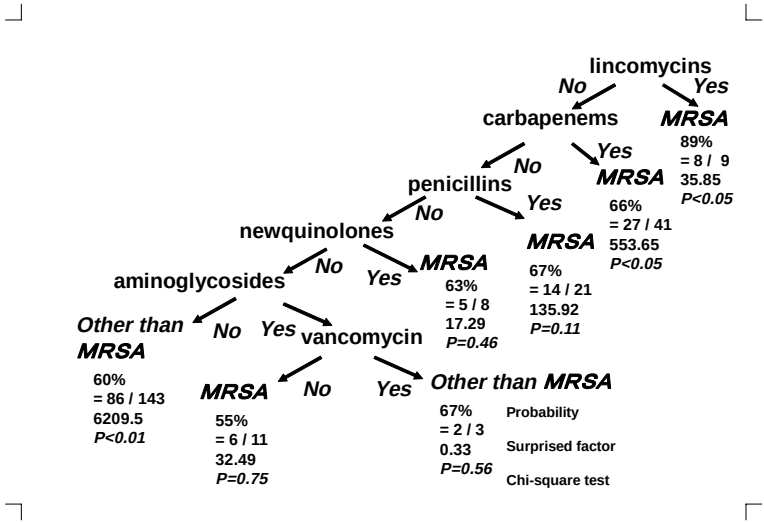


Fig. 4. Decision tree showing an association between MRSA detection and Antibiotics

5.5 Chi-square Test

Chi-square test was applied to evaluate the association between Cephem antimicrobial agent and MRSA detection, whose contingency table is shown in Table 3. The p value was 0.64 $\geq$ 0.05 and was statistically significant. Thus, this result showed that cephem medication has a weak relationship with MRSA detection.

Table 3. Contingency table between MRSA detection and Cephem medication

	Catheter	Non Catheter	Total
MRSA	25	93	118
Others	28	90	118

5.6 Rules from Decision Tree

In the decision tree(Fig. 4) , the first node is medication of Lincomycin. Here, Lincomycin medication means Lincosamides antimicrobial agent medication which includes Clindamycin medication. This decision tree begins to branch from the factor of the strongest relationship with MRSA detection. As the factor of the first node is Lincomycin medication, the following rule was obtained:

Rule-3:
If Lincomycin medication = YES, then MRSA detection = YES.
(probability: 89%= 8/9).

This rule indicates that Lincomycin medication is the least effective for MRSA thera-py. Furthermore, the decision tree shows that Carbapenem, Penicillin, and New quino-lon medications are also ineffective for MRSA therapy.

5.7 Discussion

Our study has been conducted to extract background risk factors of MRSA infection, by using data mining techniques. The decision trees and if-then rules were extracted, and chi-square tests were also applied to evaluate the association with MRSA detection. In Japan, it is generally said that there are strong relations between the MRSA detection and the cephem medication. But, in our hospital, we feel empirically that our cephem medication is not closely related with MRSA. This empirical knowledge was supported by our mining results: cephem medication was not extracted from the decision tree shown in Fig. 4. Furthermore, chi-square test shows that the dependency between cephem medication and MRSA infection was not statistically significant, as shown in Table 3. Thus, there may not be an association between MRSA detection and Cephem medication.

As shown in the above section, rule 3 extracted from Fig. 1 may show that Lincomycin medication induce MRSA infection. (Here, as the above mentioned, Lincomycin medication includes Clindamycin.) Because, according to NCCLS-2004 [4], Lincosamides Clindamycin: Macrolide-resistant isolates of *S. aureus* and coagulase-negative *Staphylococcus* spp. [methylation of the 23S rRNA encoded by the *erm* gene also referred to as MLSB (macrolide, lincosomide and type B streptogramin) resistance] may have constitutive or inducible resistance to clindamycin or may be resistant only to macrolides (efflux mechanism encoded by the *msrA* gene).

Furthermore, from the analysis of Schreckenberger, et. al. [5], the incidences of inducible clindamycin resistance at two hospital (an inner-city hospital and suburban community hospital) were 7 and 12% for methicillin-resistant *Staphylococcus aureus* (MRSA), 20 and 19 % for methicillin -susceptible *S. aureus* (MRSE), and 14 and 35% for coagulase-negative staphylococci, respectively. From their analysis, the variability of inducible resistance to clindamycin was found in those two hospitals. They concluded that susceptibility testing of staphylococci should include the disk diffusion induction test (D-test).

Therefore, rule-3 might indicate the above variability of inducible MRSA-resistance to Lincomycin. It will be important for the variability of inducible MRSA-resistance to Lincomycine in the hospital infection control. Thus, susceptibility testing of staphylococci should include the test of inducible MRSA-resistance to Lincomycin.

6 Application III: Risk Detection and Clarification in Hospital Administration

6.1 Objective

The objectives of this research is to investigate what kind of knowledge can be extracted by statistical methods from the datasets stored in the hospital information system of Chiba University Hospital, especially useful for future hospital management and decision support. Especially, since the revenue of Japanese

hospital is based on NHI points of Japanese medical care, it is important to investigate the factor which determines the amount of NHI points.

6.2 Methods

Representation of Discharge Summaries. When the hospital information system for discharge summaries is introduced in Chiba University Hospital in 1978, a discharge summary is distributed to doctors as a paper sheet for each patient admitted to the hospital. Doctors fill in each sheet just after the patient leaves the hospital, the parts of this sheet which can be coded are stored electronically. A sheet for discharge summary is composed of the items common to all the departments and the items specific to each department. For example, the items specific to neurology consists of the results of neurological examinations and the severity of diseases. The common items consist of those in which codes or numerical values should be filled in and those in which texts should be input. After the doctor in charge fill in those items and submit to the division of medical records, the staff input codes and numerical values into a database system. These processes are continued until a new hospital information system was introduced in 2000, which means that the non-text items common to all the departments has been stored for about 20 years. There are 16 items for codes or numerical values: patient ID, the department in charge, occupation, height and weight on admission, height and weight just before hospital discharge, a motivation for visit, outcome, autopsy or not, cause of death, the date of first visit, the date of admission, the date of discharge, the name of disease (ICD-9 code [6]), treatment method. However, height and weight just before hospital discharge are not input in the database. Concerning the items specific to each department, only those of surgery and ophthalmology are stored electronically.

Extraction of Datasets from Hospital Information System. The databases in the hospital information system of Chiba University Hospital are described by MUMPS [7]. MUMPS is a programming language which can construct a database with a hierarchical structure based on a binary tree. By using the characteristics of a binary tree, each item can store several data as a tree, which makes the data management and retrieval more efficient than relational databases. Datasets for analysis are extracted from the database on discharge summaries and the database on patient basic information by using patient ID and the date on admission as keys. The program for extraction is developed by the first author due to the following reasons. Since NHI points, which stands for National Healthcare Insurance points, are stored for each patient ID and each month, the total points for each admission for each patient are calculated from NHI points for each month. The total points are combined with the dataset extracted from the discharge summaries by using patient ID and the date on admission as keys. The number of the records of the dataset extracted from the global: MRMG, which is a database on discharge summaries, is 157,636 for 21 years from 1978.4 to 2000.3. The time needed for computation is about one hour by SUN Workstation (Enterprise 450). Concerning the dataset combined with

NHI points, the number of the records is 20,146 for three years from 1997.4 to 2000.3.

Methods for Statistical Analysis. Descriptive statistics, exploratory data analysis and statistical tests were applied to the dataset extracted only from the discharge summaries for the analysis of patient basic information (gender, age and occupation), outcome, the number of the days in hospitals and diseases, including their chronological trends. Concerning the datasets combined with accounting information for three years (1997.4 to 2000.3), the relations among NHI points and items in the discharge summaries were analyzed by descriptive statistics, exploratory data analysis, statistical tests, regression analysis and generalized linear model. R was used for these analyses..

6.3 Results

Due to the limitation of the spaces, the most interesting results are shown in this section. In the subsequent subsections, the results of the whole cases, and two levels of ICD-9 code, called major and minor divisions, are compared. Especially, concerning the results for the major and minor divisions, malignant neoplasm and the following three largest minor divisions of the malignant neoplasm are focused on: neoplasm of trachea, bronchus, and lung, neoplasm of stomach, and neoplasm of liver and intrahepatic bile ducts. In the subsequent sessions, neoplasm of lung, stomach and liver denotes the above three divisions for short.

Distribution of Length of Stay. Figure 5 shows the distribution of the length of stay of the whole cases, which skewed with the long tail to the right. Thus, the logarithm of the length of stay is taken to tame this skewness. Figure 6 gives the result of this transformation, whose distribution is very close to normal distribution: this means that the distribution of the whole cases is log-normal. This observation holds even when the major divisions are taken as sample. It is notable that the nature of this distribution holds even in the minor divisions, three largest three diseases.

Table 4 summarizes the descriptive statistics of length of stay with respect to the whole cases, major and minor divisions. The natures of these distributions are not significantly changed if these cases are stratified by the condition whether a surgical operation is applied to a case or not.

Distribution of NHI Points. Since the variance of raw data of NHI points are very large and the distribution is skewed, the raw data are transformed into the "median index", which is defined as the ratio of the total points to the median of the whole cases. Figure 7 and 8 show the distribution of the raw data and that of the logarithm of the raw data of median index. Those figures suggests that the NHI points of the whole cases follow log-normal distribution. On the other hand, the distributions for minor divisions are different. The same observations are obtained from the distribution of NHI Points of Neoplasm.

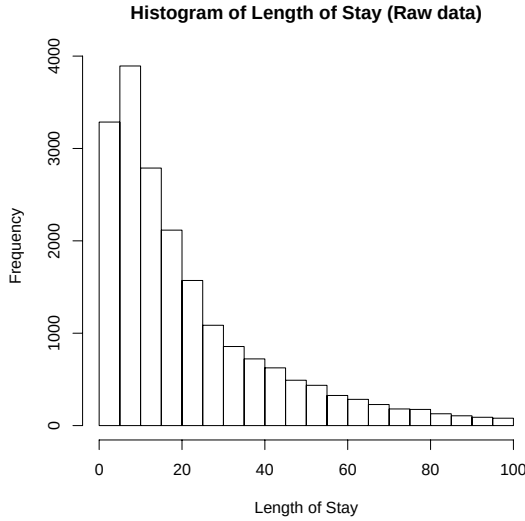


Fig. 5. Distribution of Length of Stay (Raw Data, Total Cases)

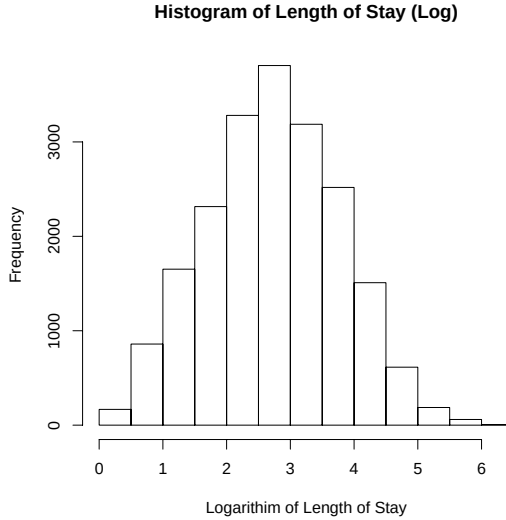


Fig. 6. Distribution of Length of Stay (Logarithm Transformation, Total Cases)

6.4 Correlation Between Length of Stay and NHI Points

Figure 9 and 10 depicts the scattergram between the length of stay and NHI points of total cases and neoplasm cases which suggest high correlation between two variables. For simplicity, the vertical and horizontal axes show the logarithm of raw values. Actually, The coefficient of correlation are calculated as 0.837 and 0.867, which mean that the correlation is very strong.

Table 4. Descriptive Statistics of Length of Stay

	Average	Median	SD	Skewness	Kuritosis
Whole Cases					
Raw Data	26.46	16.00	33.67	4.34	34.15
Logarithmic Transformation	2.74	2.77	1.06	-0.06	-0.28
Neoplasm					
Raw Data	37.54	25.00	38.72	2.90	13.21
Logarithmic Transformation	3.19	3.22	0.98	-0.32	0.08
Malignant Neoplasm of Lung					
Raw Data	49.65	39.00	43.42	2.57	10.82
Logarithmic Transformation	3.57	3.66	0.88	-0.79	2.00
Malignant Neoplasm of Stomach					
Raw Data	36.44	36.00	19.18	0.46	0.37
Logarithmic Transformation	3.40	3.58	0.72	-1.42	2.55
Malignant Neoplasm of Liver					
Raw Data	35.93	33.00	21.40	1.19	2.70
Logarithmic Transformation	3.38	3.50	0.71	-1.18	3.03

Table 5. Correlation between length of stay and NHI points (After Logarithm Transformation)

	Total	With Operation	Without Operation
Total Cases	0.837	0.829	0.779
Neoplasm	0.867	0.844	0.826
Lung Cancer	0.838	0.648	0.903
Stomach Cancer	0.827	0.738	0.801
Liver Cancer	0.711	0.577	0.755

Table 5 summarized the correlation coffecients between NHI points and Length of Stay with respect to the whole cases, neoplasm and three major types of malignant neoplasm: lung, stomach and liver. Comparison of the coefficient of correlation between the group with and without a surgical operation shows that the group without an operation has higher correlations than that with an operation, which suggests that NHI points of the treatment methods other than surgical operations should be strongly dependent on the lengths of stay.

Generalized Linear Model. Since all the items except for the length of stay are categorical variables, conventional regression models cannot be applied to the study on relations between NHI points and other items. For this purpose, generalized linear model [8] was applied to the dataset on combination of accounting data and discharge summaries. NHI point was selected as a target variable and the following four variables were selected as explanatory variables: outcome, treatment method, major division of ICD-9 codes and the categorized length of stay. The length of stay is categorized so that the distribution of the transformed variable is close to normal distribution, where the width of windows is set to 0.5 for the logarithmic value of the length of stay. Treatment, outcome and major divisions

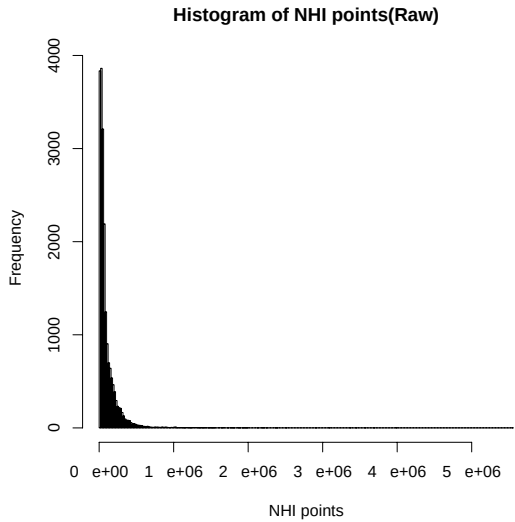


Fig. 7. Distribution of NHI Points (Raw Data, Total Cases)

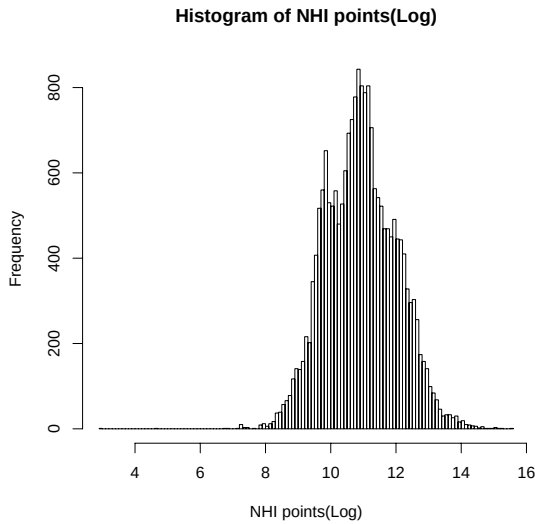


Fig. 8. Distribution of NHI Points (Logarithm Transformation, Total Cases)

of ICD codes are transformed into dummy variables to clarify the contributions of these values to a target variable. For example, the outcomes of discharge are split into the following six dummy variable: D1: recovered, D2: improved, D3: unchanged, D4: worsened, D5: dead and D6: others. Figure 11 shows the results of GLM on the total cases, whose target variable is NHI points. All the variables are sorted by the F value. The most contributing factor is the length of stay, whereas the contributions of the other factors are small.

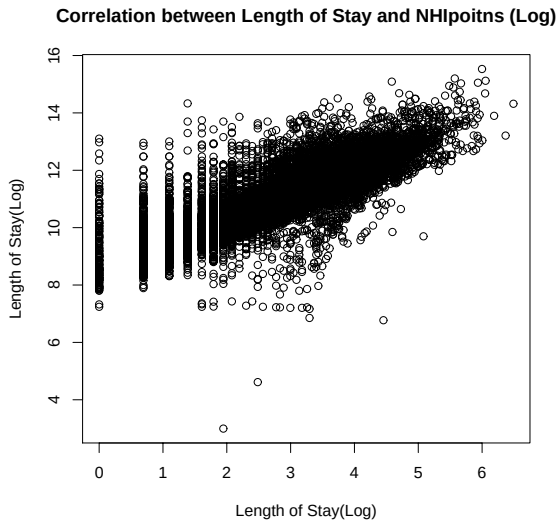


Fig. 9. Scattergram of Length of Stay and NHI Points (Logarithm Transformation, Total Data)

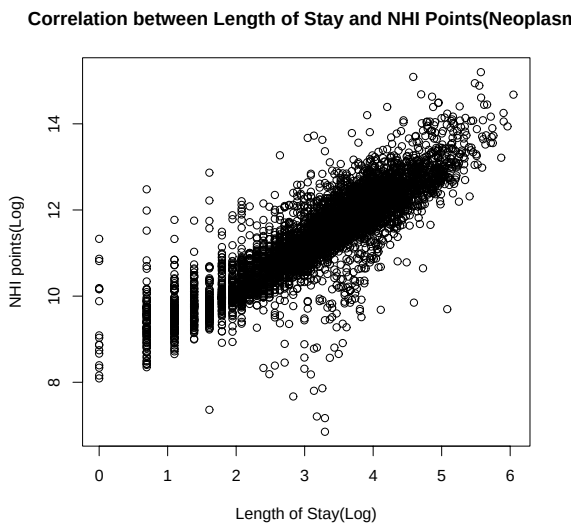


Fig. 10. Scattergram of length of stay and NHI points (Logarithm Transformation, Neoplasm)

Figure 12 gives the results of GLM on major division (malignant neoplasm), and three minor divisions, whose target variable is NHI points. Compared with Fig. 11, the number of the factors which gives the significant contributions to NHI points are very small, which suggest that the variabilities of NHI points in major and minor divisions are very low, compared with that of total cases.

```
Call:
glm(formula
     = lognhi0 ~ (loglos0 + sex + age + outcome + adtime)^2,
     data = table)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.17179  -0.34257  -0.05306   0.26980   4.37032

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.5134576   0.0517890  164.388 < 2e-16 ***
loglos0           0.8556933   0.0160016   53.476 < 2e-16 ***
sexM              0.1609546   0.0405073    3.973 7.12e-05 ***
age              -0.0038052   0.0008674   -4.387 1.16e-05 ***
outcome          -0.0083361   0.0181751   -0.459 0.646487
adtime           -0.0071641   0.0207570   -0.345 0.729992
loglos0:sexM     -0.0076588   0.0094779   -0.808 0.419061
loglos0:age       0.0006624   0.0001925    3.441 0.000581 ***
loglos0:outcome -0.0081192   0.0048621   -1.670 0.094960 .
loglos0:adtime   -0.0091114   0.0052452   -1.737 0.082392 .
sexM:age         -0.0003907   0.0004071   -0.960 0.337163
sexM:outcome     -0.0265756   0.0117403   -2.264 0.023611 *
sexM:adtime       0.0049953   0.0110712    0.451 0.651850
age:outcome       0.0011690   0.0002427    4.816 1.48e-06 ***
age:adtime        0.0011459   0.0002167    5.289 1.25e-07 ***
outcome:adtime    0.0136464   0.0056265    2.425 0.015304 *
---
Signif. codes:
    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family
 taken to be 0.3421949)

Null deviance: 18087.9 on 15425 degrees of freedom
Residual deviance: 5273.2 on 15410 degrees of freedom
AIC: 27253

Number of Fisher Scoring iterations: 2
```

Fig. 11. GLM Anaysis on NHI Points (Total Cases, Logarithmic Transformed Data)

7 Conclusion

Since all the clinical information have been stored electronically as a hospital information system(HIS), it is now expected that mining such combined data will give a new insight to medical accidents.


```
glm(formula
     = lognhi ~ (loglos + sex + age + outcome + adtime)^2,
     data = data_tumor)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.254140	-0.253159	-0.006586	0.257345	3.272261

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.3787405	0.1231678	68.027	< 2e-16 ***
loglos	0.9676587	0.0318115	30.419	< 2e-16 ***
sexM	0.0635790	0.0850749	0.747	0.4549
age	-0.0006800	0.0018249	-0.373	0.7095
outcome	0.0149955	0.0366230	0.409	0.6822
adtime	0.0212879	0.0319450	0.666	0.5052
loglos:sexM	0.0270253	0.0161205	1.676	0.0937 .
loglos:age	-0.0005277	0.0003898	-1.354	0.1758
loglos:outcome	0.0109782	0.0071636	1.532	0.1255
loglos:adtime	-0.0412971	0.0086875	-4.754	2.06e-06 ***
sexM:age	-0.0017424	0.0008612	-2.023	0.0431 *
sexM:outcome	-0.0042336	0.0172197	-0.246	0.8058
sexM:adtime	0.0017416	0.0188084	0.093	0.9262
age:outcome	-0.0007851	0.0004409	-1.781	0.0750 .
age:adtime	0.0018626	0.0003839	4.852	1.26e-06 ***
outcome:adtime	0.0053574	0.0072173	0.742	0.4579

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family
taken to be 0.2561771)

Null deviance: 4940.4 on 4729 degrees of freedom
Residual deviance: 1207.6 on 4714 degrees of freedom
AIC: 6999.4

Number of Fisher Scoring iterations: 2

Fig. 12. GLM Anaysis on NHI Points (Neoplasm, Logarithmic Transformed Data)

In order to utilize information about risk extracted from information systems, we propose risk mining which integrates the following three important processes: risk detection, risk clarification and risk utilization.

Risk Detection discovers patterns or information unexpected to domain experts, which can be viewed as a sign of large scale accidents. In risk clarification, domain experts and data miners construct the model of the hidden mechanism of risk, focusing on detected risk information. If domain experts need more

information with finer granularity, we should collect more data with detailed information, and apply data mining to newly collected data.

Risk utilization evaluated clarified risk information in a real world environment to prevent risk events. If risk information is not enough to prevention, then more analysis is required. Thus, additional data collection is evoked for a new cycle of risk mining process.

As illustrative examples, we applied risk mining process to the following three domains.

First, we applied the total process to analysis of nurses' incident data. Data collected in 6 months were analyzed by rule induction methods, which detects several important factors for incidents (risk detection). Since data do not include precise information about these factors, we recollect incident data for 6 months to collect precise information about incidents. Then, rule induction is applied to new data. Domain experts discussed all the results obtained and found several important systematic errors in workflow (risk clarification). Finally, nurses changed workflow to prevent incidents and data were recollected for 6 months. Surprisingly, the frequency of medication errors has been reduced to one-tenth (risk utilization).

Then, we applied risk detection process to MRSA infection control, where we found several useful rules for background risk factors of MRSA infections. In the antimicrobial agents, cepheems have a weak risk factor to MRSA infections and lincomycins might have the variability of inducible MRSA-resistance.

Finally, we applied risk detection and clarification process to hospital administration, where we analyzed the following two datasets extracted from the hospital information system in Chiba University Hospital. One is the dataset extracted from the database on discharge summaries stored for about twenty years from 1978 to 2000. The other is combination of data from discharge summaries and accounting information system. The analysis gave the following results: (1) malignant neoplasm is the first major category which determines the profitability of Chiba University Hospital, which is stable for twenty years. (2) In a global view, the length of stay is the principle factor for the revenue of the hospital, whose distribution follows the log-normal distribution. (3) Treatment method may be a secondary factor to determine the distribution of the length of stay for each disease, which may be correlated with the property that the length of stay follows log-normal distribution for each minor division in total. (4) Treatment without a surgical operation should be more examined by additional information, which is also important to evaluate the profitability of the university hospital.

References

1. Tsumoto, S.: Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. *Information Sciences* (124), 125–137 (2000)
2. Tsumoto, S.: G5: Data mining in medicine. In: Kloesgen, W., Zytkow, J. (eds.) *Handbook of Data Mining and Knowledge Discovery*, pp. 798–807. Oxford University Press, Oxford (2001)

3. Quinlan, J.: C4.5 - Programs for Machine Learning. Morgan Kaufmann, Palo Alto (1993)
4. for Clinical Laboratory Standards, N.C.: Performance Standards for Antimicrobial Susceptibility Tests. Supplement M100-S14. National Committee for Clinical Laboratory Standards, Wayne, Pennsylvania (2004)
5. Schreckenberger, P., Ilendo, E., et al.: Incidence of constitutive and inducible clindamycin resistance in staphylococcus aureus and coagulase-negative staphylococci in a community and tertiary care hospital. *J. Clin. Microb.* 42, 2777–2778 (2004)
6. ICD9.chrisendres.com: Online icd9/icd9cm codes and medical dictionary.
<http://icd9cm.chrisendres.com/>
7. Walters, R.: M Programming: A Comprehensive Guide. Butterworth-Heinemann, Woburn (1997)
8. McCullagh, P., Nelder, J.: Generalized Linear Models, 2nd edn. CRC Press, Boca Raton (1990)

Using Cryptography for Privacy Protection in Data Mining Systems

Justin Zhan

Carnegie Mellon University
justinzhan@andrew.cmu.edu

Abstract. To conduct data mining including mining on the web data, we often need to collect data from various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. How multiple parties collaboratively conduct data mining without breaching data privacy presents a challenge. In this paper, we present a solution for privacy-preserving k-Medoids clustering which is one of data mining tasks. The solution is based on the cryptography technology.

1 Introduction

Data mining and knowledge discovery in databases are important research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. The field connects the three worlds of databases, artificial intelligence and statistics. The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if meaningful information or knowledge cannot be extracted from it. Data mining and knowledge discovery attempt to answer this need. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding an *a priori* hypotheses. As a field, it has introduced new concepts and algorithms such as association rule mining, classification, clustering, etc. Data mining techniques are widely used and are becoming more and more popular with time.

Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from the point of view of privacy preservation. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual benefit. Despite the potential gain, this is often not possible due to the confidentiality issues which arise. It is well documented [7] that the unlimited explosion of new information through the Internet and other media has reached a point where threats against the privacy are very common and they deserve serious thinking.

Let us consider an example. There are several hospitals involved into a multi-site medical study. Each hospital has its own data set containing patient records.

These hospitals would like to conduct data mining over the data sets from all of hospitals with the goal of more valuable information would be obtained via mining the joint data set. Due to privacy laws, one hospital cannot disclose their patient records to other hospitals.

How can these hospitals achieve their objective? The goal of this paper is to present technologies to solve privacy-preserving collaborative data mining problems over large data sets with reasonable efficiency.

Next, we will provide our building blocks which contain the notion of security, definition of privacy, homomorphic encryption, digital envelope technique and attack models that we consider.

2 Building Blocks

2.1 Notion of Privacy

In the last five years, the research community has developed numerous technical solutions for the privacy-preserving data mining. However, the notion of privacy that satisfies both technical and societal concerns is unknown. In this section, we would like to refer the privacy of definition proposed by Zhan et al. [27,28].

Security and privacy are related but different [24]. To achieve privacy, we often have to depend on security. In the privacy-preserving collaborative data mining, we need secure channels to protect network attackers.

For the purpose of illustration, we categorize the protection into two layers as shown in Figure 1. One is the protection against the collaborative parties,

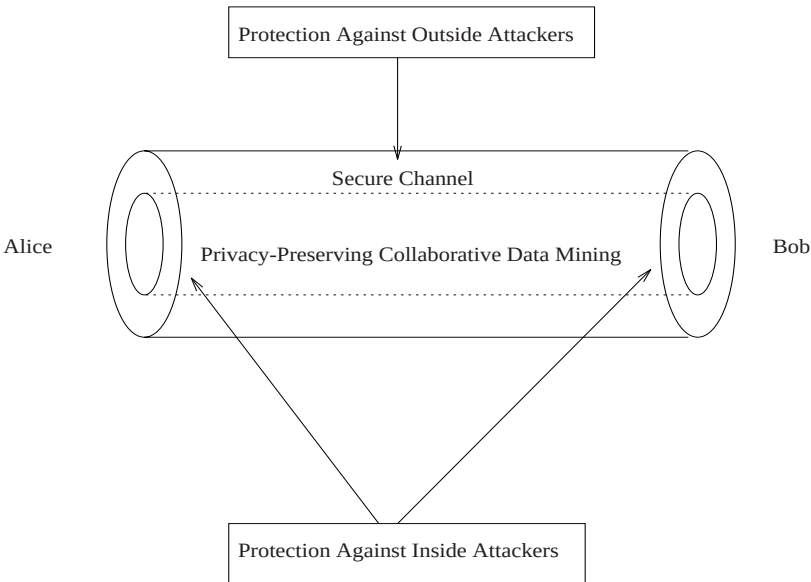


Fig. 1. Inside Attackers vs. Outside Attackers

the other is protection against network attackers. Without loss of generality, Let us call attacks from collaborative parties *inside attacks*, these parties are called *inside attackers*; let us call attacks outside the collaborative parties *outside attacks*, the attackers who conduct the attacks are called *outside attackers*.

To protect against outside attackers, we need to rely on secure channels. In this paper, we assume that the communications between collaborative parties are encrypted by a non-malleable encryption scheme. Our focus is to prevent inside attackers from knowing private data in collaborative data mining.

Prevention of inside attackers is different from prevention of outside attackers in that the inside attackers usually have more knowledge about private data than outside attackers. Furthermore, the goal of collaborative data mining is to obtain valid data mining result. However, the result itself may disclose the private data to inside attackers. Therefore, we cannot hope to achieve the same level of protection in privacy-preserving collaborative data mining as in general secure communications which protect against the outside attackers. However, we would like to prevent the private data from being disclosed during the mining stage. In order to state more precisely how we understand privacy in the data mining context, we propose the following definition:

Definition 1. *A privacy-oriented scheme S preserves data privacy if for any private data T , the following is held:*

$$|Pr(T|PPDMS) - Pr(T)| \leq \epsilon$$

where

- $PPDMS$: Privacy-preserving data mining scheme.
- ϵ : A probability parameter.
- $Pr(T|PPDMS)$: The probability that the private data T is disclosed after privacy-preserving data mining schemes being applied.
- $Pr(T)$: The probability that the private data T is disclosed without any privacy-preserving data mining scheme being applied.
- $Pr(T|PPDMS) - Pr(T)$: The probability that private data T is disclosed with and without privacy-preserving data mining schemes being applied.

We call ϵ the privacy level that the privacy-oriented scheme S can achieve. The goal is to make ϵ as small as possible.

We have defined privacy for data mining algorithms. However, often time, we need to reduce the whole privacy-preserving data mining algorithm to a set of component privacy-oriented protocols. We say a privacy-preserving data mining algorithm preserves privacy if each component protocol preserves privacy and the combination of the component protocols do not disclose private data. In the secure multiparty computation literature, a composition theorem [11] describes the similar idea.

Theorem 1. *Suppose that g is privately reducible to f and that there exists a protocol for privately computing f . Then there exists a protocol for privately computing g .*

Proof. Refer to [11].

We now formally define privacy for a component protocol.

Definition 2. A privacy-oriented component protocol CP preserves data privacy if for any private data T , the following is held:

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

where

- CP : Component protocol.
- $Pr(T|CP)$: The probability that the private data T is disclosed after a privacy-preserving component protocol being applied.
- $Pr(T) - Pr(T|CP)$: The probability that private data T is disclosed with and without privacy-preserving component protocol.

We call ϵ the privacy level that the privacy-oriented component protocol CP can achieve. The goal is to make ϵ as small as possible.

Next, we introduce the fundamental building blocks for this paper. They are homomorphic encryption, digital envelope technique and attack models that we will consider.

2.2 Homomorphic Encryption

A cryptosystem is homomorphic with respect to some operation $*$ on the message space if there is a corresponding operation $*'$ on the ciphertext space such that $e(m) *' e(m') = e(m * m')$. In our privacy-oriented protocols, we use additive homomorphism offered by [18] in which Paillier proposed a new trapdoor mechanism based on the idea that it is hard to factor number $n = pq$ where p and q are two large prime numbers. In the performance evaluation, Paillier compares the proposed encryption scheme with existing public-key cryptosystems. The results show that the encryption process is comparable with encryption process of RSA in terms of the computation cost; the decryption process is faster than the decryption process of RSA.

In this paper, we utilize the following property of the homomorphic encryption functions: $e(m_1) \times e(m_2) = e(m_1 + m_2)$ where m_1 and m_2 are the data to be encrypted. Because of the property of associativity, $e(m_1 + m_2 + \dots + m_n)$ can be computed as $e(m_1) \times e(m_2) \times \dots \times e(m_n)$ where $e(m_i) \neq 0$. That is

$$e(m_1 + m_2 + \dots + m_n) = e(m_1) \times e(m_2) \times \dots \times e(m_n). \quad (1)$$

2.3 Digital Envelope

A digital envelope [4] is a random number (or a set of random numbers) only known by the owner of private data. To hide the private data in a digital envelope, we conduct a set of mathematical operations between a random number (or a set of random numbers) and the private data. The mathematical operations could be addition, subtraction, multiplication, etc. For example, assume the private data value is v . There is a random number R which is only known by the owner of v . The owner can hide v by adding this random number, e.g., $v + R$.

Goal-Oriented Attack Model. Secure communication in networks is often goal-oriented. For example, there are two people (e.g., Alice and Bob) who want to talk to each other. Suppose that Alice would like to send message m to Bob, the goal of Alice is to let Bob receive message m ; the goal of Bob is to receive m from Alice. In the meanwhile, they want to keep the communication secure so that network attackers cannot see m or modify it in a meaningful way. To achieve this goal, they apply encryption to message m .

Privacy-preserving collaborative data mining is also goal-oriented. The fundamental goal of privacy-preserving collaborative data mining in our framework is to obtain the accurate results while keeping the data private. In other words, we need to obtain accurate data mining results without letting each party see other parties' private data. This is an ideal case. In fact, the data mining results may disclose partial information about the private data. Therefore, a practical goal of privacy preservation is to hide private data during the mining stage for a given privacy-preserving data mining problem.

In this paper, we define our attack model as *Goal-Oriented Attack Model*. In this model, all the collaborative parties need to follow their goals. The basic goal of collaborative data mining is to obtain desired data mining results which are shared among the collaborative parties. The attacks can be applied as long as they follow their basic goal and the purpose of the attacks of one party (or a group of parties) is to gain useful information about the data of the other party (or the other group of parties).

3 Privacy-Preserving Collaborative Data Mining Problems

Our approaches can be used for the typical privacy-preserving collaborative data mining problems: Multiple parties, each having a private data set which is denoted by DS_1 , DS_2 , $\dots$ and DS_n respectively, want to collaboratively conduct a data mining task on the concatenation of their data sets. Because they are concerned about their data privacy or due to the legal privacy rules, neither party is willing to disclose its actual data set to others.

In this paper, we consider the privacy-preserving collaborative k-Medoids Clustering problem which is defined as follows:

Problem 1. P_1 has a private data set DS_1 , P_2 has a private data set DS_2 , $\dots$ and P_n has a private data set DS_n . The data set $[DS_1 \cup DS_2 \cup \dots \cup DS_n]$ forms a dataset, which is actually the concatenation of DS_1 , DS_2 , $\dots$ and DS_n . The n parties want to conduct k-Medoids clustering over $[DS_1 \cup DS_2 \cup \dots \cup DS_n]$ to obtain the mining results satisfying the given constraints.

There are two types of collaborative models. In the vertical collaboration, diverse features of the same set of data are collected by different parties. In the horizontal collaboration, diverse sets of data, all sharing the same features, are gathered by different parties. The collaborative model that we consider is the *vertical collaboration*.

4 Privacy-Preserving K-Medoids Clustering

Clustering is the process of grouping a set of objects into classes or clusters so that objects within a cluster have similarity in comparison to one another, but are dissimilar to objects in other clusters [12]. In other words, clustering is a process of finding natural groupings in a set of data. It has been widely used in the applications such as marketing, city planning, insurance, medicine, chemistry, remote sensing, and so on. A complete review of the current state-of-the-art of clustering techniques can be found in [3]. There are many clustering algorithms such as k-means method, k-medoids method, probabilistic clustering, etc. We focus on k-medoids method since it allows arbitrary objects that are not limited to numerical attributes [15]. In k-medoids clustering, a cluster is denoted by one of its points. It is an easy solution in that it covers any attribute type and that medoids are resistant against outliers. Once medoids are chosen, clusters are defined as subsets of points close to respective medoids, and the objective function is described as the distance between a point and its medoid. We target k-medoids clustering instead of k-means clustering. From the privacy protection point of view, the k-medoids clustering seems more challenging than the k-means clustering [23] because we always use the real instance to compute the distances in k-medoids clustering while we use the *mean instance* whose values are the means of the real instance values to compute the distances.

4.1 Overview of k-Medoids Clustering Algorithm

The k-medoids method divides a distance-space into k clusters. A medoid [15], that is selected from the dataset, represents a cluster. The algorithm chooses k medoids to denote the k clusters. Clusters are then created by assigning each of the remaining instances to the nearest medoid. Although the k-medoids clustering algorithm is different from the k-nearest neighbor classification algorithm, both algorithms do have similarity.

- Similarity: Both algorithms need to compute and compare the distance between a special instance and each of other instances. In k-medoids clustering, we need to compute the distance between a medoid and other instances. In k-nearest neighbor classification, we need to compute the distance between a query instance and other instances.
- Difference: The purposes of both algorithms are different. The purpose of k-medoids clustering is to compute the distance between one instance and each of k medoids, then assign this instance to a medoid with the smallest distance value, and repeat this process for all the instances. However, the purpose of k-nearest neighbor classification is to compute the distance between a query instance and each instance in the dataset, select k closest instances, and assign a class label for the query instance by the majority class principle.

We describe the k-medoids clustering algorithm in the following.

Algorithm 1.

1. Arbitrarily select k instances from the dataset as medoids.
2. Assign each remaining (non-medoid) instance to the cluster with the nearest medoid.
3. Compute the compactness of a clustering, denoted by $TD_{current}$.

$$TD = \sum_{i=1}^k TD(C_i) \quad (2)$$

$$TD(C_i) = \sum_{t \in C_i} dist(t, m_{C_i}) \quad (3)$$

4. For each pair (medoid M and non-medoid NM)
 - Compute the value of TD for the partition that results from swapping M with NM , denoted by $TD_{NM \longleftrightarrow M}$.
5. Select the non-medoid NM for which $TD_{NM \longleftrightarrow M}$ is minimal.
6. If $TD_{NM \longleftrightarrow M} < TD_{current}$
 - Swap NM with M
 - Set $TD_{current}$ to be $TD_{NM \longleftrightarrow M}$.
 - Go to Step 4.

The algorithm requires a distance function. For instance, the distances can be defined in terms of standard Euclidean distance. As we will discuss, each party computes her own portion of the distance and utilization of certain distance measure does not cause privacy violation. Therefore, other distance functions can be applied as well.

4.2 Notations

We define the following notations for illustration purposes.

- k : the total number of medoids.
- t : a non-medoid instance.
- m_{C_i} : the medoid of the cluster C_i .
- M : a general term for medoids. It contains all possible medoids.
- NM : a general term for non-medoids. It contains all possible medoids.
- $TD(C_i)$: the measure of the compactness for a cluster C_i .
- TD : the measure of the compactness of a clustering that contains all the clusters.

4.3 The Scenarios Where the Private Data Maybe Exposed

The key step of the k-medoids clustering algorithm is the computation of the distance between each non-medoid t and its medoid m_{C_i} without disclosing their private data. There are two cases where we need privacy-oriented computations: (1) Assign each non-medoid instance to the cluster with the nearest medoid. (2) Compute TD . That is, for a particular cluster, computing the distances between each non-medoid instance and its medoid; then adding all the distances together to obtain $TD(C_i)$. TD can then be computed by summation of $TD(C_i)$ for all k clusters. Given a non-medoid instance t , multiple parties want to compute the distance between t and its medoid instances m_{C_i} . Privacy-oriented protocols are developed in the next section to enforce such computations without sacrificing data privacy.

4.4 Privacy-Preserving Protocols for Vertical Collaboration

In vertical collaboration, since each party holds only a portion of attributes for each instance, each party computes her portion of the distance (called the *distance portion* which is the square of the standard Euclidean distance) according to her attribute set. To decide the nearest medoid of t , all the parties need to sum their distance portions together, then compare the summation. For example, assume that the distance portions between t and the medoid instance m_{C_i} are $s_{11}, s_{12}, \dots, s_{1n}$; and the distance portions between t and the medoid instance m_{C_j} ($i \neq j$) are $s_{21}, s_{22}, \dots, s_{2n}$ where s_{1j} and s_{2j} belong to P_j for $j \in [1, n]$. To compute whether the distance between the medoid instance m_{C_i} and t is larger than the distance between the medoid instance m_{C_j} and t , we need to evaluate the expression $\sum_{i=1}^n s_{1i} \geq \sum_{i=1}^n s_{2i}$.

Problem 2. Assume that P_j has a private distance portion of the i th instance, s_{ij} , for $i \in [1, k], j \in [1, n]$, the problem is to decide whether $\sum_{j=1}^n s_{ij} \leq \sum_{j=1}^n s_{lj}$ for $i, l \in [1, k]$ ($i \neq l$) and select the smallest value $TD(C_i)$, without disclosing each distance portion.

Highlight of Protocol 1: Our protocol has four steps. (1) Key and digital envelope generation: multiple parties select one of them, e.g., P_n , as the key generator, who creates a cryptographic key pair (e, d) of a semantically-secure homomorphic encryption scheme. Each party generates k digital envelopes. (2) Computing $e(\sum_{j=1}^n (s_{ij} + R_{ij}))$ for $i \in [1, k]$: each party puts her private distance portion into a digital envelope and sends it to P_{n-1} . (3) Computing $e(\sum_{j=1}^n R_{ij})$, for all $i \in [1, k]$: each party encrypts her digital envelopes and sends them to P_1 . (4) P_1, P_{n-1} and P_n jointly compute the nearest medoid.

We present the formal protocol as follows:

Protocol 1.

Step I: Key and digital envelope generation.

1. P_j s for $j \in [1, n]$ randomly select a key generator, e.g., P_n .

2. P_n generates a cryptographic key pair (e, d) of a semantically-secure homomorphic encryption scheme and publishes its public key e .
3. Each party independently generates k digital envelopes, i.e., P_j generates k digital envelopes R_{ij} , for all $i \in [1, k], j \in [1, n]$.

Step II: To compute $e(\sum_{j=1}^n (s_{ij} + R_{ij}))$ for $i \in [1, k]$.

1. P_1 computes $e(s_{i1} + R_{i1})$, for $i \in [1, k]$, and sends them to P_2 .
2. P_2 computes $e(s_{i1} + R_{i1}) \times e(s_{i2} + R_{i2}) = e(s_{i1} + s_{i2} + R_{i1} + R_{i2})$, where $i \in [1, k]$, and sends them to P_3 .
3. Repeat sub-step 1 and 2 of step II until P_{n-1} obtains $e(s_{i1} + s_{i2} + \dots + s_{i(n-1)} + R_{i1} + R_{i2} + \dots + R_{i(n-1)})$, for all $i \in [1, k]$.
4. P_n computes $e(s_{in} + R_{in})$ for $i \in [1, k]$, and sends them to P_{n-1} .
5. P_{n-1} computes $e(s_{i1} + s_{i2} + \dots + s_{i(n-1)} + R_{i1} + R_{i2} + \dots + R_{i(n-1)}) \times e(s_{in} + R_{in}) = e(s_{i1} + s_{i2} + \dots + s_{i(n-1)} + s_{in} + R_{i1} + R_{i2} + \dots + R_{i(n-1)} + R_{in}) = e(\sum_{j=1}^n (s_{ij} + R_{ij}))$, $i \in [1, k]$. Let $e(S + R)$ denote the k encrypted elements as follows: $[e(S_1 + R_1), e(S_2 + R_2), \dots, e(S_k + R_k)]$, where $S_i = \sum_{j=1}^n s_{ij}$ and $R_i = \sum_{j=1}^n R_{ij}$.

Step III: To compute $e(\sum_{j=1}^n R_{ij})$ for all $i \in [1, k]$.

1. P_n computes $e(R_{in})$ for $i \in [1, k]$ and sends them to P_{n-1} .
2. P_{n-1} computes $e(R_{in}) \times e(R_{i(n-1)}) = e(R_{in} + R_{i(n-1)})$ for $i \in [1, k]$, and sends them to P_{n-2} .
3. Repeat the sub-step 1 and 2 of step III until P_1 obtains $e(R_{i1} + R_{i2} + \dots + R_{i(n-1)}) \times e(R_{in}) = e(\sum_{j=1}^n R_{ij})$, for all $i \in [1, k]$. The k encrypted elements are denoted by $e(R)$ that contains the following: $[e(R_1), e(R_2), \dots, e(R_k)]$ where $R_i = \sum_{j=1}^n R_{ij}$.

Step IV: To compute the nearest medoid.

1. Computation between P_1 and P_n .
 - (a) P_1 randomly permutes $e(R_1), e(R_2), \dots, e(R_k)$, then sends the permuted elements to P_n .
 - (b) P_n decrypts each element and sends them to P_1 in the same order as P_1 did.
 - (c) P_1 computes R that contains the following: $[R_1, R_2, \dots, R_k]$. Note that P_1 can do it since she has the permutation function.
2. Computation between P_{n-1} and P_n .
 - (a) P_{n-1} randomly permutes $e(S_1), e(S_2), \dots, e(S_k)$, then sends the permuted elements to P_n .
 - (b) P_n decrypts each element and sends them to P_{n-1} in the same order as P_{n-1} did.
 - (c) P_{n-1} computes $[S_1 + R_1, S_2 + R_2, \dots, S_k + R_k]$ denoted by $S + R$. Note that: (1) P_{n-1} can do it since she has the permutation function. (2) The permutation function that P_1 used is independent from the permutation function that P_{n-1} used.

3. P_{n-1} and P_1 compute $e(S_i - S_l) = e(\sum_{j=1}^n s_{ij} - \sum_{j=1}^n s_{lj})$, for $i, l \in [1, k] (i \neq l)$, and collects the results into a sequence Φ which contains $k(k-1)$ elements. This computation can be achieved via the following process:
 - (a) P_1 computes $e(R_l)$ and $e(-R_i)$ for $i, l \in [1, k] (i \neq l)$, then sends them to P_{n-1} .
 - (b) P_{n-1} computes $e(S_i - S_l)$ for $i, l \in [1, k] (i \neq l)$ as follows:
 - $e(S_i + R_i) \times e(-R_i) = e(S_i)$.
 - $e(-S_l - R_l) \times e(R_l) = e(-S_l)$.
 - $e(S_i) \times e(-S_l) = e(S_i - S_l)$.
4. Computation between P_{n-1} and P_n .
 - (a) P_{n-1} randomly permutes this sequence Φ and obtains the permuted sequence denoted by Φ' , then sends Φ' to P_n . Note that the permutation is independent from the ones she used.
 - (b) P_n decrypts each element in sequence Φ' . He assigns the element $+1$ if the result of decryption is not less than 0, and -1 , otherwise. Finally, he obtains a $+1/-1$ sequence denoted by Φ'' .
 - (c) P_n sends Φ'' to P_{n-1} who computes the smallest element. (Details are given in Protocol 4 [27].) It is the nearest medoid for a given non-medoid instance t . He then decides the cluster to which t belongs.

The Correctness Analysis of Protocol 1: To show that the protocol correctly finds the nearest medoid for a given non-medoid instance t , we analyze step by step. In step II, P_{n-1} obtains $e(s_{i1} + R_{i1}) \times e(s_{i2} + R_{i2}) \times e(s_{i3} + R_{i3}) \times \dots \times e(s_{in} + R_{in}) = e(s_{i1} + R_{i1} + s_{i2} + R_{i2} + \dots + s_{in} + R_{in}) = e(\sum_{j=1}^n (s_{ij} + R_{ij}))$, for $i \in [1, k]$. In step III, P_1 finds $e(R_{i1}) \times (R_{i2}) \times \dots \times e(R_{in}) = e(\sum_{j=1}^n R_{ij})$, for $i \in [1, k]$. In step IV, during sub-step 1-3, P_{n-1} obtains $e(\sum_{j=1}^n s_{ij} - \sum_{j=1}^n s_{lj})$ for $i, l \in [1, k] (i \neq l)$. Following the detailed description on how to compute the smallest element, we know that P_{n-1} finds the nearest medoid for a given non-medoid, which is the desired result.

The Complexity Analysis of Protocol 1: The communication cost of this protocol is $\alpha(3n + k^2 + 5k - 3)$.

The computational costs are contributed by: (1) The generation of kn digital envelopes. (2) nk additions. (3) $k^2 + k + (2n - 2)k$ multiplications. (4) $2kn$ encryptions. (5) $\frac{1}{2}k(k - 1)$ decryptions. (6) $4k + k(k - 1)$ permutations. (7) $\frac{1}{2}k(k - 1)$ assignments when P_n computes Φ'' .

Therefore, the total computation cost is $g_4kn + nk + k^2 + k + (2n - 2)k + 12kn + \frac{13}{2}k(k - 1) + g_2(4k + k^2 - k) + \frac{1}{2}k(k - 1)$.

Theorem 2. Protocol 1 preserves data privacy at a level equal to ADV_{P_n} .

Proof. We will identify the value of ϵ such that

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

holds for $T = T_{P_i}$, $i \in [1, n]$, and $CP = \text{Protocol 1}$.

According to our notation,

$$ADV_{P_n} = Pr(T_{P_i} | VIEW_{P_n}, Protocol) - Pr(T_{P_i} | VIEW_{P_n}), i \neq n,$$

and

$$ADV_{P_i} = Pr(T_{P_j} | VIEW_{P_i}, Protocol) - Pr(T_{P_j} | VIEW_{P_i}), i \neq n, j \neq i.$$

The information that P_i where $i \neq n$ obtains from other parties is encrypted by e that is semantically secure,

$$ADV_{P_i} = ADV_S.$$

In order to show that privacy is preserved according to Definition 2, we need to know the value of the privacy level ϵ . We set

$$\epsilon = \max(ADV_{P_n}, ADV_{P_i}) = \max(ADV_{P_n}, ADV_S) = ADV_{P_n}.$$

Then

$$Pr(T_{P_i} | VIEW_{P_n}, Protocol) - Pr(T_{P_i} | VIEW_{P_n}) \leq ADV_{P_n}, i \neq n,$$

and

$$Pr(T_{P_j} | VIEW_{P_i}, Protocol) - Pr(T_{P_j} | VIEW_{P_i}) \leq ADV_{P_n}, i \neq n, j \neq i,$$

which completes the proof ¹.

Next, we will discuss how to privately compute TD .

4.5 Privacy-Preserving Protocol for Computing TD

Once each non-medoid instance is assigned to the nearest medoid, we need to compute the compactness of a clustering.

$$TD = \sum_{i=1}^k TD(C_i) = \sum_{i=1}^k \sum_{t \in C_i} dist(t, m_{C_i}), \quad (4)$$

To compute TD , each party computes her local distance portions between each non-medoid instance and the corresponding medoid instance, and adds them together. For the purpose of illustration, let us assume that P_j gets a distance portion v_j for $j \in [1, n]$. In order to compute TD , we need to compute $\sum_{j=1}^n v_j$. The Protocol 3 in [27] can be applied to solve this problem.

¹ Note that all the information that P_n obtains from other parties is $\sum_{j=1}^n s_{ij} - \sum_{j=1}^n s_{lj}$, for $i, l \in [1, k] (i \neq l)$ in random order.

4.6 Privacy-Preserving Protocol for Horizontal Collaboration

In horizontal collaboration, the problem is reduced to how to compute the distance between two instances without disclosing private data.

Problem 3. Assume that P_f has an instance vector $\vec{x}_f$ and P_g has an instance vector $\vec{x}_g$. Both vectors have $\mathcal{Y}$ elements. We use x_{fi} to denote the i th element in vector $\vec{x}_f$, and x_{gi} to denote the i th element in vector $\vec{x}_g$. The goal is to compute the $dist(\vec{x}_f, \vec{x}_g)$ without disclosing their actual data to each other.

Highlight of Protocol 2: In our privacy-oriented protocol, P_f denotes the party who has the medoid instance, and P_g denotes the party who has the non-medoid instance t . P_f generates a cryptographic key pair (e, d) of a semantically-secure homomorphic encryption scheme and publishes its public key e . She encrypts each element of x_{fi} for $i \in [1, \mathcal{Y}]$ and sends them to P_g . P_g computes $e(x_{fi} - x_{gi})$, $e[(x_{fi} - x_{gi})^2]$, and $e[\sum_{i=1}^{\mathcal{Y}} (x_{fi} - x_{gi})^2]$. P_g sends $e[\sum_{i=1}^{\mathcal{Y}} (x_{fi} - x_{gi})^2]$ to P_f who decrypts it and compute the $dist(x_f, x_g)$. We describe this more formally as follows.

We present the formal protocol as follows:

Protocol 2.

INPUT: P_f 's input is a vector $\vec{x}_f = \{x_{f1}, x_{f2}, \dots, x_{f\mathcal{Y}}\}$, and P_g 's input is a vector $\vec{x}_g = \{x_{g1}, x_{g2}, \dots, x_{g\mathcal{Y}}\}$. The elements in the input vectors are taken from the real number domain.

OUTPUT: $dist(x_f, x_g)$.

1. P_f performs the following operations:
 - (a) P_f generates a cryptographic key pair (e, d) of a semantically-secure homomorphic encryption scheme and publishes its public key e .
 - (b) P_f sends $e(x_{f1}), e(x_{f2}), \dots, e(x_{f\mathcal{Y}})$ to P_g .
2. P_g computes the following operations:
 - (a) $e(x_{f1}) \times e(-x_{g1}) = e(x_{f1} - x_{g1})$, $e(x_{f2} \times e(-x_{g2})) = e(x_{f2} - x_{g2})$, $\dots$, $e(x_{f\mathcal{Y}}) \times e(-x_{g\mathcal{Y}}) = e(x_{f\mathcal{Y}} - x_{g\mathcal{Y}})$.
 - (b) $e(x_{f1} - x_{g1}) + e(x_{f1} - x_{g1}) = e[(x_{f1} - x_{g1})^2]$, $\dots$, $e(x_{f\mathcal{Y}} - x_{g\mathcal{Y}}) + e(x_{f\mathcal{Y}} - x_{g\mathcal{Y}}) = e[(x_{f\mathcal{Y}} - x_{g\mathcal{Y}})^2]$.
 - (c) $e[(x_{f1} - x_{g1})^2] \times e[(x_{f2} - x_{g2})^2] \times \dots \times e[(x_{f\mathcal{Y}} - x_{g\mathcal{Y}})^2] = e(dist(x_f, x_g))$.
 - (d) P_g sends $e(dist(x_f, x_g))$ to P_f .
3. P_f computes $d[e(dist(x_f, x_g))] = dist(x_f, x_g)$.

The Correctness Analysis of Protocol 2: To show that the protocol correctly computes $dist(x_f, x_g)$. P_g has three encrypted sequences: (1) the sequence of $e(x_{fi})$; (2) the sequence of $e(x_{fi} - x_{gi})$; (3) the sequence of $e[(x_{fi} - x_{gi})^2]$ where $i \in [1, \mathcal{Y}]$. The first sequence is sent by Alice. The second sequence is computed by $e(x_{fi}) \times e(-x_{gi}) = e(x_{fi} - x_{gi})$, for $i \in [1, k]$, according to Equation 1. The third sequence also follows Equation 1, e.g., $e(x_{fi} - x_{gi}) + e(x_{fi} - x_{gi}) = e((x_{fi} - x_{gi})^2)$. Bob computes $e[(x_{f1} - x_{g1})^2] \times e[(x_{f2} - x_{g2})^2] \times \dots \times e[(x_{f\mathcal{Y}} - x_{g\mathcal{Y}})^2]$ which is $e(dist(x_f, x_g))$ according to Equation 1 and Equation 1. P_g then sends $e(dist(x_f, x_g))$ to P_f who decrypts it and obtains $dist(x_f, x_g)$.

The Complexity Analysis of Protocol 2: The bit-wise communication cost of this protocol is $\alpha(\mathcal{T} + 1)$ where $\mathcal{T}$ is the number of attributes in each instance.

The computational cost is caused by the following: (1) The generation of a cryptographic key pair and $\mathcal{T}$ digital envelopes. (2) One decryption. (3) The total number of $2\mathcal{T}$ encryptions. (4) $2\mathcal{T} - 2$ multiplications. (5) One summation.

Therefore, the total computation cost is $g_4\mathcal{T} + 13 + 12\mathcal{T} + 2\mathcal{T} - 2 + 1 = (g_4 + 14)\mathcal{T} + 14$.

Theorem 3. *Protocol 2 preserves data privacy at a level equal to ADV_{P_f} .*

Proof. We will identify the value of ϵ such that

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

holds for $T = T_{P_i}$, $i \in [1, n]$, and $CP = \text{Protocol 2}$.

According to our notation,

$$ADV_{P_g} = Pr(T_{P_i} | VIEW_{P_g}, \text{Protocol 2}) - Pr(T_{P_i} | VIEW_{P_g}), i \neq g,$$

and

$$ADV_{P_f} = Pr(T_{P_i} | VIEW_{P_f}, \text{Protocol 2}) - Pr(T_{P_i} | VIEW_{P_f}), i \neq f.$$

Since all the information that P_g obtains is $e(x_{fi})$ for $i \in [1, \mathcal{T}]$ and e is semantically secure,

$$ADV_{P_g} = ADV_S.$$

In order to show that privacy is preserved according to Definition 2, we need to know the value of the privacy level ϵ . We set

$$\epsilon = \max(ADV_{P_g}, ADV_{P_f}) = \max(ADV_S, ADV_{P_f}) = ADV_{P_f}.$$

Then

$$Pr(T_{P_i} | VIEW_{P_g}, \text{Protocol 2}) - Pr(T_{P_i} | VIEW_{P_g}) \leq ADV_{P_g}, i \neq g,$$

and

$$Pr(T_{P_i} | VIEW_{P_f}, \text{Protocol 2}) - Pr(T_{P_i} | VIEW_{P_f}) \leq ADV_{P_f}, i \neq f,$$

which completes the proof.

Now, there are k distances. Assume that each party holds one of them, to compute the nearest medoid, k parties need to be involved in the computation.

(Note that, if any party holds more than one distance, then she only takes the smallest one to compare with other distance held by other parties.) To obtain the smallest distance, one distance needs to be compared against all other distances to decide whether it is smaller than any one of them. If it is, then the smallest is found; otherwise, we need to take another distance and compare it with all others. One distance will be always smaller than all the others, or equal to some of them. In the following, we show how one distance can be privately compared with all other distances.

Highlight of Protocol 3: A key generator is randomly selected, assume that it is P_k who generates a cryptographic key pair (e, d) of a semantically-secure homomorphic encryption scheme and publishes its public key e . P_1 generates $k - 2$ digital envelopes denoted by R_i for $i \in [2, k - 1]$. P_1 computes $e(S_1 + R_i - S_i) \times e(-R_i) = e(S_1 - S_i)$ for $i \in [2, k]$. P_k finally computes the smallest element.

We present the formal protocol as follows:

Protocol 3 . *INPUT:* P_j 's input is a distance S_j $j \in [1, k]$.

OUTPUT: Whether S_1 is the smallest element.

1. *Key and digital envelope generation*
 - (a) A key generator is randomly selected, assume that it is P_k .
 - (b) P_k generates a cryptographic key pair (e, d) of a semantically-secure homomorphic encryption scheme and publishes its public key e . Let $e(\cdot)$ denote encryption and $d(\cdot)$ denote decryption.
 - (c) P_1 generates $k - 2$ digital envelopes denoted by R_i for $i \in [2, k - 1]$.
2. *Computation between P_1 and P_i for $i \in [2, k]$*
 - (a) P_1 sends $e(S_1 + R_i)$ to P_i for $i \in [2, k - 1]$.
 - (b) P_i computes $e(S_1 + R_i) \times e(-S_i) = e(S_1 + R_i - S_i)$ for $i \in [2, k - 1]$.
 - (c) P_i sends $e(S_1 + R_i - S_i)$, for $i \in [2, k - 1]$, to P_1 .
 - (d) P_k sends $e(-S_k)$ to P_1 .
 - (e) P_1 computes $e(S_1 + R_i - S_i) \times e(-R_i) = e(S_1 - S_i)$ for $i \in [2, k]$.
3. *Communication of permuted sequence*
 - (a) P_1 randomly permutes the following sequence: $e(S_1 - S_2), e(S_1 - S_3), e(S_1 - S_4), \dots, e(S_1 - S_k)$. Let's denote the permuted sequence by E_1 . P_1 then sends E_1 to P_2 .
 - (b) P_2 randomly permutes E_1 and gets E_2 , then sends E_2 to P_3 .
 - (c) Repeat (a), (b) until P_{k-1} gets E_{k-1} and sends E_{k-1} to P_k .
4. P_k decrypts each element of E_{k-1} . If all the decrypted elements are less than or equal to 0, then P_k concludes that S_1 is the smallest element; otherwise, S_1 is not the smallest element.

To obtain the smallest element, the above protocol needs to be run k times in the worst case.

The Correctness Analysis of Protocol 3: To show that the protocol outputs yes if S_1 is the smallest element, No otherwise, we analyze step by step. In step II, P_1 obtains $e(S_1 - S_2), e(S_1 - S_3), \dots, e(S_1 - S_k)$. In step III, P_k obtains

$e(S_1 - S_2), e(S_1 - S_3), \dots, e(S_1 - S_k)$ in a permuted form. In step IV, P_k decrypts each element of the permuted sequence. P_k then gets $S_1 - S_2, S_1 - S_3, \dots, S_1 - S_k$ in the permuted order. P_k knows that if $S_1 \leq S_i$ for $i \in [2, k]$, then $S_1 - S_i \leq 0$ and S_1 is the smallest element. Otherwise, S_1 is not the smallest element. Note that there are possible more than one smallest element.

The Complexity Analysis of Protocol 3: The communication cost of this protocol is $\alpha(k^2 - 3)$ where k is the total number of parties involved into this protocol.

The computational costs are contributed by: (1) The generation of a cryptographic key pair and $k-2$ digital envelopes. (2) $k-2$ additions. (3) $k-1$ multiplications. (4) $2k-3$ encryptions. (5) $k-1$ decryption. (6) A permutation of $k-1$ random numbers of E_i in step 3.

Therefore, the total computation cost is $g_1 + g_4(k-2) + k-2 + k-1 + 12k - 18 + 13k - 13 + g_2(k-1) = (27 + g_2 + g_4)k + g_1 - g_2 - 2g_4 - 34$.

Theorem 4. *Protocol 3 preserves data privacy at a level equal to ADV_{P_k} .*

Proof. We will identify the value of ϵ such that

$$|Pr(T|CP) - Pr(T)| \leq \epsilon$$

holds for $T = T_{P_i}$, $i \in [1, n]$, and $CP = \text{Protocol 3}$.

According to our notation,

$$ADV_{P_1} = Pr(T_{P_i} | VIEW_{P_1}, \text{Protocol 3}) - Pr(T_{P_i} | VIEW_{P_1}), i \neq 1,$$

and

$$ADV_{P_k} = Pr(T_{P_i} | VIEW_{P_k}, \text{Protocol 3}) - Pr(T_{P_i} | VIEW_{P_k}), k \neq 1, k \neq i.$$

Since all the information that P_1 obtains is $e(S_1 - S_i)$ ($i \in [2, k-1]$) and e is semantically secure,

$$ADV_{P_1} = ADV_S.$$

In order to show that privacy is preserved according to Definition 2, we need to know the value of the privacy level ϵ . We set

$$\epsilon = \max(ADV_{P_1}, ADV_{P_k}) = \max(ADV_S, ADV_{P_k}) = ADV_{P_k}.$$

Then

$$Pr(T_{P_i} | VIEW_{P_1}, \text{Protocol 3}) - Pr(T_{P_i} | VIEW_{P_1}) \leq ADV_{P_k}, i \neq 1.$$

$$Pr(T_{P_i}|VIEW_{P_k}, Protocol3) - Pr(T_{P_i}|VIEW_{P_k}) \leq ADV_{P_k}, k \neq 1, k \neq i.$$

which completes the proof ².

Next, to privately compute TD , Protocol 3 in [27] can still be utilized.

4.7 Overall Complexity Overhead Analysis

In this section, we analyze the overall efficiency of our proposed solutions. We have provided the computation cost and communication cost for each component protocol. We will combine them in order to achieve privacy-preserving k-medoids clustering. The overall communication and computation overhead for vertical collaboration among multiple parties is $\alpha(3n+k^2+5k-3)i$ and $((8+g_2)k^2+(15+g_2)kn+(3g_2-8)k)i$ where i is the number of iterations that the algorithm needs to be run. The overall communication and computation overhead for horizontal collaboration among multiple parties is $\alpha(\Upsilon+k^2-2)i$ and $((14+g_2)\Upsilon+12+g_1+27k+2g_2k-3g_2+g_1-34)i$ where i is the number of iterations that the algorithm needs to be run.

4.8 Discussion

To protect actual data from being disclosed, one approach is to alter the data in a way that actual individual data values cannot be recovered, while certain computations can still be applied to the data. Due to the fact that the actual data are not provided for the mining, the privacy of data is preserved. This is the core idea of randomization-based techniques.

The random perturbation technique is usually realized by adding noise or uncertainty to actual data such that the actual values are prevented from being discovered. Since the data no longer contains the actual values, it cannot be misused to violate individual privacy. Randomization approaches were first proposed by Agrawal and Srikant [2] to solve the privacy-preserving data mining problem. Specifically, they addressed the following question. Since the primary task in data mining is the development of models about aggregated data, can they develop accurate models without access to precise information in individual data records? The underlying assumption is that a person will be willing to selectively divulge information in exchange of useful information that such a model can provide. Agrawal and Aggarwal [1] showed that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. They showed that when a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. Evfimievski et. al. [9] presented a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward *uniform*

² Note that the information that P_k obtains from other parties is $S_1 - S_i$ ($i \in [1, k]$) in random order.

randomization, the discovered rules can unfortunately be exploited to find privacy breaches. They analyzed the nature of privacy breaches and proposed a class of randomization operators that are much more effective than uniform randomization in limiting the breaches. Du and Zhan [6] proposed a technique for building decision trees using randomized response techniques which were developed in the statistics community for the purpose of protecting surveyees privacy. The randomization-based methods have the benefits of efficiency. However, the drawbacks are that post-randomization data mining results are only an approximation of pre-randomization results. There are some randomization level control parameters. It has been experimentally shown that for certain scenarios under the control of a randomization parameter, the accuracy of the results can achieve a certain level in the case of both decision tree classification [2,6] and association rule mining [9,20]. Furthermore, the randomization-based method may be invalid to protect data privacy in certain scenarios. Kargupta et. al. [14] analyzed the privacy of random perturbation techniques and showed how to attack the privacy by using random matrix-based data filtering techniques. Although Evfimievski et. al. [8] showed how to limit privacy breaches while using randomization for the privacy-preserving data mining, there are still concerns that randomization-based approaches may allow an attacker to reconstruct distributions and also give out much information about the original data values.

Secure Multi-party Computations (SMC) deal with computing any function on any input in a distributed network. Each participant holds one of the inputs while ensuring that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output. The SMC problem was introduced by Yao [26]. It has been proved that for any polynomial function, there is a secure multi-party computation solution [10]. The approach used is as follows: the function F to be computed is first represented as a combinatorial circuit, and then the parties run a short protocol for every gate in the circuit. Every participant gets corresponding shares of the input wires and the output wires for every gate. This approach, though appealing in its generality and simplicity, is highly impractical for large data sets. Following the idea of secure multiparty computation, people designed privacy-oriented protocols for the problem of the privacy-preserving collaborative data mining. Lindell and Pinkas [17] first introduced a secure multi-party computation technique for classification using the ID3 algorithm, over horizontally partitioned data. Specifically, they consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Du and Zhan [5] proposed a protocol for making the ID3 algorithm privacy-preserving over vertically partitioned data. Lin and Clifton [16] proposed a secure way for clustering using the EM algorithm over horizontally partitioned data. Kantarcioglu and Clifton [13] described protocols for the privacy-preserving distributed data mining of association rules on horizontally partitioned data. Vaidya and Clifton presented protocols for privacy-preserving association rule mining over vertically

partitioned data [21] and provided a solution for building a decision tree without compromising data privacy [22].

Encryption is a well-known technique for preserving the confidentiality of sensitive information. Comparing with other techniques described, a strong encryption scheme can be more effective in protecting the data privacy. An encryption system normally requires that the encrypted data should be decrypted before making any operations on it. For example, if the value is hidden by a randomization-based technique, the original value will be disclosed with certain probability. If the value is encrypted using a semantic secure encryption scheme [18], the encrypted value provide no help for attacker to find the original value. One of the schemes is the homomorphic encryption which was originally proposed in [19] with the aim of allowing certain computations performed on encrypted data without preliminary decryption operations. To date, there are many such systems. Homomorphic encryption is a very powerful cryptographic tool and has been applied in several research areas such as electronic voting, on-line auction, etc. [25] is based on homomorphic encryption where Wright and Yang applied homomorphic encryption to the Bayesian networks induction for the case of *two* parties. Zhan et. al. [29] proposed a cryptographic approach to tackle collaborative association rule mining among multiple parties. In this paper, we will apply homomorphic encryption [18] and digital envelope techniques [4] to the privacy-preserving data mining and use them to design privacy-oriented protocols for privacy-preserving k-nearest neighbor classification problem.

In this paper, we have proposed to use homomorphic encryption and digital envelope technique to achieve collaborative data mining without sharing the private data among the collaborative parties. In particular, we provide a solution for k-Medoids Clustering with vertical collaboration as well as horizontal collaboration. We provide efficient analysis for our solution. The solution is not only efficient but also provides decent privacy protection under our definition. In the future, we would like to examine other privacy-preserving collaborative data mining tasks and implement a privacy-preserving collaborative data mining system.

References

1. Agrawal, D., Aggarwal, C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA, May 21-23, pp. 247–255. ACM Press, New York (2001)
2. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 439–450. ACM Press, New York (2000)
3. Berkhin, P.: Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA (2002)
4. Chaum, D.: Security without identification. Communication of the ACM 28(10), 1030–1044 (1985)
5. Du, W., Zhan, Z.: Building decision tree classifier on private data. In: IEEE International Workshop on Privacy, Security, and Data Mining, Maebashi City, Japan, December 9, IEEE Computer Society Press, Los Alamitos (2002)

6. Du, W., Zhan, Z.: Using randomized response techniques for privacy-preserving data mining. In: Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, ACM Press, New York (2003)
7. Epic. Privacy and human rights an international survey of privacy laws and developments. In: Electronic Privacy Information Center (May 2003), <http://www.epic.org>
8. Evfimievski, A., Gehrke, J.E., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2003), San Diego, CA, ACM Press, New York (2003)
9. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26, pp. 217-228. ACM Press, New York (2002)
10. Goldreich, O.: Secure multi-party computation (working draft) (1998), http://www.wisdom.weizmann.ac.il/home/oded/public_html/foc.html
11. Goldreich, O.: The foundations of cryptography, vol. 2. Cambridge University Press, Cambridge (2004)
12. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
13. Kantarcioglu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. In: DMK 2002. The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Madison, WI, pp. 24-31. ACM Press, New York (2002)
14. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: ICDM 2003. Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, November 19-22, IEEE Computer Society Press, Los Alamitos (2003)
15. Kaufman, L., Rousseeuw, P.: Finding groups in data. Wiley, New York (1990)
16. Lin, X., Clifton, C., Zhu, M.: Privacy preserving clustering with distributed em mixture modeling. In: Knowledge and Information Systems (2004)
17. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, Springer, Heidelberg (2000)
18. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223-238. Springer, Heidelberg (1999)
19. Rivest, R., Adleman, L., Dertouzos, M.: On data banks and privacy homomorphisms. In: DeMillo, R.A., et al. (eds.) Foundations of Secure Computation, pp. 169-179. Academic Press, London (1978)
20. Rizvi, S., Haritsa, J.: Maintaining data privacy in association rule mining. In: Proceedings of the 28th VLDB Conference, Hong Kong, China (2002)
21. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23-26, ACM Press, New York (2002)
22. Vaidya, J., Clifton, C.: Privacy-preserving decision trees over vertically partitioned data. In: 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security Nathan Hale Inn, University of Connecticut, Storrs, CT, U.S.A (August 7-10, 2005)

23. Vaidya, J., Clifton, C.W.: Privacy preserving k-means clustering over vertically partitioned data. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24-27, ACM Press, New York (2003)
24. Weight, J.: Ensuring trust in the electronic health record. In: Electronic Health Information and Privacy Conference, Ottawa, Canada (2005)
25. Wright, R., Yang, Z.: Privacy-preserving bayesian network structure computation on distributed heterogeneous data. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), ACM Press, New York (2004)
26. Yao, A.C.: Protocols for secure computations. In: Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos (1982)
27. Zhan, J.: Privacy Preserving Collaborative Data Mining. PhD thesis, Department of Computer Science, University of Ottawa (2006)
28. Zhan, J., Matwin, S.: A crypto-based approach to privacy preserving collaborative data mining. In: Workshop on Privacy Aspect of Data Mining (PADM 2006) in conjunction with the IEEE International Conference on Data Mining (ICDM 2006), HongKong, December 1, IEEE Computer Society Press, Los Alamitos (2006)
29. Zhan, J., Matwin, S., Chang, L.: Privacy-preserving collaborative association rule mining. In: 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Nathan Hale Inn, University of Connecticut, Storrs, CT, U.S.A (August 7-10, 2005)

Author Index

- Anderson, John R. 205
Ando, Shin 341
Anicic, Darko 367
- Barrett, Natasha 267
Bhamidipati, Narayan L. 324
Bothell, Daniel 205
Bouvier, Jake 164
Brezany, Peter 353
Brezanyova, Jarmila 353
Brodie, Michael 367
- Cao, Longbing 454
Cercone, Nick 59
Chen, Guisheng 36
Chen, Min 428
Chen, Zhengxin 443
- Dai, Ruwei 454
de Bruijn, Jos 367
Duan, Qiguo 428
- Fensel, Dieter 367
- Han, Yanni 36
Hao, Jing 244
Haselwanter, Thomas 367
Hepp, Martin 367
Heymans, Stijn 367
Hirose, Masayuki 341
Hoffmann, Jörg 367
Huang, Minlie 291
- Janciak, Ivan 353
Jankowski, Andrzej 122
Jumi, Masatoshi 341
- Kerrigan, Mick 367
Knoblich, Ulf 164
Kopecky, Jacek 367
Kou, Gang 443
Krummenacher, Reto 367
- Lausen, Holger 367
Li, Deyi 36
- Li, Jiao 291
Li, Jun 267
Li, Ke 244
Li, Kuncheng 1, 223, 244
Li, Wenbin 382
Li, Xingsen 443
Li, Yuefeng 406
Liang, Peipeng 223
Lin, Cui 267
Lin, Tsau Young (T. Y.) 143
Liu, Chunnian 382
Liu, Jiming 1, 223, 382
Liu, Kun 36
Lu, Na 244
Lu, Shengfu 1, 223
Luo, Dan 454
- Matsuoka, Kimiko 471
Miao, Duoqian 428
Mocan, Adrian 367
- Nishida, Toyooki 73
- Ohsuga, Setsuo 97
- Pal, Sankar K. 324
Peng, Yi 443
Poggio, Tomaso 32, 164
Pryczek, M. 305
- Qin, Yulin 1, 205
- Scicluna, James 367
Sekino, Yuko 254
Shan, Baoci 244
Shi, Yong 443
Shirao, Tomoaki 254
Skowron, Andrzej 122
Sun, Ron 190
Suzuki, Einoshin 341
Szczepaniak, P.S. 305
- Tjoa, A Min 353
Toma, Ioan 367
Tomczyk, A. 305
Tsumoto, Shusaku 471
Tsumoto, Yuko 471

Vo, Mong-Hang 143

Wah, Benjamin 1

Wang, Guoyin 278

Wang, Wei 244

Washburn, David A. 267

Wei, Zhihua 428

Wu, Jinglong 1

Xiao, Liping 36

Xu, Jianyang 244

Yan, Bin 244

Yang, Yanhui 223, 244

Yao, Y.Y. 382

Yao, Yiyu 1, 223

Yokoyama, Shigeki 471

Zaremba, Michal 367

Zhan, Justin 494

Zhang, Chengqi 454

Zhang, Yan-Qing 267

Zhong, Ning 1, 223, 382, 406

Zhu, Xiaoyan 291